Spring 5-17-2010

# The Effects of Sanction Intensity on Criminal Conduct: A Randomized Low-Intensity Probation Experiment

Charlotte E. Gill
*University of Pennsylvania*, cegill00@gmail.com

# The Effects of Sanction Intensity on Criminal Conduct: A Randomized Low-Intensity Probation Experiment

**Abstract**
Probation is a well-established part of our criminal justice toolkit, but we know surprisingly little about the circumstances under which it is effective. Attempts to increase supervision intensity for crime- and cost-saving purposes have yielded mixed results at best. This dissertation examines the theory and scientific evidence on the effectiveness of probation, and the impact of changing the intensity of probation sanctions on recidivism.

First, we conduct a rigorous search and synthesis of the existing literature on intensive probation programs. We utilize meta-analysis to identify the circumstances under which such programs might be effective. We find no evidence that probationers in these programs fare better than their counterparts under traditional supervision. We call for further research into supervision approaches that emphasize behavioral management over contact frequency and caseload size. Second, we employ a range of statistical procedures to examine the viability of saving resources by reducing supervision for low-risk offenders. In a randomized controlled trial comparing low-intensity probation to traditional practice, we find no evidence that reducing supervision increases recidivism. We find that low-risk probationers are heterogeneous in their characteristics but homogeneous in their propensity to reoffend. They appear to respond well regardless of the intensity of the sanction. Finally, we use epidemiological methods to evaluate the low-risk prediction model used in the experiment. We find that the model successfully identifies offenders who are unlikely to commit serious offenses, and is therefore a useful tool for diverting probationers to low-intensity supervision. In turn, low-intensity supervision is not associated with changes in offending severity. Chapters 2 and 3 both conclude that low-intensity supervision is a safe strategy that works very well for a probation agency's lowest-level offenders.

This dissertation contributes to knowledge by changing perceptions of the characteristics of offenders and resource allocation in criminal justice supervision. We find that 'more' does not always mean 'better,' and there is no need to distribute expensive services equally. In a given probation population, the majority of offenders will respond well no matter how little supervision they receive, so it makes sense to focus our attention on the minority that will not.

**Degree Type**
Dissertation

**Degree Name**
Doctor of Philosophy (PhD)

**Graduate Group**
Criminology

**First Advisor**
John M. MacDonald

**Second Advisor**
Lawrence W. Sherman

**Third Advisor**
Paul D. Allison

**Keywords**
probation, intensive probation, risk assessment, experimental criminology, meta-analysis

**Subject Categories**
Criminology

# THE EFFECTS OF SANCTION INTENSITY ON CRIMINAL CONDUCT: A RANDOMIZED LOW-INTENSITY PROBATION EXPERIMENT

Charlotte Elizabeth Gill

A DISSERTATION

in

Criminology

Presented to the Faculties of the University of Pennsylvania

In

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2010

Supervisor of Dissertation

_____

John M. MacDonald, Associate Professor of Criminology

Graduate Group Chairperson

_____

William Laufer, Professor of Criminology, Sociology, and Legal Studies & Business Ethics

Dissertation Committee:

John M. MacDonald, Associate Professor, Criminology

Lawrence W. Sherman, Director, Jerry Lee Center of Criminology

Paul Allison, Professor, Sociology

In memory of Edward Routledge & Allan Routledge

# ACKNOWLEDGMENTS

This dissertation could not have been completed without the support of so many wonderful people along the way. I am extremely thankful to:

First and foremost, Jerry Lee. I am so grateful that you gave me the opportunity to study at Penn, and your contribution to the field of criminology is an inspiration.

My committee, John MacDonald, Larry Sherman, and Paul Allison. Your ideas, comments, advice, and teaching really have made this work the best it could be.

Everyone who has worked so hard on the probation experiments, especially Geoff Barnes, Suzanne McMurphy, Jordan Hyatt, Richard Berk, Ellen Kurtz, Lindsay Ahlman, Kevin Reynolds, Janet McHale, and Bob Malvestuto.

My dear friends and colleagues from the UK restorative justice experiments, especially Sarah Bennett, Dorothy Newbury-Birch, Heather Strang, Caroline Angel, Nova Inkpen, Brian Dowling, Helen Orros, Kim Smith, and Taryn Goff, for teaching me about research, policing, the real world, not leaving my valuables on display, and what to do when you get chased by the K-9 unit, and encouraging me to follow what I was good at. Turns out I won't be taking that job at Tesco after all.

All the members of the Campbell Crime and Justice Group steering committee, who have always taken a warm and genuine interest in my progress despite all those nagging emails I send them. Special thanks go to David Weisburd, to whom I am incredibly grateful for recognizing my potential, and Dave Wilson, who despite having literally a million other things to do has answered enough of my stupid questions to qualify as a fourth committee member.

Janel McCaffrey and Knakiya Hagans, who got me through the last four years with both practical advice and sanity maintenance (and a strong dose of humoUr). You guys rock. I'll be sure to put 'coming back to visit you' on my shhhedule.

All my phabulous Philly friends, especially the Brit girls, Faye, Lucy, and Jo, who have been so encouraging and have always been there when I've needed the healing powers of several… cups of tea. Laura D, thank you and your wonderful family for making me an honorary family member and American and helping me settle in long enough to get this done!

My amazing family, whose love, support, and sacrifice got me all the way here. Mum, Dad, and Grannie, thank you so much for understanding (and financing…) my crazy educational aspirations. James (the Met's finest), thank you for laughing at me for "writing about it while I get out there and do it." Perhaps we can compare notes? Grandpa, I wish you were here to see this but you know your photo has been beside me throughout. Allan, I think I finally count as a scientist and I hope I made you proud.

Last but definitely not least, my partner Dan Woods, who has stuck around despite repeated requests for advice, comments, and Mexican food; self-doubt; mood swings; and being woken up at 1a.m. to talk about meta-analysis (but since he's "forever in my debt," I guess he had to). I love you.

# ABSTRACT

THE EFFECTS OF SANCTION INTENSITY ON CRIMINAL CONDUCT:

A RANDOMIZED LOW-INTENSITY PROBATION EXPERIMENT

Charlotte Elizabeth Gill

John MacDonald

Probation is a well-established part of our criminal justice toolkit, but we know surprisingly little about the circumstances under which it is effective. Attempts to increase supervision intensity for crime- and cost-saving purposes have yielded mixed results at best. This dissertation examines the theory and scientific evidence on the effectiveness of probation, and the impact of changing the intensity of probation sanctions on recidivism.

First, we conduct a rigorous search and synthesis of the existing literature on intensive probation programs. We utilize meta-analysis to identify the circumstances under which such programs might be effective. We find no evidence that probationers in these programs fare better than their counterparts under traditional supervision. We call for further research into supervision approaches that emphasize behavioral management over contact frequency and caseload size. Second, we employ a range of statistical procedures to examine the viability of saving resources by reducing supervision for low-risk offenders. In a randomized controlled trial comparing low-intensity probation to

traditional practice, we find no evidence that reducing supervision increases recidivism. We find that low-risk probationers are heterogeneous in their characteristics but homogeneous in their propensity to reoffend. They appear to respond well regardless of the intensity of the sanction. Finally, we use epidemiological methods to evaluate the low-risk prediction model used in the experiment. We find that the model successfully identifies offenders who are unlikely to commit serious offenses, and is therefore a useful tool for diverting probationers to low-intensity supervision. In turn, low-intensity supervision is not associated with changes in offending severity. Chapters 2 and 3 both conclude that low-intensity supervision is a safe strategy that works very well for a probation agency's lowest-level offenders.

This dissertation contributes to knowledge by changing perceptions of the characteristics of offenders and resource allocation in criminal justice supervision. We find that 'more' does not always mean 'better,' and there is no need to distribute expensive services equally. In a given probation population, the majority of offenders will respond well no matter how little supervision they receive, so it makes sense to focus our attention on the minority that will not.

# TABLE OF CONTENTS

# LIST OF TABLES

# LIST OF FIGURES

# PREFACE

This dissertation presents an examination of the impact of changing the intensity of the probation sanction on the recidivism of adjudicated probation clients. Historically, research on probation supervision has revolved around the assumption that more intensive (increased) supervision is the best way to ensure public safety and prevent crime, while saving money relative to incarceration. However, that assumption was not, for the most part, evidence-based. Rigorous research has shown that intensive probation programs yield mixed results at best, and may be even less effective for the least serious offenders. This dissertation addresses two broad questions: What scientific evidence is there for the effectiveness of probation supervision? Does variation in its intensity affect crime outcomes? The following three chapters represent three stand-alone papers looking at different aspects of these questions: a review of the existing evidence; a randomized controlled trial of low-intensity supervision; and a method for predicting low-risk offenders suitable for minimal supervision.

Chapter 1 reports the results of a rigorous search and synthesis of the existing literature on intensive supervision probation (ISP). Although ISP is one of the most thoroughly tested criminal justice interventions, the evaluations have never been identified and synthesized in a methodologically rigorous way. We examine forty-seven experimental and quasi-experimental studies conducted over the last fifty years using meta-analytic techniques to investigate whether and under what circumstances ISP might be effective. We find no evidence that the mostly high-risk probationers on ISP fare any better than their counterparts who receive traditional supervision. Increased supervision

intensity makes no difference to recidivism, and tends to increase the rate of technical violations (which can lead to returns to jail and further criminalization) due to the increased surveillance inherent in the process. However, we find several more recent studies that show more promising reductions in recidivism. These programs tend to focus more closely on the content of supervision, which remains a largely neglected aspect of probation, rather than contact frequency and caseload size. We conclude that the assumption that "more is better" does not necessarily hold true, and that it is more important to ask what probation officers are expected to achieve during supervision.

Chapters 2 and 3 build on the findings from an experiment conducted with the Philadelphia Adult Probation and Parole Department (APPD), in which the agency restructured supervision activities along risk-based lines. Sixty per cent of APPD's caseload was filtered into reduced-intensity supervision based on a statistical prediction that they were at low risk of serious recidivism. The removal of resources from some probationers has met with criticism, despite the longer-term goal of the project: to free up staff to work more closely with the dangerous offenders who increase the public's fear of crime. The goal of these chapters is to provide a rigorous evaluation of low-intensity supervision and the suitability of the prediction model, to ensure first that no serious offenders inadvertently receive too little supervision, and second that reduced supervision in itself is not criminogenic. While Chapter 1 shows no evidence that *increased* supervision prevents serious offenders from committing crimes, it remains important to show that *reduced* supervision does not lead to unfavorable outcomes. The policy cannot work as a resource-saving strategy if agencies view it as too politically risky.

From a theoretical standpoint, Chapters 2 and 3 also provide an insight into the nature of low-risk offenders and the types of crime they commit. Understandably, a great deal of attention is paid in the criminological literature to unpacking the characteristics of more serious offenders, but in developing low-intensity supervision the Philadelphia APPD hypothesized that the majority of its caseload could be classified as low risk. If it is true that the majority of offenders pose little risk of serious recidivism, it is important to learn how they compare to the minority who pose a greater threat.

In Chapter 2 we utilize a range of statistical procedures to break down the main results of the low-intensity supervision experiment, in order to ensure that there are no circumstances under which reduced supervision increases recidivism. We find no evidence that this was the case. We use instrumental variables techniques to model the characteristics of low-risk offenders and their relationship to take-up and outcomes of low-intensity probation. We find that low-risk offenders represent a much broader range of society than the traditional 'young male' offender. However, they have an extremely low propensity to reoffend and appear to perform well regardless of the degree of supervision they receive. Thus, we conclude that it is not necessary to treat all offenders equally when it comes to probation supervision. Standards can be relaxed for most offenders to allow probation officers to spend more time with higher-risk clients, working to identify and address their needs. For such a model to work in practice, a good prediction model is needed to identify who can safely be diverted to the low-intensity unit.

Chapter 3 examines whether the prediction model used in the low-intensity supervision experiment, which classifies offenders as low and non-low risk based on their

risk of committing just the most serious crimes, is an effective tool for the type of risk management strategy employed by Philadelphia APPD. Using methods from the epidemiology field, we assess the sensitivity of the model to the patterns of offending severity exhibited in a sample of probationers. We find that low-risk offenders have a very low propensity for serious offending, while those receiving a non-low risk prediction are much more likely to engage in these offense types. The model appears to successfully recognize this over several different definitions of severity. Furthermore, there is no change in offending severity when supervision is reduced for predicted low-risk offenders. We find further evidence of the homogeneity in this sample's propensity to reoffend regardless of supervision intensity. Predicted low-risk offenders with a history of serious offending respond just the same to reduced supervision as they do to regular supervision. For those without a serious offending history, reduced supervision is even associated with reductions in recidivism compared to the status quo. Thus, we conclude that low-intensity supervision is a safe and effective strategy that works particularly well for a probation agency's lowest-level offenders.

# CHAPTER 1. Is 'More' Really 'Better'? The Impact of Intensive Probation Supervision on Recidivism.

## Introduction

Probation is one of the most frequently-used criminal sanctions in the United States (American Correctional Association, 2006). At the end of 2008, nearly 5.1 million adults were on probation alone – 84 per cent of all adults under community supervision. In all, one in forty-five U.S. adults is on probation or parole.[1] Although growth slowed slightly in 2008, the population under community supervision has been steadily rising for some time, increasing by more than half a million between 2000 and 2008 (Glaze & Bonczar, 2009).

Despite the extent of its use, probation has suffered from image problems, particularly a public perception that it is a 'soft' approach to crime for often serious offenders who are highly likely to recidivate.[2] Subsequently, many probation agencies have struggled to access sufficient funding (Petersilia, 1997). This highlights a clear need for probation agencies to identify supervision practices that are effective at reducing recidivism, and at the same time represent an efficient use of scarce resources. Taxman (2002) notes that considerable research has been dedicated to programming and services that are often provided in conjunction with or on referral from probation, such as cognitive-behavioral therapy, drug courts, and skill-building programs (see also MacKenzie, 2006a; 2006b). Yet comparatively little attention has been paid to the

impact of probation supervision itself on crime: the number of cases a probation officer handles, the frequency of contact between officer and client, and the nature of the interaction. Supervision is perhaps considered an uninteresting part of the probation process, "in the background of other programming" and therefore "inconsequential to effectiveness" (Taxman, 2002, p. 179).

On the contrary, supervision is a crucial aspect of probation not only because it is the bedrock of programming, but also because in a chronically under-funded enterprise it may constitute the only interaction between client and agency. In this regard it may directly impact the client's future criminal behavior. If a probation officer with a caseload of 150 clients has inadequate time to spend with each one, s/he may find it impossible to build an accurate picture of individuals' needs in order to target programming most effectively. Supervision levels vary widely, from weekly or twice-weekly meetings for high-risk or delinquent probationers, to telephone reporting for those near the end of their sentences. In some busy agencies 'supervision' may constitute nothing more than a mail-in contact detail confirmation card (Petersilia & Turner, 1993, p. 285). It is not always clear whether supervision intensity is related to the client's needs or risk, or whether it is simply determined by operational capabilities.

In this paper, we conduct a systematic search for literature on probation supervision intensity and synthesize the results using meta-analytic techniques to present the most current knowledge about the effect of changing intensity on probationers' subsequent criminal conduct. We find that the amount of supervision in itself does not appear to be associated with recidivism outcomes. More supervision may in fact increase

probation violation rates because offenders are at greater risk of detection. The literature on intensive probation also sheds very little light on the nature of effective practices.

## Background

Intensive supervision probation (ISP) is one aspect of probation that has received considerable research attention. ISP programs usually consist of small caseloads and enhanced reporting requirements. However, interest in the practice has evolved from a need to find punitive alternatives to imprisonment rather than a general desire to understand more about supervision practices. As a result, there has been very little articulation of the theoretical basis for its hypothesized effectiveness beyond the assumption that 'more is better.' Indeed, Bennett (1988) described ISP as "a practice in search of a theory."

Skeem and Manchak (2008) propose that probation supervision may follow one of three broad guiding philosophies: control/surveillance, treatment, or a hybrid of both. ISP programs developed over the last fifty years have fallen into all three of these categories, but the 'classic' model has been a surveillance strategy designed to keep track of serious offenders who would otherwise be incarcerated. As such, ISP appears rooted in traditional theories of formal social control and deterrence. Offenders are offered the opportunity to remain in the community on the understanding that they are being constantly monitored, and the consequence of failure is the loss of liberty. Several qualitative studies have noted that most offenders express a preference for incarceration over intermediate sanctions like ISP (e.g., Crouch, 1993; Petersilia & Deschenes, 1994),

which perhaps suggests that ISP is a more unpleasant prospect than prison for adjudicated offenders and could therefore have a strong deterrent effect against future offending. MacKenzie and Brame (2001) suggested an alternative mechanism by which social controls operate through ISP. They proposed that increased supervision intensity could lead to increased involvement in conventional and therapeutic activities, and found some support for that hypothesis through empirical testing. Overall, ISP studies have usually focused on the field testing of programs and avoided any explication of the theoretical foundations of probation supervision.

Clear and Hardyman (1990) describe two waves of interest in ISP research: the first in the 1960s, and another in the mid-1980s. More recently, a third wave of research has refined the application of increased supervision intensity, considering its relationship with carefully matched programming and treatment. The earliest set of field studies of what may be characterized as ISP programs focused on the impact of reducing probation officers' caseload sizes, and followed the 'treatment' philosophy. At the time, the rehabilitative ideal prevailed in corrections, and it was believed that smaller caseloads allowed probation officers more time to help their clients (Petersilia & Turner, 1990). However, these initiatives appeared to make little impact on recidivism, and even increased probation failures and technical violations. Clear and Hardyman (1990) suggest that one important reason for the lack of effectiveness of these initiatives was a lack of insight into how probation supervision activity could best serve the treatment goal. Probation officers simply did not know how to use the additional time made available to them.

The collapse of the rehabilitative ideal and the subsequent 'nothing works' paradigm of the 1970s, along with a sharp rise in crime, led to an exponential increase in prison growth (and the cost of corrections) that has persisted ever since (e.g., Ruth & Reitz, 2003). The probation population was also growing at a similar pace, and probation officer caseloads were becoming too large to allow them to serve the increasing number of serious and high-need offenders being granted probation or parole (Petersilia & Turner, 1993). By the 1980s there was renewed interest in ISP as part of a battery of 'intermediate sanctions' that sought to alleviate prison overcrowding and save money, while maintaining the appearance of being tough on offenders who would otherwise have been incarcerated. The focus was on surveillance and control of the offender through small caseloads, frequent contacts, increased drug testing, and mandatory employment. The new ISP was rooted in the classical theory of deterrence through swift, certain punishment, effected by close supervision (Petersilia & Turner, 1990).

Georgia was the first state in the U.S.A. to implement this new generation of ISP program. Participants had very low recidivism rates, maintained employment, and paid probation fees that helped offset the cost of supervision. The Georgia model was subsequently adopted elsewhere in the United States, with mixed results. The Bureau of Justice Assistance (BJA) responded to the interest in and uncertainty about the Georgia model by funding a large, multi-site randomized controlled trial in the mid-1980s, which was evaluated by the RAND Corporation. Twelve of the fourteen experiments compared ISP to routine supervision, while two compared ISP to incarceration. By and large, the results of the evaluations were disappointing, again showing little impact on new crimes and an increase in technical violations compared to usual practice. Furthermore, a

program intended to reduce the strain on the prison system actually resulted in more incarcerations, as increased surveillance and drug testing raised the likelihood of probation failure (Petersilia & Turner, 1993).

The inability of ISP to demonstrate potential as a crime prevention program under the scrutiny of a rigorous research design largely killed off interest in the surveillance/control model of probation supervision by the 1990s. ISP was listed in the influential University of Maryland report to the United States Congress, *Preventing Crime: What Works, What Doesn't, What's Promising*, as a program that did not work (Sherman et al., 1997; MacKenzie, 2006b). However, the 'what works' movement also led to an increased focus on the factors that influence successful programming. Andrews, Bonta, and Hoge (1990) introduced what are now commonly described as the 'principles of effective intervention' (PEI), which posit that programs should be designed to be responsive to offenders' specific risk and need levels (the risk-need-responsivity, or RNR, model: see also Taxman & Thanner, 2006). The risk principle in particular suggests that more intensive supervision and treatment should be targeted at higher-risk offenders, an idea that is strongly supported by empirical research (see Lowenkamp, Latessa, & Holsinger, 2006, for a summary). The PEI suggest that ISP might be more effective if, through increased contact and control, the probation officer were able to establish offenders' risk and need levels and direct them into appropriate treatment.

Treatment provision was not a priority of the BJA/RAND-evaluated programs, and few participants received such services (Latessa et al., 1998). However, results from some of the study sites indicated that intensive supervision combined with treatment might have a positive effect on crime, which led the evaluators to call for more research

into such interaction effects (Petersilia, Turner, & Deschenes, 1992a; Petersilia & Turner, 1993). Several more recent studies also suggest that ISP programs that adhere to the PEI and offer a balance of treatment and surveillance (the 'hybrid' philosophy) show promise in improving offender outcomes (e.g., Latessa et al., 1998; Paparozzi & Gendreau, 2005). A recent meta-analysis of a wide range of correctional interventions also supports the contention that modern treatment-focused ISPs are more effective at reducing recidivism than surveillance-based programs (Aos, Miller, & Drake, 2006). MacKenzie (2006b), in a detailed update to the University of Maryland report, lists intensive supervision with a treatment component as a 'promising' strategy in corrections, which means that further rigorous research is needed but several studies have produced encouraging results.

Uncertainty about the effectiveness of ISP indicates a clear need for work to unpack the complex relationships between surveillance and treatment, probation officer and client. Taxman (2008a) notes that efforts are now under way to effect organizational change in probation departments that will allow for greater rapport-building between officers and offenders, which is intended to lead to behavioral change. She is currently leading experimental research into "proactive" and "seamless" criminal justice supervision and treatment programs that embody these new directions and have so far shown substantial reductions in recidivism for participants (Taxman, 2008b). A recent randomized controlled trial in Hawaii indicated that intensive probation programs rooted in the classical deterrence tradition may be effective when a consistent, incentive-based structure is implemented. The Hawaii HOPE program combined increased drug testing with swift, certain adjudication and shock incarceration for violations. A novel aspect of the program was the handling of violations. Non-compliant offenders continued their

supervision with probation officers trained in therapeutic techniques, and repeat violators were directed to treatment services as well as being punished (Hawken & Kleiman, 2009).

Taken together, the research on ISP to date suggests a complex dynamic that goes beyond earlier assertions that the programs do not work. Furthermore, even less is known about the converse of ISP: increasing caseloads and reducing contacts ('low-intensity' supervision). The PEI would suggest that ISP be reserved for the highest-risk offenders, with reduced surveillance and services for those at the lowest end of the risk-need spectrum. There is some speculation that increased caseloads can lead to harmful reductions in supervision, putting society at risk from offenders whose probation officers have too many clients to ensure that each one is not a threat to public safety (e.g.,Worrall et al., 2004;[3] Lemert, 1993). However, Glaser (1983) speculated that reduced frequency of contact would not adversely affect low-risk or low-need clients. This suggestion is supported empirically, notably by a recent randomized experiment (Barnes et al., forthcoming; also Johnson, Austin, & Davies, 2003; Wilson, Naro, & Austin, 2007). Additionally, several studies have indicated that more intensive supervision can have unfavorable effects on the recidivism of low-risk offenders (Erwin, 1986; Hanley, 2006; Lowenkamp, Latessa, & Holsinger, 2006). We still have much to learn about probation and parole supervision, and the circumstances under which its use is effective in reducing crime.

## The Present Study

The overall aim of the present study is to undertake a comprehensive review and synthesis of the most rigorous research available on the effects of probation supervision intensity on recidivism. The focus of the review is programs that include among their primary features a change in the ratio of probationers to probation officers (caseload size), frequency of contact between officers and clients, or other 'frontline' supervisory behavior, such as drug testing. The effects of these changes are tested against a counterfactual of 'supervision as usual' – offenders who remained part of standard probation caseloads. The primary outcome measure is recidivism, as measured by arrests, charges, or convictions. We also examine the impact of probation intensity on technical violations.

As we have seen, there is conflicting evidence about the effectiveness of increasing the intensity of probation supervision. It may depend on the specific philosophies and components of the programs and how they interact with supervision levels. The risk and need levels, and other characteristics, of offenders who participated in ISP research studies may also impact the relative effectiveness of the programs. We systematically code the characteristics of each program and sample to examine which, if any, of these characteristics moderate the overall effect of the change in intensity.

The specific research questions we address in this systematic review are:

1. How does the degree of probation supervision intensity affect probationers' subsequent offending and technical violations?
2. To what extent does program philosophy (treatment, surveillance, or hybrid) influence the success or failure of changes in supervision intensity?

3. To what extent do the risk/need levels of program participants affect their response (in terms of reoffending and violations) to changes in supervision intensity?
4. Which other program components or offender characteristics moderate the overall effect of supervision intensity on crime?

## Systematic Review Methodology

Criteria for inclusion and exclusion of studies in the review

*Types of Interventions*

Eligible studies will test the effect of a change in intensity of probation supervision on subsequent crime. A change in intensity could be brought about by increasing or decreasing the ratio of clients to probation officers (changing caseload size); increasing or decreasing the frequency of contact between clients and their officers; or increasing or decreasing the frequency of other forms of supervisory control effected by probation officers, such as drug testing.[4] Studies in which the primary purpose of the research design is to estimate the impact of these specific measures on recidivism and/or technical violations are considered. Most studies have tested increases in intensity rather than decreases, but changes in both directions are eligible for inclusion in the review.

We impose a number of restrictions on program type in order to preserve comparability between what we already know will be a highly diverse set of studies. Some programs have examined the provision of supervision as part of a 'team' approach; for example, multi-agency collaboration between probation officers, police officers, and

treatment providers. Evaluations of these programs are eligible as long as the probation officer is the primary supervisor. This limitation allows us to maintain a degree of equivalence between treatment providers and settings, and between treatment and control group conditions. For example, we included a study in which probation officers provided increased supervision by frequently visiting clients' homes accompanied by a police officer (Piquero, 2003). However, we excluded a study in which the only difference in supervision intensity between the treatment and control groups was that treatment group probationers were assigned police officers who made unannounced visits during their regular patrol shifts to monitor probation compliance (Giblin, 2002).

We also restrict our analysis to the study of adjudicated offenders sentenced to probation or granted parole. Probation services may also be provided at the pretrial stage, or as part of diversion strategies for first-time juvenile arrestees or 'pre-delinquent' adolescents. We hypothesize that there may be substantial differences in the offending propensities of participants in these programs compared to adjudicated offenders, particularly because offenders at the pretrial stage are not guaranteed to receive any conviction or sentence. There is also no straightforward comparison condition to pretrial probation in the same way that 'supervision as usual' simply involves more or less of the same intervention.

*Types of studies*

We attempt to maximize internal validity in our selection of studies by limiting the sample to studies meeting at least a 'high' Level 4 on the Maryland Scientific

Methods Scale (SMS: Farrington et al., 2006). The SMS is a 5-point methodological rating scale, on which 1 indicates the least reliable research design (a one-group study with only post-intervention outcomes), and 5 represents the most rigorous design (random assignment of multiple units to treatment and comparison groups). Level 3 designs (non-comparable treatment and control units) are generally accepted as the 'minimum interpretable' research design (Cook & Campbell, 1979). Level 4 studies are quasi-experimental designs including multiple treatment and comparison units, pre- and post-program measures of offending behavior, and controls for potential bias from confounding factors through matching of treatment and comparison subjects or multivariate statistical controls. We limit our analysis only to those studies that utilize strong quasi-experimental designs involving at least subject-level matching, or randomized controlled trials (RCTs). We justify these strict inclusion criteria on the basis of *a priori* knowledge of a large body of the highest-quality research on ISP. The BJA/RAND studies alone were the largest randomized experiment in corrections undertaken in the United States at the time (Petersilia & Turner, 1993, p. 292). Thus, we expect to find sufficient numbers of experimental and quasi-experimental studies meeting our other eligibility criteria to permit a meta-analysis to be conducted.

The control condition must be regular probation or parole supervision ('supervision as usual'). This may vary widely between studies in terms of number and type of contacts, caseload size, and so on, as long as the control group participants are exposed to the regular practices of the probation agency. The specific components of the control group are coded. In some evaluations, ISP programs based on the 'Georgia model' were compared to the agency's existing intensive supervision program, rather

than 'routine' probation (e.g., Ventura County, California: Petersilia & Turner, 1990). We consider these studies for inclusion as long as there are differences between the existing and experimental ISPs that meet the requirements set out in the previous section. Evaluations in which ISP is compared to incarceration or a different program (e.g., a boot camp) are excluded. The aim of this review is to investigate the impact of changing probation/parole supervision intensity, so our baseline for assessing such change must be probation/parole supervision of a different intensity than that received by the treatment group.

*Types of Participants*

We include both juvenile and adult probationers in the review. Since probation agencies supervise a broad range of offenders, most studies will include mixed caseloads of male and female offenders with different risk and need levels and varying offending histories. However, we expect that most participants will be the moderate to high-risk male offenders usually targeted in high-intensity probation programs. Some experimental ISPs were directed at specific offending problems (e.g., focusing on drug-involved offenders), while others accept a range of offender types. Many probation and parole agencies do not have different policies for the supervision of probationers as compared to parolees, so studies may include mixed caseloads. Specific details about all these variations are coded.

*Types of Outcomes*

Eligible studies measure recidivism in terms of new arrests and/or convictions. Technical violations of probation, such as absconding or failing a drug test, are also included as a separate outcome measure. While technical violations do not inevitably result in a recorded arrest or charge for a new offense, they represent a failure to comply with probation conditions that could be affected by the intensity of supervision.

The use of technical violations as an outcome measure comes with the caveat that increased supervision intensity could increase the likelihood of a violation being detected through increased surveillance, rather than simply a failure to comply. This caveat applies to new criminal cases too, but to a lesser extent. New crimes are more likely to be detected by the police than by probation officers, so future arrests are less likely to be affected by the offender's probation status. This also makes arrest a preferable outcome measure to charges or convictions that come further along the criminal justice process and may be more affected by disclosure of prior sentences. Of course, police officers in smaller beat areas probably know the repeat offenders too and will adjust their discretion to arrest accordingly. All recidivism measures suffer from inherent limitations.

Offending measured by self-report is not excluded, but most ISP studies use official records. This is a limitation of our research: it is well-known that official records can underestimate the prevalence of reoffending, and there may be confounding between the treatment and response that could be partly overcome by using self-reports. However, these data were simply not available to the extent needed to conduct a meaningful analysis.

Search strategy for identification of relevant studies

We used several strategies to conduct a comprehensive search for literature on probation intensity. Our main source of information was the Internet. References were found through keyword searches of online abstract databases and the websites of research organizations and government agencies (see Appendix A for lists of keywords and the databases and websites searched). Specialist search engines like Google Scholar also provide a rich source of 'grey literature.'[5] We also consulted lists of references from existing reviews of probation supervision and intensity, and of randomized trials in general (Petersilia & Turner, 1993; Phipps et al., 1999; Taxman, 2002; Weisburd, Sherman, & Petrosino, 1990), and book and microfilm collections at the University of Pennsylvania library. We supplemented the online searches with hand searches of key journals in the field.[6] Every effort was made to locate unpublished material where possible. Most agencies now make reports available online for review or download, meaning we were able to obtain most of our references very quickly. However, many of the older reports that have not been digitally archived are harder to access. The University of Pennsylvania's Inter-Library Loan service proved useful in locating some of these studies. Electronic references were downloaded to Zotero, a web-based program that captures, stores, and manages references.[7] Eligibility of studies was assessed by reading titles and abstracts, and obtaining the full text of documents that appeared to be relevant.

Description of methods used in primary research

We located a wide range of evaluations testing a change in probation intensity. However, the BJA/RAND experiments from the 1980s (Petersilia & Turner, 1993) represent the 'classic' ISP model, and serve as a convenient illustration of a typical study design. The BJA/RAND studies were a fourteen-site randomized controlled trial of largely surveillance/control-oriented ISP programs. Two of the study sites compared ISP to incarceration (so were not eligible for inclusion in this review), while the remaining twelve contrasted ISP with supervision as usual (SAU) or existing intensive supervision models. Enhancements of both probation and parole supervision were tested. The exact nature of the program depended on the study site – each jurisdiction selected components of the Georgia ISP model for inclusion as it saw fit. Key common features of all the evaluations included smaller caseloads of around 25-30 offenders per officer (usually compared to 100 or more in SAU), increased frequency of contact (usually at least once a week at first, gradually decreasing in phases), drug testing, and mandated employment.

Participants in the ISP evaluations had to be adults. Their risk levels varied, but they were generally more serious offenders. Petersilia and Turner (1993) state: "People placed on enhancement ISPs [as opposed to prison diversion or early release] are generally deemed too serious to be supervised on routine caseloads" (p. 292). However, persons convicted of homicide, robbery, or sex crimes were excluded as a matter of policy from the experiment. Participants were primarily males in their late twenties to early thirties, with extensive criminal records. A substantial proportion of participants were drug dependent. The study sites set their own eligibility criteria for participants beyond these initial requirements. Participants were randomly assigned to treatment and

control conditions by RAND researchers. The study sites implemented the randomization sequence.

Data collection occurred in several waves. A baseline assessment of demographic characteristics and criminal history was conducted shortly after assignment. Supervision details and services received were recorded at six and twelve months; and recidivism (proportion with new technical violations, arrests, convictions, and incarcerations) was recorded at twelve months. Data on drug testing were collected monthly. Cost data and calendars for assessing time at risk were also collected. Each site obtained its own data, and procedures were checked for validity by RAND staff. Recidivism data came from official records rather than self-reports.

Criteria for determination of independent findings

Many ISP studies report data on multiple outcome measures, which cannot be considered independent treatment effects for the purposes of quantitative meta-analysis because they are taken from the same sample of participants. In this review we do not attempt to pool outcome measures. As described above, the different outcome measures can be affected in different ways by the offenders' probation status. We initially take the more conservative approach of handling different types of outcome measure separately. However, we combine arrests and convictions in some analyses. In these cases, arrest outcomes take precedence over convictions so that multiple outcomes from the same study are not used. We prioritize arrest because a successful conviction is dependent on many external factors and may not represent the most accurate picture of the offender's

actual behavior. We analyze technical violations separately because of the strong likelihood that they will be related to the treatment condition due to the increased surveillance inherent in ISP programs.

In the event that samples or outcomes are broken down by subgroups (e.g., new arrests are reported for the full sample and then broken out into drug, property, and violent crime arrests), we use the data for the full sample or outcome only. Where enough studies provide results broken down by the same types of subgroups, we analyze those outcomes separately. Some studies report only felony or misdemeanor arrests or convictions, but do not combine the two. In these cases we prioritize felony offenses, given that ISP is generally used with more serious offenders.

A related threat to the independence of findings is the measurement of follow-up outcomes for the same sample at multiple time periods. In such cases, the longest follow-up period is preferred. However, in some studies we obtained, sample sizes decreased significantly over time as cases were lost to follow-up. In those cases we selected the follow-up period with the closest number of cases to the original sample size to minimize bias from attrition.

Where multiple reports are based on the same dataset or sample, we combine results where possible, counting the sample as one study. The study containing the longest follow-up period and/or the most detail is considered the primary study, and other reports are used to supplement the data from the primary study where necessary. We checked each coded document carefully to ensure that re-analyses of the same datasets were not inadvertently included with primary evaluation data from the same research project. We do include several studies conducted in the same jurisdiction (usually no

smaller than the county level) where the dates of the study periods indicate that there is little risk of project or participant overlap. Where multiple treatment groups are compared to a single control group (e.g., Haapanen & Britton, 2002), we select one treatment group only to maintain independence of control group samples. The treatment group is selected at random to avoid bias in the overall effect size.

Details of study coding categories

A systematic review can be thought of as a survey in which the respondents are studies rather than people. Each retrieved report is 'interviewed' using a survey instrument (coding protocol) to obtain information relevant to our analysis. The coding protocol developed for this study is reproduced in Appendix B. It is designed to capture the hierarchical nature of evaluation data: a single study may report separate effect sizes for multiple outcome constructs for multiple samples in multiple treatment-comparison contrasts or study sites ('modules'). We recorded a range of methodological details about each study to assist in decision-making about eligibility and study quality. A host of items capturing information about program, setting, and participant characteristics served as both determinants of eligibility and potential moderator variables. We did not expect all these factors to influence outcomes and did not test each one to minimize the risk of finding results that were statistically significant merely by chance. However, we also aimed to be as inclusive as possible so that potentially relevant information was not missed.

Treatment of qualitative research

Qualitative research studies are not included in the systematic review results, but relevant qualitative data are used to inform the background, framing, and analysis of our questions. The broad definition of our search terms allows qualitative studies to be systematically identified in the literature searches.

## Analytic Strategy

Meta-analytic procedures are used to quantitatively combine effect size data from the eligible studies where appropriate (i.e., where two or more studies were available that measured a common outcome, such as arrests, and contained sufficient information to calculate an effect size). Effect sizes for each outcome measure in the studies are encoded according to procedures outlined in Lipsey and Wilson's (2001) guide to meta-analysis. The type of effect size chosen depends on the form of the original outcome measure. Most evaluations of ISP include dichotomized measures of the prevalence of recidivism or technical violations (e.g., the proportion of offenders arrested/not arrested). This type of data is suitable for calculating odds ratios (OR). The odds ratio compares two groups on the relative odds[8] of an event (e.g., arrest) occurring (see Appendix C). The odds ratio is centered at 1, so OR=1 indicates no difference between the treatment and control groups on the outcome measure. In our analyses, OR > 1 indicates a result that favors the control group (i.e. recidivism increases following assignment to intensive probation), and OR < 1 indicates a result that favors the treatment group (assignment to intensive probation is associated with reduced recidivism). The events of interest here

(arrests, convictions, violations, etc.) are unfavorable and the intention of the change in supervision intensity is to reduce their prevalence. Thus, a smaller effect size implies fewer events, which is the goal of the programs being tested.[9]

The synthesis of effect sizes in a meta-analysis also requires the calculation of a weight for each effect size. Without the inclusion of the weight, each study's effect size is assumed to contribute equally to the overall (mean) effect size. This is unjustified because smaller studies have greater sampling error and should not contribute as much to the mean outcome as larger, relatively more reliable studies. Lipsey and Wilson (2001, p. 36) suggest that the optimal study weight is based on the inverse of the squared standard error of the effect size (called the 'inverse variance weight'). Formulas for calculating the standard error and inverse variance weight for the OR are presented in Appendix C.

Computations of effect sizes and inverse variance weights, and calculation of the mean effect sizes and corresponding confidence intervals and statistical tests, are performed using specialized meta-analysis macros written for STATA software (Wilson, 2002). We use RevMan software (Cochrane Collaboration, 2008) to construct forest plots for the graphical representation of meta-analysis results. The forest plot shows the weighted mean effect size and associated 95 per cent confidence interval for each study. A square represents the point estimate of the effect size, the size of the square represents the study weight, and the lines on either side of the square are the confidence intervals. The mean effect size across all studies is displayed as a diamond, whose far left and right points represent the lower and upper bounds of that estimate's confidence interval. The plot is centered around 1, the point at which no difference is observed between treatment and control groups. Point estimates to the left of center represent outcomes favoring the

treatment group, and those to the right favor the control group. When the confidence intervals do not touch or cross the center line, the point estimate is statistically significant.

We assume a random effects model, rather than fixed effects, for all analyses (see Appendix C for details). Fixed effects models in meta-analysis assume that the only random error in the distribution of effect sizes arises from *within*-study sampling error. We do not consider this to be theoretically justified in our analysis because of the considerable *between*-study differences (heterogeneity: see below) in program characteristics, settings, and populations. Further, because we know we have not been able to capture all the available research on intensive probation programs, we can consider our set of studies a sub-sample of a larger 'population' of studies, with its own sampling error. Both of these factors justify the use of the random effects model (Lipsey & Wilson, 2001, pp. 117-120).[10]

Meta-analytic methods are also used to investigate whether the overall mean effect size is moderated by other factors. We are interested in the potential impact of certain program and offender characteristics on the variation in effect sizes across studies. Because all our moderator variables are categorical and we have a small set of *a priori* hypotheses about potential moderators, such as risk/need level and supervision philosophy, we use the meta-analytic analog to the analysis of variance (ANOVA) to test whether these factors might account for any variability in the observed effect sizes from each study (Lipsey & Wilson, 2001, pp. 120-122). We assess each categorical variable separately using this strategy. Even though we include a substantial number of studies in this meta-analysis, cell frequencies became very small when they were broken out by

research design, outcomes, and different levels of each moderator variable. Therefore, we focus only on bivariate comparisons on each moderator, and do not attempt to model outcomes any further. This is a limitation of our moderator analysis: the results we present do not control for the presence of any additional moderators.

The analog to the ANOVA is based on the *Q*-statistic calculated as part of the main random effects model. *Q* is the weighted sum-of-squares of each effect size around the grand mean. It represents the extent to which differences between the effect sizes are statistically related to differences in moderators (a statistically significant *Q*-statistic indicates evidence of between-study heterogeneity). We use the random effects analog to the ANOVA (also called a 'mixed effects' model), which assumes there is still unmeasured variability after moderators are modeled. We justify this on the basis of our limited set of moderators, which are unlikely to explain all the variability between studies.[11] The relevant formulas are presented in Appendix C. These analyses are also performed using the STATA macros.

Due to the greater risk of bias in non-randomized studies, experimental and quasi-experimental results are treated separately in all analyses. Randomized experiments that indicate large baseline differences between participants on characteristics likely to be related to outcomes (such as prior offending history), or which experienced substantial attrition of participants or other implementation problems are analyzed with the quasi-experiments. The concern with such experiments is that the attrition may be caused by reasons related to the treatment and/or outcome; for example, higher-risk offenders may be more likely to abscond from probation and be subsequently lost to follow-up, thus offending outcomes for the remaining lower-risk offenders are biased.

## Discussion of Results

Systematic search results

We present detailed search results in Appendices A and D. Appendix A includes a flowchart describing how the number of database 'hits' obtained through the systematic search translated into the 47 studies included in the final analysis. Our search terms initially produced 30,591 hits. Recall that we deliberately left the search terms broad to pick up background information as well as evaluations, so this number by no means reflects the total number of studies available. It also includes a substantial number of duplicate hits both across and within databases. We identified 528 references to potential evaluations of changes in probation intensity, of which 410 were put forward for more detailed title and abstract screening.

We identified 239 references requiring full coding, and obtained 81 per cent (n=194) of these. Most of the studies we could not obtain were the tests of caseload size variation conducted by local government agencies (mostly the state of California) in the 1950s and 1960s. We do not feel that our results are greatly biased by these missing studies, because literature reviews of ISP have indicated that they found similar results to the research of the 1980s and early 1990s – ISP had no effect on recidivism and increased technical violations.

Of the 194 reports we coded in full, 21 were eligible for inclusion in the meta-analysis, 102 were evaluations that did not meet our eligibility criteria (this number reflects some multiple reports of the same study), and the remainder were either relevant background literature or additional reports of eligible studies that were used as

supplements to the primary report. We list the supplemental reports and ineligible studies (with reasons for exclusion) in Appendix D.[12] Most of the excluded studies tested changes in probation intensity, but had no comparison group or unmatched controls. A few involved high-quality research designs but the comparison groups did not include regular probationers (in most of these cases ISP was compared to incarceration). Many of the 21 eligible reports contained data for multiple study sites that could be treated as separate evaluations for the purposes of the review. For example, Petersilia and Turner (1990) reported on ISP experiments in three California counties. Thus, our final sample contains 47 independent evaluations of probation intensity variation.

Description of eligible studies

Of the 47 evaluations we include in our review, 38 were randomized trials and 9 were either matched-pairs designs or RCTs that reported high attrition. The unusually large number of RCTs for a systematic review of a criminal justice system intervention reflects the interest in obtaining rigorous data on the effectiveness of intermediate sanctions (largely funded by the U.S. government) as crime and the incarceration rate rose between the 1970s and 1990s. All the evaluations tested increased probation intensity.[13] We present specific details about each eligible study in Appendix D, and summarize their characteristics in Table 1.1. Studies are listed in Appendix D first by research design (RCTs then quasi-experiments), then by the report date.

Table 1.1 shows that almost all of the studies were conducted in the United States (N = 42). Five experimental studies were conducted in the United Kingdom, all of which

were evaluated in the same report (Folkard, Smith, & Smith, 1976). More than half of the studies were reported in government or technical reports rather than traditional academic sources. Almost all the studies were conducted in the 1990s or earlier.[14]

Almost all the ISP studies were enhanced probation or parole initiatives rather than prison diversion programs. Reduced caseloads were the primary component of the test of increased intensity, although most studies also involved increased contact as a natural consequence of the smaller caseload, even if there was no set protocol for contact frequency. In a few cases increased drug testing was the only difference in supervision intensity between the treatment and control groups (e.g., Haapanen & Britton, 2002; Hawken & Kleiman, 2009). Following Skeem and Manchak's (2008) model, we assessed the prevailing supervision philosophy as control/surveillance in approximately 40 per cent of studies, treatment in 16 per cent of studies, and hybrid in 45 per cent.

Comparison groups in the eligible studies almost always experienced routine probation supervision. A handful of programs compared ISP to existing, less restrictive ISPs. "Routine probation" covered a myriad of conditions that were often not reported in great detail, but as much information as possible about what it comprised is recorded in the 'Comparison' column of the table in Appendix D. It usually involved larger caseloads, less frequent contacts, and fewer services. Most evaluations targeted general offender caseloads comprised mostly of probationers rather than parolees.

Because ISP is a general supervision strategy that can be implemented across the board in probation agencies, we found lots of variation in participant characteristics across studies. For example, we do not analyze the racial composition of participants because this was highly dependent on the characteristics of the general population in the

agency's jurisdiction. Table 1.1 shows some limited participant characteristics that could be generalized across studies. We found a mix of studies evaluating programs for either juvenile probationers, or youth (18 and over) and adults. Study samples comprised mostly male offenders. Only two studies examined ISP versus SAU in exclusively female caseloads.[15] Most samples comprised high or mostly high risk offenders (as assessed by classification instruments or offending history), reflecting the fact that ISP has usually been used as a means of community supervision for more serious offenders who might otherwise have gone to prison. Needs assessments were not frequently discussed so we do not present these results, but where needs assessments were carried out most ISP offenders were classified as high need for services, often based on drug and alcohol dependencies.

Overall mean effects of probation supervision intensity on recidivism

Table 1.2 shows the results of the main analysis examining how probation supervision intensity is related to subsequent offending and technical violations. Each row of the table, along with the estimated effect sizes for each included study, is visually represented in separate forest plots (Figures 1.1 to 1.7). Across the 47 studies, we obtained a total of 213 different outcome measures. The present study makes use of those that measure the prevalence of arrests, drug arrests, convictions, and technical violations.[16]

The results reported in Table 1.2 are consistent with what we already know about intensive supervision. None of the mean effect sizes is statistically significant, and the

direction of effects is as we would expect given prior research. In the RCTs, assignment to intensive supervision made no difference to the prevalence of rearrest or reconviction (mean OR for arrests = .93; $p \leq .72$; mean OR for convictions = .98, $p \leq .80$). We also see no significant effect of ISP in the quasi-experiments, although note that the raw effect sizes (especially for convictions) show moderate reductions in recidivism associated with ISP and the small number of studies in these categories reduce the likelihood of a finding being statistically significant (mean OR for arrests = .83, $p \leq .10$; mean OR for convictions = .60, $p \leq .10$). The forest plot for conviction outcomes (Figure 1.4) suggests that there is a lot of uncertainty in this model: the confidence interval around the mean effect size is clearly very large.

Our analyses also indicate an increase in technical violations associated with ISP. Across the RCTs, intensive supervision was associated with a 54 per cent increase in the odds of a technical violation (mean OR = 1.54, $p \leq .06$). A smaller, non-significant increase of 29 per cent was observed across the quasi-experiments (mean OR = 1.29, $p \leq .22$) but is again based on a much smaller subset of studies. Finally, we found no effect of ISP for the subset of studies reporting drug related effects (mean OR = 1.14, $p \leq .10$). The mean effect sizes should be interpreted with caution given the very small number of events in some of the studies (see Figure 1.7).

Table 1.2 indicates substantial heterogeneity across studies in four of our seven analyses, as evidenced by the highly significant $Q$ statistics (RCT arrests: $Q = 61.55$, $p < .001$; RCT technical violations: $Q = 120.11$, $p < .001$; quasi-experiment technical violations: $Q = 18.91$, $p \leq .004$; quasi-experiment convictions: $Q = 22.74$, $p < .001$). This indicates that there is more variability between studies than we would expect from the

sampling error within each study alone. This suggests that other unique characteristics of each study, such as program and offender characteristics and study setting (all of which could also be confounded with each other), might explain the heterogeneity. Thus, the overall effects of ISP on recidivism compared to SAU may be moderated by some of these explanatory variables.

Moderator analyses

We focus only on arrests/convictions and technical violations for the experimental sample in these analyses. We only have a maximum of 9 quasi-experiments, so frequently have insufficient observations in each category of the moderator variables to conduct a meaningful analysis. We justify combining arrests and convictions for the recidivism analysis because it substantially increases our sample size (seventeen RCTs reported only conviction outcomes) and our main effects analysis showed that there was no difference in the effects of ISP on either arrests or convictions. Furthermore, we did not observe significant heterogeneity across conviction outcomes, so we have no reason to believe that the effect of ISP on convictions could be moderated by other factors that are not related to arrest outcomes. The overall mean effect size and *Q*-statistic for the combined RCT arrest/conviction studies were very similar to those for RCT arrests.[17]

*To what extent does program philosophy influence the effect of changes in supervision intensity on recidivism?*

Table 1.3 indicates no effect of program philosophy on recidivism (arrests and convictions) in the 38 RCTs included in the study, based on the meta-analytic analog to the ANOVA ($Q_B$ = 1.73, $p \leq .421$). Most of the studies were of surveillance/treatment hybrid programs (N = 17). In these programs, the odds of failure in the treatment group were marginally smaller than in the treatment group, but no real effect is observed (mean OR = .93, $p \leq .525$). Similar results were observed in the 15 surveillance-based programs (mean OR = .92, $p \leq .475$). Interestingly, offenders in treatment-based programs did have 20 per cent greater odds of recidivism than their counterparts in regular probation, but the effect size is not statistically significant, and its reliability is questionable because it is only based on 6 studies, 5 of which were conducted by the same evaluators (mean OR = 1.20, $p \leq .314$). Unfortunately, we were unable to build a good picture of the moderating effect of supervision philosophy on ISP and technical violations because none of the studies reporting technical violation outcomes followed a treatment-based model (Table 1.4). As we would expect, technical violations were higher in both control-based (N = 13) and hybrid (N = 3) programs. For both program types the odds of a technical violation was about fifty per cent greater in the treatment groups than the control groups (mean OR for control-based: 1.55, $p \leq .091$; mean OR for hybrid: 1.46, $p \leq .481$). However, the statistical difference between these estimates is negligible ($Q_B$ = .01, $p \leq .923$) .

*To what extent do the risk/need levels of program participants affect their response to changes in supervision intensity?*

Based on the principles of effective intervention (PEI), we speculated that supervision may be more effective if more intensive programs are targeted at the highest risk/need offenders, and vice versa. There are several impediments to assessing this question in great detail. First, we could not examine whether low-intensity probation is effective for low-risk probationers across a range of studies. Second, very few studies included needs assessments, and those that did lacked detail. Thus, we decided not to use the need variable in our analysis. Finally, many of the studies discussed the PEI or evaluated programs that had made an attempt to target higher-risk offenders, so we do not have much variation in our data. We are only able to examine whether programs including either all or a majority of high-risk offenders were more or less likely to prevent reoffending than those including offenders of any risk level or not utilizing a risk assessment. A further caveat related to this final point is that programs that did not formally assess participants for risk may still have had inclusion criteria that targeted more serious offenders.

The results in Table 1.3 may lend some support to this caveat. There is no difference between programs that targeted higher-risk offenders and those that accepted any offender type ($Q_B < .01$, $p \leq .981$), and neither risk category is associated with recidivism outcomes. The odds ratio for studies involving higher risk offenders was .96 (N = 11, $p \leq .798$), and for all risk levels .97 (N = 27, $p \leq .735$). For technical violations, there is also very little variation in the odds of failure by risk level ($Q_B = .04$, $p \leq .836$; high risk: N = 10, OR = 1.59, $p \leq .092$; mixed risk: N = 6, OR = 1.45, $p \leq .292$).

*Which other components of programs or offender characteristics moderate the overall effect of supervision intensity on crime?*

We examined a range of other potential moderators of the effect of supervision intensity on recidivism and technical violations, the results of which are listed in Tables 1.3 and 1.4 respectively. As before, our analyses are somewhat limited by the reduced cell frequencies when studies are broken out by each level of the moderator variables. We group our outcomes into three main categories: study characteristics, program characteristics, and sample characteristics. The varied program characteristics are particularly important because of the range of different activities that constitutes ISP in each study.

Among selected study characteristics (Table 1.3), we found that only the type of publication was significantly associated with ISP recidivism outcomes ($Q_B$ = 11.86, $p \leq$ .003). Of course, the publication type does not directly influence the outcome of a study, but these results are important because they show that our other results are not affected by publication bias (for example, non-publication of unfavorable or null-effect results). In government reports, ISP did not have any effect on recidivism on average (N = 13, OR = 1.01, $p \leq$ .949), but among the other unpublished papers we found, the odds of recidivism were significantly reduced by ISP programs (N = 9, OR = .70, $p \leq$ .002), while in academic articles there was a marginally significant increase (N = 16, OR = 1.20, $p \leq$ .085). If we had only examined the published literature we might have deduced that ISP does not work, and while these results are not exactly promising they do not lend themselves to such a drastic conclusion. We could not reliably assess different study settings because almost all of the studies were conducted in the U.S., and the five that

were not were all conducted in the U.K. at the same time by the same evaluators, and reported in the same paper.

We see similar results for technical violations, although the smaller number of studies in this category affected the analyses we could run (Table 1.4). We combined government and other unpublished reports to compare them with published academic articles. Although the between-group difference was non-significant ($Q_B$ = 2.31, $p \leq$ .129), there was a large, statistically significant increase in technical violations for ISP participants in published studies, compared to a slight increase in unpublished studies (published: N = 8, OR = 2.08, $p \leq .007$; unpublished: N = 8, OR = 1.19, $p \leq .472$). We only had technical violation data for studies conducted in the 1980s and 1990s, and although studies from both decades had increased violations among ISP participants, the increase was larger and significant in the 1980s (N = 11, OR = 1.81, $p \leq .023$; 1990s: N = 4, OR = 1.22, $p \leq .631$). The between-group difference is non-significant ($Q_B$ = .63, $p \leq$ .429). Although the number of studies in the 1990s is very small, limiting the conclusions we can draw from these results, it may be the case that 1980s studies showed more technical violations because control/surveillance was the prevailing supervision philosophy in that era.

We examined the moderating effects of a limited set of sample characteristics on ISP outcomes. We found no effect of age on the relationship between supervision intensity and recidivism ($Q_B$ = .07, $p \leq .794$; juveniles: N = 16, mean OR = .97, $p \leq .778$; youth and adults: N = 20, OR = .93, $p \leq .465$). We could not assess the effects of age on technical violations because too few studies reported this outcome for juveniles. We also found no effect for gender on either recidivism or violations (recidivism: $Q_B$ = 1.59, $p \leq$

.450; technical violations: $Q_B < .01$, $p \leq .987$). The odds of recidivism were 50 per cent greater among ISP participants in the all-female studies, but this was based on only two studies and was non-significant (mean OR = 1.50, $p \leq .299$).

The additional program characteristics we examined were program type (enhanced probation/parole or prison diversion); the population (probationers, parolees, or both) and offense types (any offenses, or a specialized caseload such as drugs) targeted by the program, and some more specific effects of changes in intensity. Again, small cell frequencies limit our ability to draw any firm conclusions from these analyses. We found that the odds of recidivism were lower in prison diversion programs than probation enhancement programs (there were no technical violation data for prison diversion programs), but with only two studies aiming to divert offenders from prison it is not possible to say that this reduction was due to intensive supervision. Not all of the studies we included accounted for time at risk in their reporting of outcomes, so it is possible that prison diversion programs appear more successful because participants had a higher likelihood of being reincarcerated and thus incapacitated ($Q_B = 1.44$, $p \leq .231$; enhancement: N = 36, OR = .99, $p \leq .952$; diversion: N = 2, OR = .69, $p \leq .207$). Few substantial differences were observed between target populations and offense types either, beyond what might be expected given the nature of the categories. The odds of recidivism and violations were slightly higher among parolees compared to probationers and mixed caseloads, and among specialized caseloads compared to mixed offense types. Parolees may be more likely to reoffend than offenders who were sentenced to probation; and offenders who have been singled out for offense-specific caseloads (e.g., specialized supervision for drug offenders) have already been designated as posing a greater risk of

failing on a certain type of crime. Since many of the specialized caseloads were drugs-focused, offenders may have been monitored more closely for substance abuse and violated for failing drug tests. This could explain the significant difference in technical violations between specialized and general caseloads ($Q_B = 3.95$, $p \leq .047$; specialized: N = 6, OR = 2.44, $p \leq .002$; general: N = 10, OR = 1.19, $p \leq .430$).

The final set of moderator variables we examined relate to the type and 'dosage' of supervision intensity, which are a key part of our inquiry. We attempted to examine how the programs we studied attempted to increase intensity, and whether it is possible to draw any conclusions about what magnitude of increased intensity is necessary to really affect outcomes. Too few studies reported enough information to assess the second question, so we examine this separately using only the studies that did report target caseload sizes and numbers of contacts and drug tests.

Tables 1.3 and 1.4 show the moderator analyses for the effects of intensity type on recidivism and technical violations respectively. On recidivism, we see little effect of any of the three main types of intensity increases (reduced caseloads, increased contacts, and mandatory drug tests), either within treatment groups or compared to controls. For technical violations, the raw effect sizes are considerably different between studies reporting and not reporting specific increases in intensity, but not in the direction expected. Studies that did not report reduced caseloads or increased contact reported higher odds of violations in the treatment groups. For increased drug testing, we also see significantly higher odds of violations in the treatment groups where drug tests were not a component of the program. These results may need no further explanation than the fact that there are very few studies involved, so they may be highly skewed. In both analyses,

it should be noted that those programs not reporting a particular type of increase (e.g., caseload size) will still have increased intensity in other ways. Furthermore, studies assessing reduced caseloads will most likely also have involved more contacts simply because the probation officers had more time with their clients, even if increased contacts were not a stated component of the program. Thus, we cannot say much about the effects of specific changes in intensity with these results.

We examined whether studies reporting increases in one or more components of intensity reported actual or planned ratios (e.g., planned caseload size in the treatment group compared to average caseloads on regular probation). Due to the amount of variability and the fact that planned amounts of supervision did not always translate into practice, we simply dichotomized dosage variables according to whether or not the dosage was changed by more or less than 100 per cent in the treatment group compared to the control group. Tables 1.5 and 1.6 show the results of this investigation for recidivism and technical violations respectively. 'High' dosage programs are those in which the number of contacts or drug tests was more than 100 per cent greater than control group standards. We were unable to include caseload size in the analysis because only one study reported a planned caseload difference of less than 100 per cent.

Table 1.5 shows that neither high nor low dosages of contacts or drug tests appeared to greatly affect recidivism compared to regular probation. The results are similar for technical violations (Table 1.6). While we observe some large effects for drug test dosage and violations, the number of studies is very small. The finding for contact frequency and technical violations is unsurprising: the odds of failure for probationers in high contact ISPs compared to controls were higher than for those in low contact ISPs

compared to controls; however, there is no statistical difference between these groups ($Q_B$ = .04, $p \leq .835$; high: N = 11, OR = 1.67, $p \leq .06$; low: N = 4, OR = 1.49, $p \leq .407$).

## Conclusion

The aim of this study was to systematically review and synthesize the most rigorous available evidence on the effects of changing probation intensity on probationers' criminal conduct. We identified and coded 239 potential evaluations of increased intensity (intensive supervision probation or ISP), and assessed a total of 47 individual treatment-comparison contrasts – 38 randomized trials and 9 quasi-experiments – as eligible for inclusion in a meta-analysis.

Despite our comprehensive approach to identifying a body of research spanning over fifty years, we were unable to find any evidence to contradict prior reports that suggest ISP 'does not work' (e.g., Petersilia & Turner, 1993; Sherman et al., 1997; MacKenzie, 2006b). Although in general the experience of ISP does not appear to substantially *increase* reoffending among participants, they do not appear to fare any better than their counterparts on regular probation for the extra supervision they receive. In addition, and again consistent with the prior research, we found that ISP was associated with an overall increase in technical violations across the studies we reviewed. In our examination of potential moderator variables, we found no policy-relevant program features that indicated any circumstances under which ISP may be more successful. Our only significant finding that is not affected by small cell frequencies or substantial statistical uncertainty is that ISP appears more successful in programs written

up in unpublished reports than those in academic articles. This only enables us to say that our findings account for potential publication bias.

We do not believe that our results should be taken as conclusive evidence that intensive probation supervision is a failed intervention. There is clearly a great deal of variation in the types of programs studied that we could not capture with our limited set of moderator variables. While the common components we were able to identify do not appear to have any great effect on recidivism, there are many more that we could not compare. Although we were able to include more rigorous studies in our meta-analysis than many other systematic reviews in the social sciences, there are so many variations on the ISP theme that many more studies including the same components would be needed to draw any meaningful conclusions. ISP effects may also be particularly sensitive to implementation issues. All experimental programs may suffer from inadequate or problematic implementation in the field, but ISP may be especially susceptible because it involves changing the practices of an extant agency. Even research funding (which may be limited compared with the general operating costs of a probation agency) might not be sufficient to enable officers to reach caseload or contact targets, or know what to do when they get there. Indeed, some studies we reviewed did include the actual as well as intended caseload sizes and contact frequencies. They usually showed that caseloads were generally slightly larger and contacts less frequent in reality than the numbers called for in the evaluation design, although it should be added that the planned *ratios* of intensity between the treatment and control groups were often similar because agencies also found it difficult to meet the contact and caseload size standards they set for regular probation.

Several reviews have indicated that some ISP programs, particularly those involving a treatment component, show more favorable results (e.g., Aos, Miller, & Drake, 2006; MacKenzie, 2006b). Our own analysis did not support this finding, but the individual point estimates in our forest graphs clearly show that some ISP programs were very successful. This suggests that further exploration of programs and designs may be needed to fully understand the potential benefits of ISP.

Two recent successful ISP studies, the Maryland Proactive Community Supervision program (Taxman, Yancey, & Bilanin, 2006), and Hawaii's HOPE (Hawken & Kleiman, 2009), appear very different on the surface, but share certain elements that might be the key to understanding the 'optimal' approach to intensive supervision. The Maryland program focuses on service brokerage and individual case planning by the probation officers: in the broad philosophical scheme, it is more treatment-based, although it includes surveillance and enforcement components too. In contrast, the Hawaii program is much more enforcement- and deterrence-focused. Probationers are notified daily whether or not they have been selected for random drug testing, and failed drug tests are met with swiftly delivered sanctions: a brief period of imprisonment (usually a weekend), the duration of which increases in response to further violations. However, multiple violators are also directed to residential drug treatment. Underlying both programs is a behavioral management model, which is articulated in the Maryland program and implicit in Hawaii's approach. The principles of behavioral management include incentive/sanction schemes; a focus on criminogenic factors that leads to tailored, rather than mandated, treatment and services; and offender accountability through behavioral contracts. In Hawaii, for example, offenders who were sent to jail for

violations first went before a judge who reminded them of their responsibilities on probation in a manner that reinforced the desire of the whole criminal justice system to see offenders succeed rather than fail. Criminal justice programming that emphasizes a combination of treatment and accountability, and incentive/sanction-based models, has shown promise in other settings, such as drug courts (Marlowe, 2003; MacKenzie, 2006b). Given the extent to which ISP programs have been directed toward drug-involved offenders, it may be particularly informative to draw comparisons with drug offending research. Gendreau, Goggin, and Fulton (2001) also lend more support to the general contention that such a balanced approach to offender supervision, emphasizing relationship-building, incentives, and adherence to the PEI, is effective.

One striking element of both programs just discussed is the lack of any emphasis of specific alterations to program intensity. Whereas prior experiments have searched for optimal caseload sizes or mandated certain numbers of contacts or drug tests, these more recent studies seem to focus on the *content* of supervision and responses to violations. This is in contrast to the surveillance-based programs in which (at their most extreme), probation officers "go out in the field actively looking for violations" (Pearson, 1988). This begs the larger question: what do probation officers actually do when required to supervise offenders more closely? The present study has not brought us much closer to answering that question than when Clear and Hardyman (1990) considered it twenty years ago. A reduction in caseload size is not automatically accompanied by a guarantee that officers will actually be able to spend more time with their clients. Fewer cases do not necessarily equate to more *intensive* treatment of probationers. Indeed, for all the

research we uncovered on ISP, there is little qualitative inquiry into the nature of the probation officer-client interactions.

Bonta et al. (2008) are among the few researchers who attempted to get inside what they call the "black box" of supervision. They discovered that the officers they studied spent "too much" time on enforcement rather than service delivery; did not account for the PEI or criminogenic need in their supervision strategies; and were not equipped with the necessary tools to effect behavioral change. The elements Bonta et al. found lacking seem to match the characteristics we suggest may be the key to successful ISP programs. Enforcement may well be easier, if not less time-consuming, than the service-oriented elements of supervision: identifying non-compliance may be more clear-cut than identifying individualized needs and tailoring case plans accordingly.

We therefore call for more research into probation supervision in general, ideally combining quantitative and qualitative methods, to uncover exactly what characteristics and processes influence successful outcomes. This review has shown that 'more' does not equal 'better' – in most cases intensive probation does not improve recidivism, and may even increase technical violations. However, we cannot yet conclusively say whether more of the 'right stuff' is better, and to do that we first need a much greater understanding of what the 'right stuff' is.

## Notes

[1] In the subsequent narrative we do not differentiate between probation and parole. 'Probation' is used as shorthand for both unless otherwise stated. In many agencies there is little difference in supervision practices for both probation and parole clients.

[2] One U.S. estimate indicates that over 40 per cent of probationers and more than half of parolees do not complete their supervision terms successfully, and that parole violators account for nearly 35 per cent of admissions to state prisons (Solomon et al., 2008).

[3] Worrall et al. conducted a cross-sectional study that indicated an increase in property crime rates across the state of California as that state's average probation caseloads increased.

[4] There are multiple ways in which supervision can be intensified, particularly in the light of advances in information technology. Electronic monitoring, satellite tracking (GPS), and voice verification systems are popular methods for 'passively' managing offender caseloads. Because such a wide range of automated systems are available, some of which have been the focus of systematic reviews in their own right (e.g., Renzema & Mayo-Wilson, 2005, on electronic monitoring), we do not include evaluations that focus solely on passive monitoring technology. However, many intensive supervision programs use technology as part of a range of surveillance measures implemented alongside direct contact with probation officers, and these studies will be considered if the monitoring technology is not the only difference in intensity between treatment and comparison cases.

[5] Grey literature refers to studies that are not commercially published or available through traditional sources, such as technical reports and dissertations. Failure to allow for the identification of grey literature in systematic searches can lead to publication bias, which occurs when the published or otherwise readily available literature is not representative of all studies. This is a real possibility: for example, some authors and journal editors may be more inclined to submit or accept statistically significant findings, whereas studies that show no discernible effect may be written up for funding agencies but never published in peer-reviewed academic journals (Rothstein & Hopewell, 2009).

[6] Important journals include: *British Journal of Criminology; Crime & Delinquency; Crime & Justice; Criminology; Criminology & Public Policy; Federal Probation; Journal of Criminal Justice; Journal of Experimental Criminology; Journal of Offender Rehabilitation; Journal of Quantitative Criminology; Journal of Research in Crime & Delinquency; Justice Quarterly; Probation Journal.*

[7] http://www.zotero.org.

[8] The odds of the event occurring are given by $p/(1-p)$ (the probability of the event occurring divided by the probability of the event not occurring).

[9] Note that although we present our results as odds ratios, analyses are actually performed on the natural log of the OR, which is centered around 0 rather than 1 and has a standard error that is easier to calculate (Lipsey & Wilson, 2001, p. 54).

[10] Current thinking in meta-analytic methods states that the random effects model should always be used. Previously, fixed effects models were considered acceptable when the *Q*-statistic from the main effects analysis was non-significant, indicating homogeneity between effect sizes. However, the assumptions of the random effects model are probably more defensible for many criminological applications. The random effects model also converges on the fixed effects as the distribution becomes homogeneous (see Appendix C) (Lipsey & Wilson, 2001, p. 120; David B. Wilson, personal communication, December 2009). Most of our analyses displayed substantial heterogeneity. We obtained both fixed and random effects estimates for the mean effect sizes and did not observe much difference between the two. The random effects estimates were generally more conservative (results not shown).

[11] The mixed effects analog to the ANOVA has a lower risk of Type I error than the fixed effects, which assumes that differences are systematic and thus does not perform well when the distribution is very heterogeneous. We employ a method of moments estimator of the random effects variance component. We also use this estimator for the main effects model (Appendix C; Lipsey & Wilson, 2001, pp. 124-5; Wilson, 2010, pp. 195-8). This is the least biased estimator available in the current version of the STATA macro, but it is less efficient than the alternative maximum likelihood approach. However, it is well-suited

to most applications, including sets of studies with relatively small sample sizes (Wilson, 2010, p. 196). We did not observe substantial differences in the results of our ANOVA tests depending on whether method of moments or maximum likelihood estimators were used (results not shown).

[12] Note that the excluded studies listed in Appendix D are only those that were coded in full and found to be ineligible. Many studies gave sufficient information about the research design or nature of the comparison group in their titles and abstracts and as such did not make it past the initial screening stage.

[13] We know of two studies on decreased probation intensity (one RCT and one rigorous quasi-experiment) that have not yet been published. We do not include them in the review because they are not comparable to the other evaluations, and are based on the same sample so cannot be compared to each other (Barnes et al., forthcoming; Berk et al., forthcoming).

[14] Although we identify individual studies according to the date of the report, the "research timeframe" measure in Table 1.1 reflects the actual year in which the research was conducted. If a study spanned two decades it is classified according to the year in which the study period began.

[15] One was a program specifically designed for women (Guydish et al., 2008), and the other was an evaluation in which results were reported separately for male and female offenders in one of the study sites (Folkard, Smith, & Smith, 1976). Data on gender composition were missing in 12 of the 47 studies, but we expect that they also reflect mixed, mostly male caseloads typical in any probation agency.

[16] Effect sizes excluded from this analysis were either continuous measures with insufficient data reported to calculate an effect size, or based on specific offense types that were either not available in enough studies, or were not theoretically relevant. For example, we saw no basis for reporting outcomes for property offenses separately from all offense types as we had no reason to believe that ISP would affect property offenses differently. Drug offense measures were the exception. Because many of the ISP programs we examined were targeted specifically at drug-involved offenders or included increased drug testing among the control measures employed, we examined this outcome separately. Eleven randomized trials reported separate data for arrests for drug-related crimes. Twenty-eight studies (21 RCTs and 7 quasi experiments) reported arrest data for any offense type; 23 studies reported technical violations (16/7), and 32 studies reported convictions (27/5).

[17] OR = .97 ($p \leq .645$). $Q = 72.80$ ($p < .001$).

## Tables

Table 1.1: Characteristics of Included Studies

| | Proportion of studies with characteristic | | |
|---|---|---|---|
| | **RCTs (N=38)** | **Quasi-Experiments (N=9)** | **Total (N=47)** |
| **Study Characteristics** | | | |
| Type of publication | | | |
| Academic publication | 42.1 | 44.4 | 42.6 |
| Government/technical report | 34.2 | 44.4 | 36.2 |
| Other unpublished | 23.7 | 11.1 | 21.3 |
| Research timeframe | | | |
| 1970s and earlier | 15.8 | 0.0 | 12.8 |
| 1980s | 34.2 | 11.1 | 29.8 |
| 1990s | 47.4 | 77.8 | 53.2 |
| 2000s | 2.6 | 11.1 | 4.3 |
| Study conducted in USA | 86.8 | 100.0 | 89.4 |
| **Program Characteristics** | | | |
| Program type | | | |
| Enhanced probation/parole | 94.7 | 88.9 | 93.6 |
| Prison diversion | 5.3 | 11.1 | 6.4 |
| Program involves caseload size reduction | 89.5 | 88.9 | 89.4 |
| Program involves contact frequency increase | 55.3 | 44.4 | 53.2 |
| Program involves drug test requirement increase | 29.0 | 11.1 | 25.5 |
| Supervision philosophy | | | |
| Control/surveillance | 39.5 | 11.1 | 34.0 |
| Treatment | 15.8 | 11.1 | 14.9 |
| Hybrid | 44.7 | 77.8 | 51.1 |
| Control group received regular supervision | 92.1 | 100.0 | 93.6 |
| Target population | | | |
| All probationers | 71.1 | 11.1 | 59.6 |
| All parolees | 13.2 | 55.6 | 21.3 |
| Mixed | 15.8 | 33.3 | 19.2 |
| Target offending type[a] | | | |
| Any offenses | 76.3 | 77.8 | 76.6 |
| Specialized caseloads | 23.7 | 11.1 | 21.3 |

Continued

Some sections do not add up to 100% due to rounding or missing data.

[a] Data not reported in 1 study.

Continued from previous page

| | Proportion of studies with characteristic | | |
|---|---|---|---|
| | **Experiments (N=38)** | **Quasi-Experiments (N=9)** | **Total (N=47)** |
| **Sample Characteristics** | | | |
| Age of sample[b] | | | |
| Juveniles | 42.1 | 44.4 | 42.6 |
| Youth and adults | 52.6 | 55.6 | 53.2 |
| Gender[c] | | | |
| All males | 29.0 | 44.4 | 31.9 |
| All females | 5.3 | 0.0 | 4.3 |
| Mixed | 34.2 | 55.6 | 38.3 |
| Offender risk level | | | |
| High/mostly high risk | 71.1 | 77.8 | 72.3 |
| Mixed risk levels/no assessment | 29.0 | 22.2 | 27.7 |
| **Outcome Characteristics** | | | |
| Arrest outcomes reported | 55.3 | 77.8 | 59.6 |
| Technical violation outcomes reported | 42.1 | 77.8 | 48.9 |
| Conviction outcomes reported | 71.1 | 55.6 | 68.1 |
| Other outcomes available | 39.5 | 0.0 | 31.9 |
| Length of follow-up period | | | |
| 12 months or less | 60.5 | 88.9 | 66.0 |
| More than 12 months | 39.5 | 11.1 | 34.0 |

Some sections do not add up to 100% due to rounding or missing data.

[b] Data not reported in 2 studies.

[c] Data not reported in 12 studies.

Table 1.2: Overall Mean Effect Sizes for Crime Outcomes

| | **N** | **Mean OR** | **95% C.I.** | | ***Q*** |
|---|---|---|---|---|---|
| | | | **Lower** | **Upper** | |
| **RCTs** | | | | | |
| Arrests | 21 | .93 | .74 | 1.17 | 61.55*** |
| Convictions | 27 | .98 | .85 | 1.13 | 20.38 |
| Technical violations | 16 | 1.54 | .99 | 2.39 | 120.11*** |
| Drug arrests | 11 | 1.18 | .86 | 1.61 | 11.50 |
| **Quasi-experiments** | | | | | |
| Arrests | 7 | .83 | .66 | 1.04 | 7.55 |
| Convictions | 5 | .60 | .33 | 1.11 | 22.74*** |
| Technical violations | 7 | 1.29 | .86 | 1.94 | 18.91** |

** $p \leq .01$; *** $p \leq .001$.

Table 1.3: Moderator Variable Effects (Arrest/Conviction, RCTs)

| | N | $Q_B$ | Mean OR | 95% C.I. | |
|---|---|---|---|---|---|
| | | | | Lower | Upper |
| **Study Characteristics** | | | | | |
| Publication type | | | | | |
| Academic publication | 16 | | 1.20 | .98 | 1.48 |
| Government/technical report | 13 | 11.86** | 1.01 | .78 | 1.30 |
| Other unpublished | 9 | | .70** | .55 | .88 |
| Research timeframe | | | | | |
| 1970s and earlier | 6 | | 1.09 | .77 | 1.55 |
| 1980s | 13 | .87 | 1.00 | .76 | 1.33 |
| 1990s and 2000s [a] | 19 | | .90 | .72 | 1.13 |
| Location of study | | | | | |
| USA | 33 | 1.10 | .93 | .80 | 1.10 |
| Non-USA | 5 | | 1.17 | .79 | 1.74 |
| **Program Characteristics** | | | | | |
| Program type | | | | | |
| Probation enhancement | 36 | 1.44 | .99 | .84 | 1.17 |
| Prison diversion | 2 | | .69 | .39 | 1.23 |
| Reduced caseload component | | | | | |
| Yes | 34 | 3.42 | 1.01 | .87 | 1.17 |
| No | 4 | | .65 | .41 | 1.01 |
| Increased contact component | | | | | |
| Yes | 21 | .49 | 1.01 | .83 | 1.22 |
| No | 17 | | .90 | .71 | 1.15 |
| Increased drug testing component | | | | | |
| Yes | 11 | 2.76 | .80 | .62 | 1.04 |
| No | 27 | | 1.05 | .88 | 1.24 |
| Prevailing supervision philosophy | | | | | |
| Control/surveillance | 15 | | .92 | .73 | 1.16 |
| Treatment | 6 | 1.73 | 1.20 | .84 | 1.72 |
| Hybrid | 17 | | .93 | .73 | 1.17 |
| Target population | | | | | |
| Probationers | 27 | | .92 | .76 | 1.12 |
| Parolees | 5 | .65 | 1.07 | .75 | 1.51 |
| Mixed | 6 | | 1.04 | .69 | 1.56 |
| Target offense types | | | | | |
| Any offending | 29 | .92 | .93 | .78 | 1.10 |
| Specialized caseloads | 9 | | 1.10 | .81 | 1.49 |

Continued

Mixed effects (method of moments) mean odds ratios

* $p \leq .05$; ** $p \leq .01$.

[a] Categories combined: only one study in 2000s, which had large effect favoring treatment group.

Continued from previous page

| | N | $Q_B$ | Mean OR | 95% C.I. | |
|---|---|---|---|---|---|
| | | | | Lower | Upper |
| **Sample Characteristics** | | | | | |
| Age | | | | | |
| Juveniles | 16 | .07 | .97 | .76 | 1.23 |
| Youth/adults | 20 | | .93 | .75 | 1.14 |
| Gender | | | | | |
| All males | 11 | | .99 | .75 | 1.30 |
| All females | 2 | 1.59 | 1.50 | .70 | 3.21 |
| Mixed | 13 | | .89 | .06 | 1.17 |
| Risk level | | | | | |
| High/mostly high risk | 11 | < .01 | .96 | .73 | 1.27 |
| Mixed risk levels | 27 | | .97 | .80 | 1.17 |

Mixed effects (method of moments) mean odds ratios.

Table 1.4: Moderator Variable Effects (Technical Violations, RCTs)

| | N | $Q_B$ | Mean OR | 95% C.I. | |
|---|---|---|---|---|---|
| | | | | Lower | Upper |
| **Study Characteristics** | | | | | |
| Publication Type | | | | | |
| Academic publication | 8 | 2.31 | 2.08** | 1.22 | 3.54 |
| Government/other unpublished report | 8 | | 1.19 | .74 | 1.94 |
| Research timeframe[a] | | | | | |
| 1980s | 11 | .63 | 1.81* | 1.09 | 3.03 |
| 1990s | 4 | | 1.22 | .54 | 2.80 |
| **Program Characteristics** | | | | | |
| Reduced caseload component | | | | | |
| Yes | 14 | .17 | 1.49 | .93 | 2.38 |
| No | 2 | | 2.04 | .48 | 8.61 |
| Increased contact component | | | | | |
| Yes | 13 | .41 | 1.44 | .89 | 2.34 |
| No | 3 | | 2.16 | .70 | 6.67 |
| Increased drug testing component | | | | | |
| Yes | 9 | 1.37 | 1.19 | .62 | 2.28 |
| No | 7 | | 2.13* | 1.02 | 4.45 |
| Prevailing supervision philosophy[b] | | | | | |
| Control/surveillance | 13 | .01 | 1.55 | .93 | 2.58 |
| Hybrid | 3 | | 1.46 | .51 | 4.24 |
| Target population | | | | | |
| Probationers | 7 | .31 | 1.66 | .77 | 3.55 |
| Parolees | 4 | | 1.72 | .67 | 4.40 |
| Mixed | 5 | | 1.24 | .50 | 3.08 |
| Target offense types | | | | | |
| Any offending | 10 | 3.95* | 1.19 | .78 | 1.82 |
| Specialized caseloads | 6 | | 2.44** | 1.38 | 4.32 |
| **Sample Characteristics** | | | | | |
| Gender | | | | | |
| All males | 5 | < .01 | 1.54 | .68 | 3.52 |
| Mixed | 11 | | 1.53 | .85 | 1.41 |
| Risk level | | | | | |
| High/mostly high risk | 10 | .04 | 1.59 | .93 | 1.69 |
| Mixed risk levels | 6 | | 1.45 | .73 | 2.89 |

Mixed effects (method of moments) mean odds ratios.

* $p \leq .05$; ** $p \leq .01$.

[a] No observations for 2000s. The single observation for 1970s and earlier (a study conducted in the 1950s) was dropped rather than being combined into another category due to potentially excessive influence from its large overall weight.

[b] No observations for 'Treatment.'

Table 1.5: Effect of Intensity Variation (Arrest/Conviction, RCTs)

| | N | $Q_B$ | Mean OR | 95% C.I. | |
|---|---|---|---|---|---|
| | | | | Lower | Upper |
| **Contact frequency difference** | | | | | |
| High | 13 | .42 | 1.08 | .87 | 1.35 |
| Low | 4 | | .91 | .56 | 1.47 |
| **Drug test frequency difference** | | | | | |
| High | 9 | .30 | .81 | .52 | 1.27 |
| Low | 2 | | 1.28 | .26 | 6.26 |

Mixed effects (method of moments) mean odds ratios.
Caseload size variation not included because only one study had low caseload variation. That study indicated large effect in favor of control group.

Table 1.6: Effect of Intensity Variation (Technical Violations, RCTs)

| | N | $Q_B$ | Mean OR | 95% C.I. | |
|---|---|---|---|---|---|
| | | | | Lower | Upper |
| **Contact frequency difference** | | | | | |
| High | 11 | .04 | 1.67 | .98 | 2.83 |
| Low | 4 | | 1.49 | .58 | 3.80 |
| **Drug test frequency difference** | | | | | |
| High | 7 | 2.03 | 1.05 | .80 | 1.38 |
| Low | 2 | | 2.07 | .85 | 5.01 |

Mixed effects (method of moments) mean odds ratios.
Caseload size variation not included because only one study had low caseload variation. That study indicated large effect in favor of treatment group.

## Figures

Figure 1.1: Effect of Intensive Probation on Arrests (RCTs)

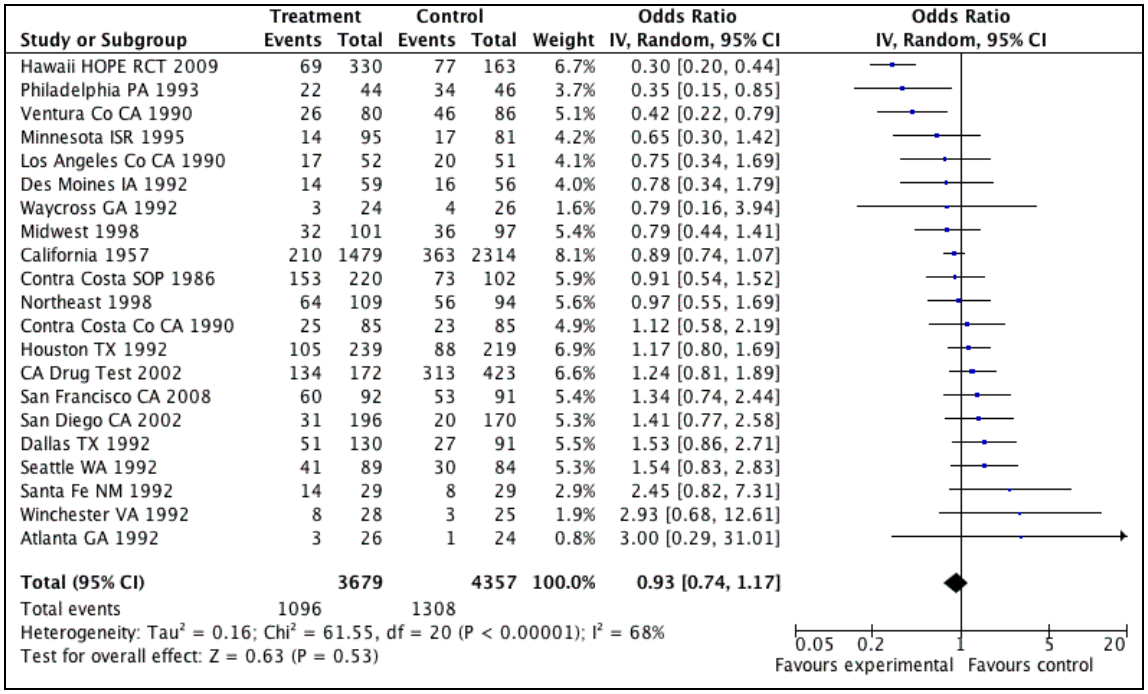



Figure 1.2: Effect of Intensive Probation on Arrests (Quasi-Experiments)

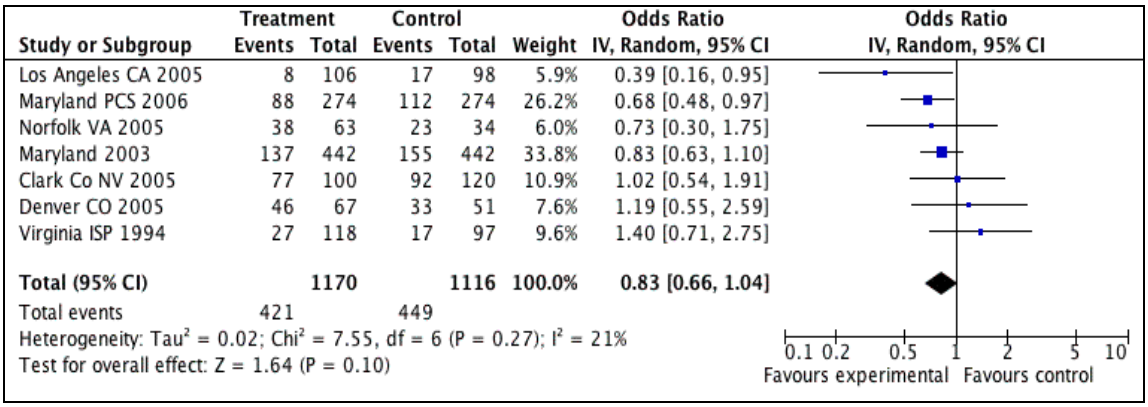
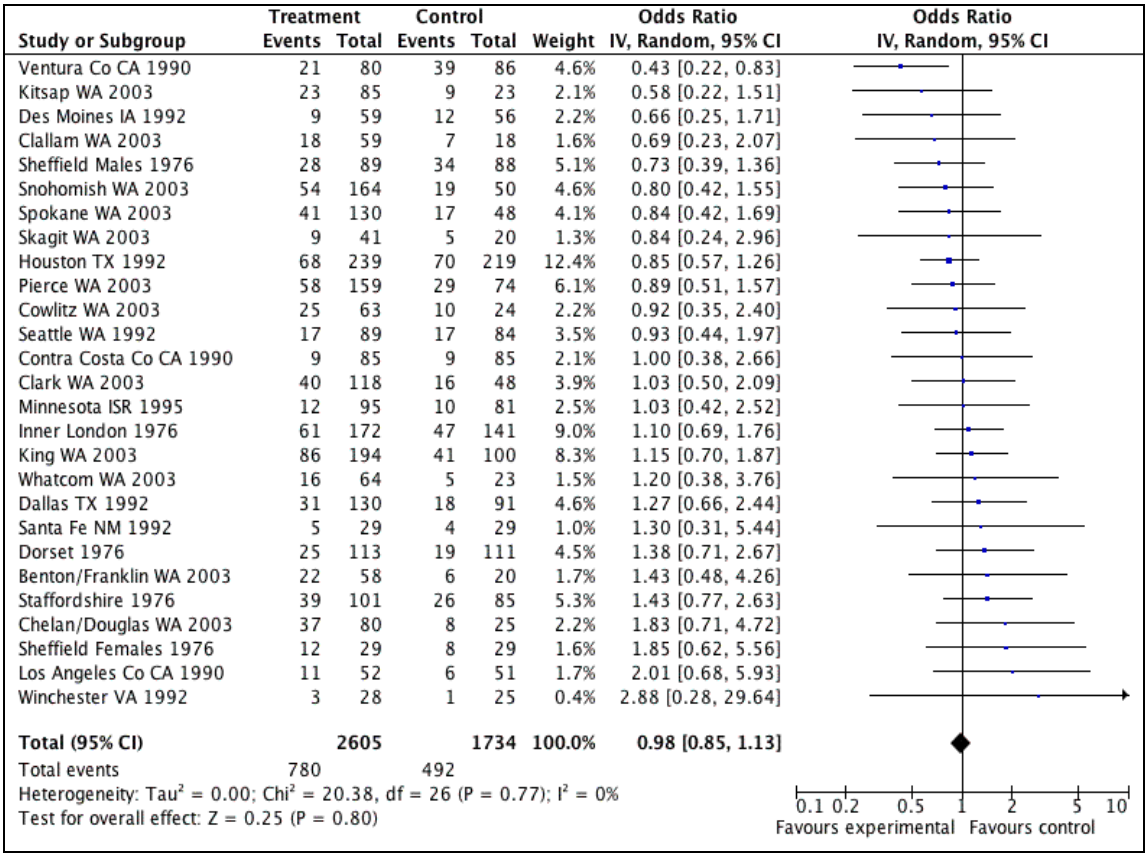


Figure 1.3: Effect of Intensive Probation on Convictions (RCTs)

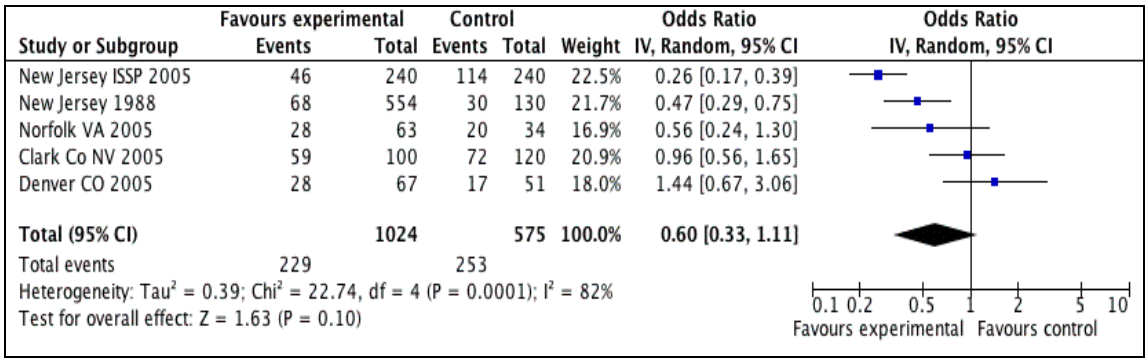


| Study or Subgroup | Treatment Events | Total | Control Events | Total | Weight | Odds Ratio IV, Random, 95% CI |
|---|---|---|---|---|---|---|
| Ventura Co CA 1990 | 21 | 80 | 39 | 86 | 4.6% | 0.43 [0.22, 0.83] |
| Kitsap WA 2003 | 23 | 85 | 9 | 23 | 2.1% | 0.58 [0.22, 1.51] |
| Des Moines IA 1992 | 9 | 59 | 12 | 56 | 2.2% | 0.66 [0.25, 1.71] |
| Clallam WA 2003 | 18 | 59 | 7 | 18 | 1.6% | 0.69 [0.23, 2.07] |
| Sheffield Males 1976 | 28 | 89 | 34 | 88 | 5.1% | 0.73 [0.39, 1.36] |
| Snohomish WA 2003 | 54 | 164 | 19 | 50 | 4.6% | 0.80 [0.42, 1.55] |
| Spokane WA 2003 | 41 | 130 | 17 | 48 | 4.1% | 0.84 [0.42, 1.69] |
| Skagit WA 2003 | 9 | 41 | 5 | 20 | 1.3% | 0.84 [0.24, 2.96] |
| Houston TX 1992 | 68 | 239 | 70 | 219 | 12.4% | 0.85 [0.57, 1.26] |
| Pierce WA 2003 | 58 | 159 | 29 | 74 | 6.1% | 0.89 [0.51, 1.57] |
| Cowlitz WA 2003 | 25 | 63 | 10 | 24 | 2.2% | 0.92 [0.35, 2.40] |
| Seattle WA 1992 | 17 | 89 | 17 | 84 | 3.5% | 0.93 [0.44, 1.97] |
| Contra Costa Co CA 1990 | 9 | 85 | 9 | 85 | 2.1% | 1.00 [0.38, 2.66] |
| Clark WA 2003 | 40 | 118 | 16 | 48 | 3.9% | 1.03 [0.50, 2.09] |
| Minnesota ISR 1995 | 12 | 95 | 10 | 81 | 2.5% | 1.03 [0.42, 2.52] |
| Inner London 1976 | 61 | 172 | 47 | 141 | 9.0% | 1.10 [0.69, 1.76] |
| King WA 2003 | 86 | 194 | 41 | 100 | 8.3% | 1.15 [0.70, 1.87] |
| Whatcom WA 2003 | 16 | 64 | 5 | 23 | 1.5% | 1.20 [0.38, 3.76] |
| Dallas TX 1992 | 31 | 130 | 18 | 91 | 4.6% | 1.27 [0.66, 2.44] |
| Santa Fe NM 1992 | 5 | 29 | 4 | 29 | 1.0% | 1.30 [0.31, 5.44] |
| Dorset 1976 | 25 | 113 | 19 | 111 | 4.5% | 1.38 [0.71, 2.67] |
| Benton/Franklin WA 2003 | 22 | 58 | 6 | 20 | 1.7% | 1.43 [0.48, 4.26] |
| Staffordshire 1976 | 39 | 101 | 26 | 85 | 5.3% | 1.43 [0.77, 2.63] |
| Chelan/Douglas WA 2003 | 37 | 80 | 8 | 25 | 2.2% | 1.83 [0.71, 4.72] |
| Sheffield Females 1976 | 12 | 29 | 8 | 29 | 1.6% | 1.85 [0.62, 5.56] |
| Los Angeles Co CA 1990 | 11 | 52 | 6 | 51 | 1.7% | 2.01 [0.68, 5.93] |
| Winchester VA 1992 | 3 | 28 | 1 | 25 | 0.4% | 2.88 [0.28, 29.64] |
| **Total (95% CI)** | | **2605** | | **1734** | **100.0%** | **0.98 [0.85, 1.13]** |
| Total events | 780 | | 492 | | | |

Heterogeneity: $Tau^2 = 0.00$; $Chi^2 = 20.38$, df = 26 (P = 0.77); $I^2 = 0\%$
Test for overall effect: Z = 0.25 (P = 0.80)



Figure 1.4: Effect of Intensive Probation on Convictions (Quasi-Experiments)



| Study or Subgroup | Favours experimental Events | Total | Control Events | Total | Weight | Odds Ratio IV, Random, 95% CI |
|---|---|---|---|---|---|---|
| New Jersey ISSP 2005 | 46 | 240 | 114 | 240 | 22.5% | 0.26 [0.17, 0.39] |
| New Jersey 1988 | 68 | 554 | 30 | 130 | 21.7% | 0.47 [0.29, 0.75] |
| Norfolk VA 2005 | 28 | 63 | 20 | 34 | 16.9% | 0.56 [0.24, 1.30] |
| Clark Co NV 2005 | 59 | 100 | 72 | 120 | 20.9% | 0.96 [0.56, 1.65] |
| Denver CO 2005 | 28 | 67 | 17 | 51 | 18.0% | 1.44 [0.67, 3.06] |
| **Total (95% CI)** | | **1024** | | **575** | **100.0%** | **0.60 [0.33, 1.11]** |
| Total events | 229 | | 253 | | | |

Heterogeneity: $Tau^2 = 0.39$; $Chi^2 = 22.74$, df = 4 (P = 0.0001); $I^2 = 82\%$
Test for overall effect: Z = 1.63 (P = 0.10)

Figure 1.5: Effect of Intensive Probation on Technical Violations (RCTs)

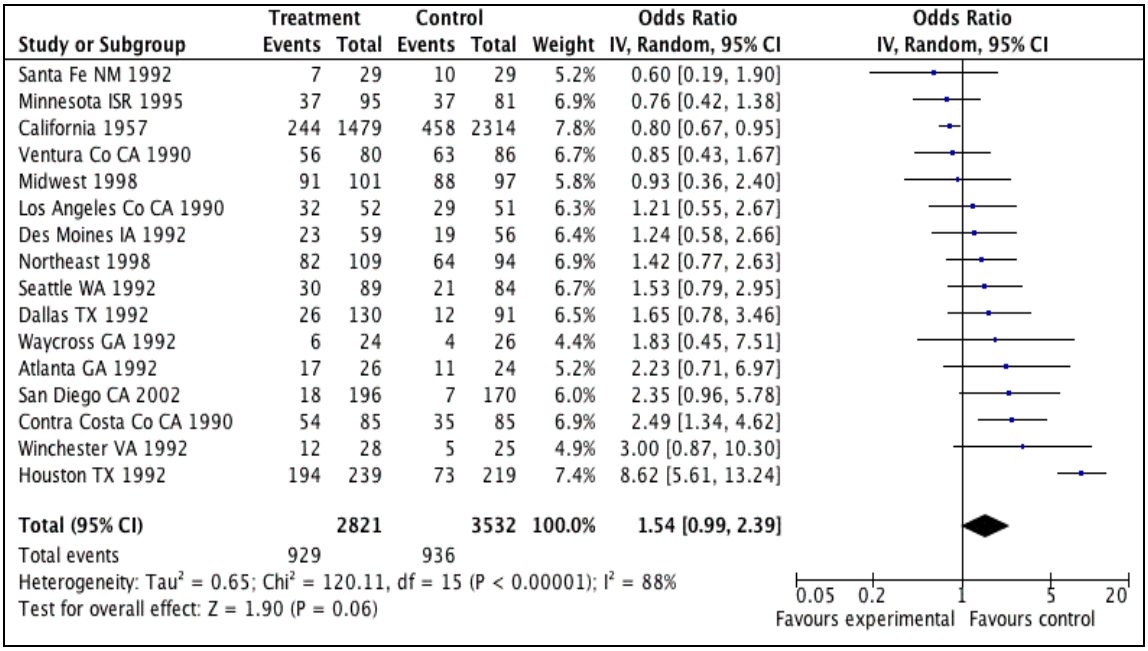



Figure 1.6: Effect of Intensive Probation on Technical Violations (Quasi-Experiments)

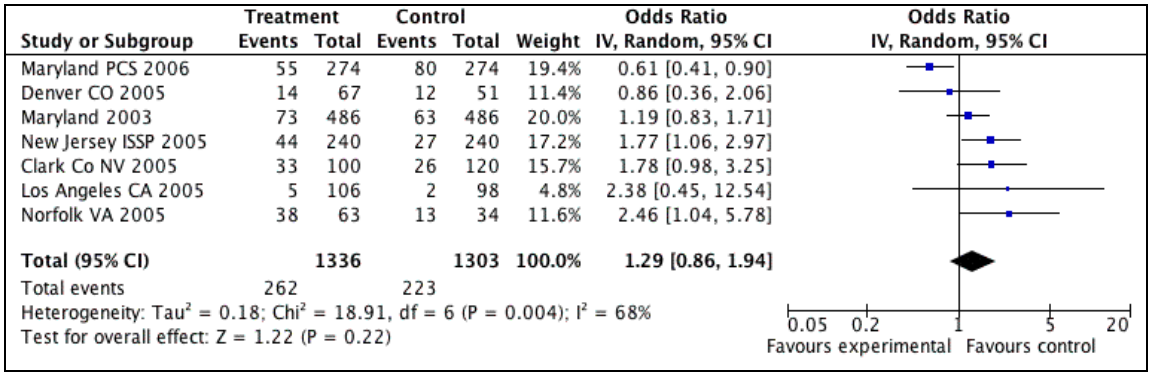
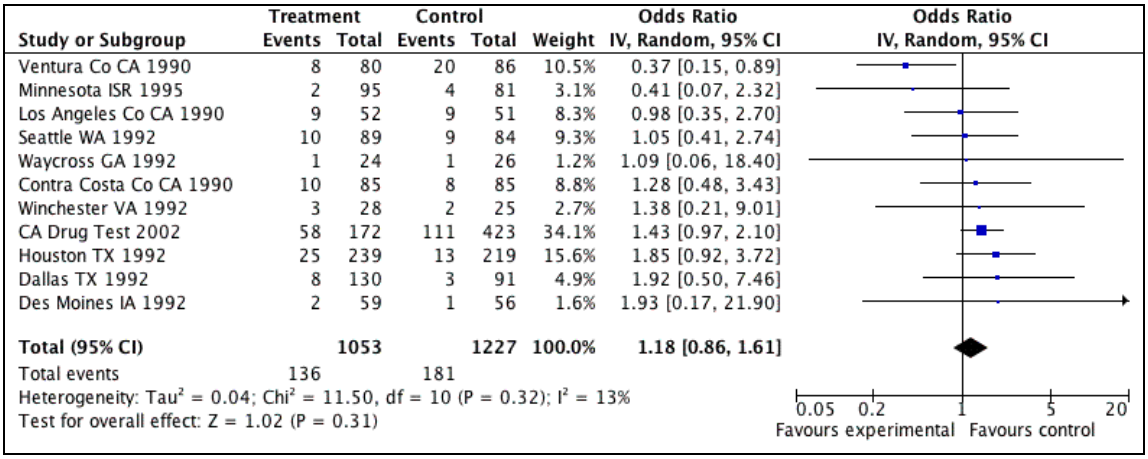


Figure 1.7: Effect of Intensive Probation on Drug Arrests (RCTs)



# CHAPTER 2. 'Low-Intensity' Probation: Is It a Viable Policy for Low-Risk Offenders?

## Introduction

What works in probation supervision? Researchers have been struggling with this question since the 1960s, yet the nature of supervision itself remains under-studied (Taxman, 2002; 2008a). Much of the research that is available has focused on initiatives to step up the power of probation as a punitive sanction. Such programs usually involve imposing strict, frequent reporting requirements on offenders, and providing supervision in small caseloads designed to increase surveillance as much as service provision. These intensive supervision probation (ISP) regimes have generally proven unsuccessful. For example, the field experiments with ISP in the 1980s showed null effects on crime and increases in technical violations (Petersilia & Turner, 1993). Despite more recent efforts to link supervision to treatment and services showing some promise (Taxman, 2008b), probation appears to suffer from an identity crisis. It was not originally intended for use with high-risk or serious offenders or as a punitive alternative to prison. Historically, probation was a 'second chance' for low-level, often first time offenders who posed little threat of serious harm (Clear & Braga, 1995). Yet, intensive probation has proved even more unsuccessful with low-risk offenders than their higher risk counterparts (Erwin, 1986; Hanley, 2006; Lowenkamp, Latessa, & Holsinger, 2006). Is this simply because ISP was not designed to target low-risk offenders, and the high levels of control and

scrutiny provoked defiant responses (Sherman, 1993) resulting in failure? Or do failure rates increase for low-risk probationers under more intensive supervision because the increase in the overall probability of detection is compared to a lower baseline?

This paper considers whether an experiment designed to test the premise that low-risk offenders can safely receive less intensive supervision than the 'standard' model of probation is sensitive to heterogeneity in the type of low-risk offender receiving treatment. As the willingness to use probation for offenders eligible for prison persists, more resources will be needed to serve their needs and adequately protect the public. If the low-risk, low-need offenders can be supervised more efficiently, probation officers will be freed up to focus on more serious cases. This sensitivity analysis mounts a comprehensive 'attack' on the low-intensity supervision model to ensure there is no way in which a policy reducing supervision for a section of the criminal population could increase the threat of harm to society. More broadly, we investigate the nature of low-risk offenders and their propensity to reoffend. The criminological literature has largely focused on the risk factors and characteristics of higher-level offenders, but these serious cases make up a much smaller proportion of the criminal population. This paper examines the characteristics of low-level offenders and their relationship with treatment effects. If the observed heterogeneity in the sample does not impact outcomes on low-intensity supervision, it seems reasonable to conclude that low-intensity supervision is a viable policy option compared to 'treatment as usual.'

## What Works in Probation Supervision?

Despite being one of the most widely-used criminal sanctions in the U.S.A., with one in forty-five adults on probation or parole (Glaze & Bonczar, 2009), much of the research on probation has failed to shed light on the characteristics of effective supervision practice. Correctional research has mainly focused on programming and treatments provided in addition to criminal justice sanctions like probation orders. In many probation agencies, standard practice is driven by resource constraints more than evidence-based strategies. With caseloads often averaging 150 to 200 offenders per probation officer in a given agency, supervision levels vary from weekly or twice-weekly meetings for the highest-risk or delinquent probationers to telephone reporting for those towards the end of their sentences. Some probationers simply mail in a card to confirm their current address (Petersilia & Turner, 1993).

Over the last thirty years, the use of probation has grown in response to (and as a result of) increasing prison populations. In this climate, intensive supervision probation (ISP) emerged as a supervision strategy that was deemed punitive enough to be used with offenders who would otherwise have been incarcerated, yet cheaper than the cost of keeping someone in prison. While few ISP programs are exactly alike, they usually involve a reduction in caseload size and increased frequency of contact with the probation officer, increased drug testing requirements and service provision or brokerage.

The first wave of research in the 1960s was generally described as the "search for the magic number" (Carter & Wilkins, 1976) because it involved experimentation with caseload size to find the optimal ratio of offenders to probation officers. The rationale was that smaller caseloads would allow probation officers to spend more time helping

their clients (Petersilia & Turner, 1990). However, these studies showed little difference in recidivism rates by caseload size, and technical violations increased for those offenders receiving more supervision (e.g., Neithercutt & Gottfredson, 1974; Banks et al., 1976; Carter & Wilkins, 1976). The caseload variation strategy was deemed unsuccessful.

The re-emergence of ISP in the 1980s was largely independent of prior interest in the topic. ISP at this time was at the forefront of an array of so-called 'intermediate sanctions' intended to provide a cost-effective but punitive alternative to incarceration. The focus of these programs was on crime deterrence through surveillance and control, which was effected through small caseloads and frequent face-to-face contacts and drug testing. A study of this model in Georgia, U.S.A., in the early 1980s showed early promise (Erwin, 1986), and in 1986 the Bureau of Justice Assistance (BJA) funded the RAND Corporation to conduct a multi-site randomized controlled trial of the 'new' ISP in comparison to regular probation or incarceration (Petersilia & Turner, 1993). Again, the results were disappointing: programs had little impact on recidivism, and technical violations and re-incarcerations increased due to increased surveillance.

Thus, intensive supervision probation appears to be an ineffective way to supervise offenders (Sherman et al., 1997; MacKenzie, 2006). The early ISP models were designed on the assumption that probation supervision in itself is 'good' for offenders, so more of it must be better. However, it appears that no theoretical basis for this assumption was ever articulated (Bennett, 1988). Clear and Hardyman (1990) suggest that the early experiments in caseload size variation failed because of a lack of knowledge about how probation supervision could serve the ultimate goal of offender treatment. Probation officers had more time to spend with clients, but did not know what

to do with it. Furthermore, the more intensive strategies were not always targeted at the highest-risk offenders in the greatest need of treatment, a policy now known to be crucial to successful correctional programming (Andrews, Bonta, & Hoge, 1990). There is no standard definition of 'high' or 'low' risk in the earlier studies, and the most serious offenders were often excluded. For example, in the RAND studies, participants were generally more serious offenders, but risk levels varied (Petersilia & Turner, 1993). Offenders convicted of homicide, robbery, and sex offenses were excluded from the studies for safety reasons, even though it is not unusual for such offenders to appear in probation and parole caseloads. In some cases, ISP was reasonably effective for high-risk offenders but backfired for low-risk offenders (e.g., Erwin, 1986; Hanley, 2006; Lowenkamp, Latessa, & Holsinger, 2006).

More recent research on probation supervision has accounted for some of these problems, making careful assessments of risk and targeting supervision appropriately. For high-risk offenders, intensive supervision has shown more positive results when it includes a greater emphasis on treatment and service brokerage by probation officers in addition to small caseloads and frequent contact (e.g., Latessa et al., 1998; Paparozzi & Gendreau, 2005; Aos, Miller, & Drake, 2006; Taxman, 2008b). Probation agencies have also begun experimenting with *reducing* the intensity of supervision for those offenders at the lowest risk of reoffending. New York City's probation department piloted an electronic kiosk reporting system for a considerable portion of its caseload that was considered to be low-risk. These offenders checked in regularly using an ATM-style device, and could request or be compelled to see a probation officer if adverse circumstances arose. Two-year rearrest rates for all crimes for low-risk probationers

declined from 31 per cent to 28 per cent after the kiosks were introduced, suggesting that it made little difference from traditional supervision for that population. The system also freed up probation officers to supervise high-risk offenders more intensively. Rearrest rates for high risk offenders subsequently decreased from 52 per cent to 47 per cent (Wilson, Naro, & Austin, 2007). Another recent study in Oregon used a similar model, assigning low-risk offenders to a "casebank" caseload where they received minimal face-to-face contact with their probation officer, again with the purpose of allowing high-risk offenders to receive intensive supervision (Johnson, Austin, & Davies, 2003). Although the analysis did not separate results for low- and high-risk offenders, a pre-post analysis of crime rates among probationers on community supervision in the county indicated that overall crime rates decreased after the implementation of risk-based supervision.

The recent attempts at implementing low-intensity probation supervision are important steps in developing effective supervision practices across the board. The latest research on intensive supervision has begun to unpack the relationship between surveillance, treatment, risk, and need. It suggests that intensive supervision may be an effective strategy for dealing with high-risk offenders. At the same time, probation departments remain chronically under-resourced. Caseloads are large and money to employ new probation officers to downsize high-risk caseloads is rarely available. Low-intensity supervision could be a vital resource-saving strategy. Allowing low-risk offenders to receive minimal supervision in a large caseload means that existing officers can be reallocated to concentrate on higher-risk clients who pose a greater public safety risk. However, the 'more is better' approach that guided prior research suggests that reducing supervision, even to low-risk offenders, could increase their reoffending. Prior

to the New York City and Oregon studies, large caseloads have always been portrayed as detrimental to crime prevention (e.g.,Worrall et al., 2004; Lemert, 1993).

## Theories of Probation

There are several logical theoretical mechanisms by which a policy of low-intensity supervision for low-risk probationers could in fact offer a safe way to allocate resources more efficiently. In order to understand how any probation practice might work, one must first consider the fundamental purpose of probation. That discussion has not featured prominently in the literature on community corrections. Probation today is usually recognized solely as a sanction. Indeed, the popularity of intensive probation is largely due to its place at the forefront of intermediate sanctions for punishing more serious offenders without sending them to prison. However, the roots of probation supervision lie in rehabilitation rather than retribution. John Augustus, who is credited with the invention of probation in 1841, intended it as a diversion from court allowing defendants to prove their desire to reform prior to trial, underpinned by the threat of criminal sanctions if they failed (Petersilia, 1997). He derived the term 'probation' from the Latin *probare*, meaning to prove or demonstrate. Petersilia (ibid.) notes a shift in the probation officer's role, starting in the mid-twentieth century, from social worker to the "eyes and ears of the local court" (p. 157). There remains a tension between the social work and surveillance/control philosophies. Thus, in considering the deterrent effect of sanctions, one must consider probation not only as a punishment designed to deter future crime, but as a 'second chance' to go straight and avoid harsher sanctions.

A low-intensity probation model involving minimal contact with a probation officer allows low-risk offenders who do not pose a threat of serious future harm to rebuild their lives relatively unencumbered by probation office visits and programming. From a punishment perspective, such a low-level sanction is commensurate with their offending and the risk they pose to society. From a rehabilitation perspective, it may be argued that low-risk offenders have less need for services and programs than their high-risk counterparts, and thus need less attention from probation officers. In these respects, the 'carrot' of low-intensity supervision, along with the 'stick' of being returned to increased supervision or jail for failure, may act as a deterrent to future offending. The deterrence literature lends some support to this idea. Studies of the perceived certainty and severity of punishment have indicated that individuals with greater experience of criminal offending perceive a lower risk of punishment than individuals with little or no experience (Paternoster et al., 1983; Nagin, 1998). If we assume that low-risk offenders, in general, have less extensive criminal careers than higher-risk offenders, we could make the argument that low-risk offenders might be more likely to be deterred by the threat of losing the relative 'freedom' of low-intensity supervision if they reoffend. Furthermore, one might argue that low-risk offenders will commit *more* crime if they receive more supervision than they need. The perception that the sentence is disproportionate to their risk level and thus unfair may weaken offenders' respect for the criminal justice system, leading to a defiant response expressed as an escalation of recidivism (Sherman, 1993). Thus, assignment to low-intensity supervision may help to ensure that low-risk offenders perceive the sanction as legitimate.

Barnes et al. (forthcoming) have suggested that being placed on probation supervision could increase the risk of reoffending for low-risk offenders through deviant peer contagion (DPC). Having antisocial and delinquent associates is one of the most important and consistently reported risk factors for crime (e.g., Andrews, 1989). One proposed mechanism by which association with delinquent peers increases the risk of crime is through DPC: contact between offenders or at-risk juveniles who come together in group-based interventions and programs that leads to reinforcement and support of delinquent values (Dishion, McCord, & Poulin, 1999; Dishion & Dodge, 2006). The fact that DPC (which is largely rooted in social psychology) has so many parallels in classic criminological theories of differential association and social learning, which focus on the relationship of crime with attachment to and behavior of antisocial peers (e.g., Sutherland, 1947; Akers, 1973; Agnew, 1991; Warr & Stafford, 1991), suggests that the concept could be extended to other environments in which offenders gather. Barnes et al. (forthcoming) propose that the probation department could create a similar dynamic. In the probation department they observed, offenders spent a great deal of time waiting in line together outside the office and talking to each other in waiting areas and elevators. Although the content of discussions between probationers in this environment has not been studied, it is reasonable to assume that at least part of the discussion focuses on the reason for their presence at the office and their opinions about the sanction.

The DPC literature suggests that lower-level delinquents are most susceptible to the influence of delinquent peers (Rosch, 2006; Lowenkamp, Latessa, & Holsinger, 2006). If DPC occurs in probation departments, it perhaps makes sense to focus first on limiting the exposure of low-risk offenders who do not currently pose a threat of serious

offending. This can be achieved through a low-intensity supervision model in which the need for low-risk offenders to attend the probation department is minimal. Of course, a caveat to using this reasoning to justify low-intensity supervision is that research on DPC and related criminological theories has largely focused on juveniles and it is not clear whether the same mechanisms operate for adults. More importantly, we can limit exposure to other probationers at the probation office but it is not possible to control offenders' access to delinquent networks in their home neighborhoods. However, informal social controls may also operate there that do not exist in the probation waiting room – for example, family or a job opportunity – that allow offenders to engage in pro-social activity without the encumbrance of regular probation visits. Lowenkamp and Latessa (2004) note that such pro-social networks play an important role in explaining why some offenders remain at low risk for future criminal behavior. Subjecting low-risk offenders to increased supervision may disrupt their positive social networks.

A further theoretical mechanism by which low-intensity probation may successfully operate is through the principles of effective intervention (PEI). The PEI were introduced by Andrews, Bonta, and Hoge (1990) in response to the 'nothing works' attitude to correctional treatment that persisted in the 1970s and 1980s. The PEI state that correctional treatment can in fact be effective when programs are designed to be responsive to offenders' specific risk and need levels (the risk-need-responsivity, or RNR, model: see also Taxman & Thanner, 2006). Thus, high intensity interventions are best reserved for high-risk, high-need offenders. Although the principles seem obvious, Andrews (1989) explains that thinking prior to the elucidation of the PEI held that treatment did not work for high-risk offenders. Andrews calls this the "social work

paradox." He suggests that 'nothing works' proponents did not consider the fact that high-risk offenders, by definition, will always reoffend more than low-risk offenders. They mistakenly took higher recidivism rates to mean that treatments were not effective, but rigorous research comparing intensive treatment to non-intensive programs actually shows that intensive treatment programs can help to reduce reoffending for high-risk offenders. Low-risk offenders continue to reoffend at a lower rate than high-risk offenders in both intensive and non-intensive treatment, but studies frequently find that their recidivism increases when they are subjected to intensive programs. Drawing on a wide body of Canadian research, Andrews (1989) concludes that "lower risk cases may be assigned safely to the least restrictive settings" (p. 15). Since then, numerous meta-analyses of correctional treatment have consistently shown that both treatment and supervision work better when the PEI (particularly the risk principle) are adhered to; that is, when a larger proportion of high-risk offenders are served. More importantly for the present study, they have shown that low-risk offenders tend to have less favorable outcomes when they receive higher-intensity programming or supervision (see Lowenkamp & Latessa, 2004 for a summary of the research). As we have seen, this finding is also borne out in studies of intensive probation (Erwin, 1986; Hanley, 2006).

While the idea of placing high-risk offenders in low-intensity supervision clearly seems inadvisable, the commonly-held notion that 'more is better' also means that the thought of reducing supervision of low-risk offenders is not intuitive to policymakers or researchers. Nevertheless, we have presented several theories – deterrence/defiance, deviant peer contagion, and the principles of effective intervention – suggesting that assigning lower-level offenders less supervision may be more appropriate. Low-intensity

probation allows deterrence to work because in terms of sanctions, the consequences of failure are much greater than they are under more intensive conditions. The lower-intensity sanction may also result in less stigma for the offender, which may lead to a defiant criminal response. We also suggested that low-intensity probation, by reducing required attendance at the probation office, may reduce the likelihood that probationers will associate with more serious offenders and strengthen ties with pro-social networks closer to home. Finally, we noted that the idea of low-intensity supervision is consistent with strongly-established principles of correctional treatment in which only those in the most need of services receive them.

## The Philadelphia APPD Low Risk Experiment

The Philadelphia APPD Low Risk Experiment was designed as a rigorous test of a policy of reducing the intensity of probation supervision for low-risk offenders. The Philadelphia experiment is the first to test this proposition using a randomized controlled trial (RCT) design. The following section describes the study design and main results. Additional details may be found in Barnes et al. (forthcoming).

### Background to the experiment

Philadelphia's Adult Probation and Parole Department (APPD), like other probation departments in the U.S., grapples with the problem of increased caseloads and limited resources. The average caseload of a Philadelphia probation officer is 150 to 200 offenders (Berk et al., 2009). Around 19 per cent of people arrested for fatal and non-

fatal shootings in Philadelphia are under APPD supervision at the time of arrest (Ahlman et al., 2008). Although the amount and intensity of supervision does vary across cases and some offenders are mandated (by judges or probation officers) to specialized units, there is little systematic variation in Philadelphia in the investment of resources according to risk. Thus, in most cases, offenders are assumed to be at similar risk of serious reoffending at baseline, and judgments are modified according to information that becomes available later in the supervision process.[1]

Philadelphia's APPD has around 50,000 clients under supervision at any time. Supervision is usually organized according to the sector of the city in which the offender lives, with a smaller number assigned to specialized units for certain types of offenders or needs (such as drug-involved or sex offenders). Intake decisions are made by administrative staff based on court orders or the offender's residence. Within departmental standards and judicial constraints, supervision of offenders is highly discretionary. Clients usually see their probation officer once a month, and receive routine drug tests at some visits, but the officer can increase or decrease the frequency of office visits as s/he sees fit. Reporting frequency may be increased to weekly or biweekly as a result of noncompliance. It may be reduced to as infrequently as once every three months toward the end of a successful term, or based on the officer's judgment that the offender is not at high risk of recidivism. For similar reasons, the officer may vary the type of supervision between office, telephone, and non-reporting.

Starting in 2005, APPD worked with the University of Pennsylvania to develop a new approach to supervision. They aimed to allocate the highest risk offenders to more intensive supervision, with a small ratio of officers to clients so that more resources could

be put into assessing and addressing those clients' needs. In order to do so, the lowest risk offenders in the agency needed to be assigned to large caseloads with minimal supervision, so that officers would be freed up to work more closely with high-risk clients. This represented a departure from the initial 'one size fits all' approach to supervision previously used by the agency (see Fig. 2.1). A risk prediction model was developed to assess which offenders were at low and high risk of offending. The Low Risk Experiment then randomly assigned offenders predicted to be low risk to a low-intensity model of supervision ('LIS') or the normal model of supervision as described above (supervision as usual: 'SAU'). The results of the experiment indicated that LIS can safely be used with low-risk offenders without increasing the risk of serious recidivism. APPD next plans to test the allocation of high-risk offenders to high-intensity or regular supervision.

Forecasting model

The statistical model used to forecast the risk of serious offending is described in full in Berk et al. (2009). Random forests methods were applied to a dataset of all probation and parole cases in Philadelphia between 2002 and 2004 to predict the risk of being charged with a new serious crime[2] within two years of the probation or parole case start date. The prediction was based only on the type of data that would be available to probation officers at intake.[3] At the request of APPD, the model was designed to stratify 61 per cent of cases as low risk, with the remainder either high risk (approximately 10 per cent) or neither low nor high (approximately 30 per cent) (Fig. 2.1). APPD also deemed

the proportions of false positives and false negatives expected in the final model to be operationally acceptable. The proportion of false positives (offenders erroneously identified as low risk) was set at 5 per cent, and the proportion of false negatives (offenders erroneously identified as high risk) was 20 per cent. A higher false negative rate was accepted given the lesser public safety concerns around this type of error.

## Experimental design

### *Selection of Cases*

APPD selected the West and Northeast regional supervision units as the sites from which experimental participants would be drawn. All cases active on probation in these two units on July 27, 2007 were extracted. The random forests model was applied to each case to produce an individual risk assessment (some probationers had multiple cases) in the form of a 'reliability score.' The reliability score is a number between 0 and 1. Cases with a reliability score above 0.5 were designated as low risk. From this assessment, 2,859 offenders were serving probation terms for low-risk offenses. They were pre-screened for eligibility for the experiment. Low-risk cases were excluded from the random assignment pool if any one of the following factors made them ineligible for low-intensity supervision:

- The case was due to expire within thirty days of the extraction date.
- The offender was placed under the supervision of a specialized unit by court order after the extraction date.
- The offender was in an existing low-risk caseload.[4]

- The offender was potentially in direct violation of their probation (had been arrested for a new offense after the start of their term of supervision).[5]
- The offender had multiple active probation cases, one or more of which was not classified as low-risk.

*Random Assignment*

Following the exclusions, a final sample of 1,559 offenders was put forward for random assignment on October 1, 2007. Note that almost half of the probationers with low-risk cases were excluded from the random assignment pool. The most common reason for exclusion was having fewer than thirty days remaining on probation (see Fig. 2.2). APPD wanted to test a low-intensity caseload of 400 clients per officer, so the random assignment sequence was designed to allocate 800 offenders to the treatment (LIS) group, with 400 each in the West and Northeast regions. The control (SAU) group consisted of 759 offenders (401 in the West and 358 in the Northeast). A considerable number of the sample were later found to be ineligible for the experiment, some due to potential violations that occurred between the time of pre-screening and random assignment, and most others for reasons that arose later that made low-intensity supervision too difficult.[6] A flowchart (Moher et al., 2001) showing exclusions and case flow through each stage of the project may be found in Fig. 2.2. The experiment followed an 'intention-to-treat' analysis (Montori & Guyatt, 2001), so those offenders who were randomly assigned to LIS but subsequently excluded were analyzed in their assigned groups rather than according to the type of supervision they actually received.

*Interventions and Follow-Up*

Probation clients assigned to the treatment group (LIS) were placed in a caseload of four hundred. Two probation officers handled the entire low-intensity caseload of 800 offenders. Clients received a considerably reduced level of supervision compared to the standard model described above. The full supervision protocol, including details about contact frequency and type, may be found in Appendix E. At their first visit with the low-intensity probation officer, treatment group subjects were informed that they were in a low-risk caseload and subject to these new reporting requirements. They were told that they would be transferred back to standard supervision if they were arrested for a new crime. Low-intensity officers were not expected to handle new offenses. However, they were expected to deal with technical violations that did not result in an arrest or warrant (e.g., missed contacts). In order to maintain low-intensity caseload sizes at 400, probationers who were transferred back to standard supervision were replaced by 'backfill' cases. These were offenders from the general caseload who had been predicted low-risk but were not initially randomly assigned. These cases are not analyzed as part of the experiment, but they ensured the integrity of the low-intensity model by keeping caseloads too large for the officers to spend more time with their low-risk clients.[7]

The control group received SAU according to the description above. While probation officers in this group had smaller caseloads and could in theory spend more time working to address offenders' needs, in practice caseloads were still large enough that the content of meetings was essentially the same in both the treatment and control

groups. However, control group offenders saw their probation officers more frequently. Control group offenders continued their regular appointments with their usual probation officer and were not informed of their low-risk prediction or experimental status.

Reoffending data, including charges for new offenses, technical violations, and wanted card issuances, were collected for each participant. New charge data are available for the two years post-random assignment. For the low-intensity model to be considered a success, failure rates in the treatment group had to be at least the same, if not lower, than in the control group. As long as recidivism was not *worse* in the treatment group, the APPD deemed low-intensity supervision an acceptable policy.

Main results

Barnes et al. (forthcoming) report experimental outcomes one year post-random assignment. They note that treatment group cases received approximately 45 per cent fewer contacts than they had in the year prior to random assignment, while the amount of contact in the control group did not change. Assuming control group offenders had a face-to-face meeting with their probation officers once a month, treatment group offenders were expected to receive one face-to-face contact for every six control group contacts, or one contact of any type (face-to-face or telephone) for every three control group contacts. This standard was not quite met, but LIS participants still received a lower-intensity intervention. They received about half the number of contacts as the control group overall. In practice, some control group members met their probation officers less than once a month, and some treatment group members saw their probation

officers more often than the experimental protocol required.[8] Nonetheless, as Barnes et al. note: "The two groups were clearly subjected to different numbers of contacts with their probation officers, and the experimental treatment appears to have been delivered as designed in strategy, if not in dosage."

No significant differences in new offending were found between the treatment and control groups after one year. Sixteen per cent of the treatment group and 15 per cent of the control group were charged with a new offense of any type ($p \leq .593$). Similarly, 15 per cent of treatment group offenders and 16.5 per cent of the control group were incarcerated during the same time period ($p \leq .426$). Overall, it appeared that low-intensity probation did not lead to more crime compared to supervision as usual, and is a safe strategy for restructuring probation supervision according to APPD's plans.

Sensitivity analysis of main results

The idea of reducing the amount of resources made available to low-risk offenders may be controversial to policymakers and the public, however logical it may seem to concentrate probation efforts on the higher end of the risk spectrum. Distinctions are rarely drawn between high- and low-risk offenders in popular dialogue. Although murderers and sex offenders usually arouse stronger emotions than low-level thieves or drug offenders, the idea that some people who are involved with the criminal justice system at any level may 'get away with' minimal supervision may offend the public's sense of fairness. Unequal distribution of resources may also make policymakers uneasy, despite being a somewhat obvious money-saving proposition. Sherman (2007, p. 303)

notes that concentrating on the 'power few'[9] may at the same time be considered "perfectly rational and morally reprehensible." Although only a small proportion of probation clients are at the highest risk of involvement in serious crime (as offenders or victims), why should those deemed to be at low risk of serious reoffending be denied the same level of attention from a probation officer if they need it? Not only could low-risk offenders be denied help and services they still need, but they may still pose a serious threat in the future. As one anonymous Philadelphia probation officer told a local newspaper after the restructuring of Philadelphia APPD: "Anybody is capable of anything. You can't just assume [low-risk offenders] won't pose a problem." Another stated: "We don't want to give people a chance to go out and commit more crimes" (Gambacorta, 2009).

These concerns cannot be discounted if Philadelphia's model of low-intensity supervision is to become a viable policy beyond the RCT. However rigorous the experimental design, the main outcomes may mask subtle variations in effects. Different conceptualizations of the outcome measure, differential treatment take-up, and differential subgroup effects may all affect our conclusions about the efficacy of the policy. This paper examines the extent to which different outcome measures and heterogeneity in offender characteristics explain any differences in recidivism outcomes in the sample of probationers assigned to LIS, compared to the control group who received 'SAU.' Our analysis also extends the main results discussed above by increasing the follow-up period to two years post-random assignment, to examine whether the null findings are sensitive to a longer follow-up period.

The specific research questions addressed in this paper are:

1. Does the effect of LIS on recidivism (new charged offenses) differ if we consider recidivism frequency as well as participation?
2. Does assignment to LIS affect the time to failure?
3. Does differential treatment take-up affect the probability of recidivism?
4. Do the effects of LIS differ across offender subgroups?
5. Do the outcomes for the above research questions hold across specific offense types?

## Analytic Strategy

Offending participation versus frequency

A common concern in criminal career research is the distinction between offending participation and offending frequency. Participation refers to whether or not a person was involved in criminal behavior (a dichotomous outcome), while frequency refers to the number of crimes committed within a certain period of time. These two conceptualizations of crime outcomes are open to different interpretations. As Blumstein et al. (1988) note in the context of lifetime offending: "Participation distinguishes active offenders from non-offenders within a population; frequency is a reflection of the degree of individual criminal activity by those who are active offenders" (p. 4). Thus, an experimental intervention could produce differential effects on outcomes depending on how the outcome is measured before and after treatment.

Ideally, we want an intervention that reduces participation *and* frequency of recidivism, resulting in some offenders desisting from crime completely and those that do not at least reoffending less often. However, in practice we may see an impact on one but

not the other. For example, the results of a set of experiments carried out in the United Kingdom to test the effect of face-to-face restorative justice on recidivism showed no change in the number of offenders who participated in crime after the program compared to before. However, those offenders committed 27 per cent fewer crimes on average than they did prior to the program (Shapland et al., 2008). Looking at participation alone, one might conclude that restorative justice was no more effective than regular court processing. However, society still benefits if fewer crimes are committed overall. Conversely, the success of the low-intensity supervision strategy might be doubtful if the lack of difference between groups in the proportion of offenders participating in crime masked an increase in the frequency of offending for the LIS group compared to those receiving SAU. Measures of participation and frequency also produce different policy-relevant estimates of the treatment effect. In the present experiment, the participation measure gives the more accurate effect of assignment to LIS versus SAU, since LIS offenders are returned to SAU after their first new offense. However, if we continue to follow experimental participants after they complete or fail LIS, the number of offenses they commit in the longer term offers an indication of whether spending any amount of time in LIS has a deterrent effect on subsequent criminal behavior.

We assess the effect of the experimental treatment on participation and frequency using regression models designed for binary and count data. We construct a binary logit model for participation. Frequency of offending is analyzed according to a Poisson regression model and several of its variants. When there is evidence of over-dispersion – excess variation not captured by the Poisson distribution – a negative binomial regression model is examined. Because a substantial proportion of our sample did not reoffend at

all, resulting in a large number of zero counts in our data, we explore whether zero-inflated Poisson or negative binomial models fit the data better.[10] Zero-inflated models correct for the large share of zero observations by allowing the zeros to be predicted by two different theoretical processes (e.g., Sarkisian, 2009): the "always zeros" (offenders who will never reoffend, regardless of changes in other conditions); and the "possible zeros" (those who might have reoffended but did not do so during the follow-up period). We present the results of several diagnostic tests used to assess the appropriateness of each model. All the models control for baseline offender characteristics, and account for time at risk according to the time during which offenders were not in jail one year pre- and one year post-random assignment. These features are described in detail below.

Time to failure and survival analysis

An alternative approach to assessing experimental outcomes is to look at the time to failure (time to first offense) rather than simple proportions or counts of new offenses. An experimental intervention may affect time to failure as well as, or even independently of, its effect on participation and frequency. In a probation agency, an understanding of how quickly probationers tend to recidivate after the probation term begins may be important for the allocation of supervision resources. Whether we analyze participation or frequency, we can only say that the treatment group was more or less likely to offend than the control group. We lose important information about the timing of events, and cannot account for the participants who did not reoffend. As Allison (1984) notes: "One might suspect … that someone arrested immediately after release [from prison] had a

higher propensity toward criminal activity than someone arrested 11 months later" (p. 11). Survival (event history) analysis techniques overcome the problems of truncation and omitting the time element by incorporating special regression techniques that allow for the censoring of cases. If low-intensity supervision had no effect on participation or frequency compared to SAU, but survival analysis revealed that low-intensity participants were likely to fail *more quickly* than the control group, it may still be necessary to re-evaluate the policy to avoid turning the low-risk unit into a 'revolving door' that sends participants right back into the criminal justice system.

We use the Kaplan-Meier (KM) method to compare the average risk of failure over time for the treatment and control groups. The KM survival estimator gives the estimated probability of the offender surviving to the end of each time interval for which failure events are calculated. In the present analysis, we measure time to failure in days post-random assignment. We then construct a Cox proportional hazards regression model to explore the risk of failure depending on experimental status and other covariates. The Cox model also allows us to control for post-random assignment time at risk. The proportional hazards model is a semi-parametric model, meaning that it does not impose a specific shape for the hazard function, or probability of failure over time (although we assume that each individual's hazard is proportional to those of others). This allows a greater degree of flexibility than parametric models for time, which require us to choose a particular distribution for the hazard function. Our choice of covariates and methods for accounting for time at risk are described in detail below.

Differential treatment take-up and subgroup effects

Treatment take-up (who actually receives the treatment, regardless of random assignment) and subgroup effects are two different but related issues that could affect the impact of LIS on recidivism. Very few experiments conducted in 'real world' settings operate perfectly (Berk, 2005). Characteristics or circumstances of the offender, probation officer, or agency (as well as errors and individual overrides or ethical concerns) could prevent the delivery of the treatment to those assigned to receive it, or lead to control group members receiving the treatment ('crossover'). Both situations affect the conclusions we are able to draw about experimental outcomes. Similarly, it is conceivable that these characteristics may also interact with treatment, leading to differential outcomes that could masked by the average effect for the full sample. For example, a treatment could prove to be more effective for women than it is for men.

As described above, the analyses of both the main first year results of this experiment and the other research questions presented here are based on the randomly assigned treatment condition (ITT) rather than the treatment actually delivered (TAD). ITT is the preferred method of analysis of the two, because it reduces the possibility of bias resulting from differences in treatment compliance. For example, we know that some treatment group members did not receive LIS because they were later found to be wanted for absconding. Their noncompliance may place them at higher risk of reoffending than other LIS participants. Excluding them from the analysis of the treatment group outcomes could introduce an upward bias in the effectiveness of LIS. The ITT approach avoids this bias by retaining these offenders in the treatment group. As such, ITT provides a better estimate of the *policy* of LIS, because in the real world

some probationers will be eligible for LIS but will not receive it because they abscond or for some other reason. However, the ITT approach involves an unavoidable trade-off between this and a test of the actual effectiveness of the treatment. Because most experiments will include some degree of non-delivery and/or crossover, any treatment effect will be attenuated (Angrist, 2006). In this experiment it is very likely that non-delivery of treatment affected the outcome: 17.8 per cent of probationers assigned to LIS did not receive it (see above and Fig. 2.2). Crossover poses a smaller problem, but does exist: 3.2 per cent (N = 24) of the control group were assigned to LIS at some time. It is not known why the crossover occurred, but it is possible that offender characteristics could be associated with the non-delivery. Factors like gender, race, age, and prior offending history may well be linked to reasons for non-delivery such as absconding or transfer to a specialized caseload for more intensive supervision.

A powerful technique for modeling heterogeneity in treatment effects created by treatment non-delivery and differential subgroup effects is the instrumental variables (IV) method (e.g., Angrist, 2006; Angrist & Pischke, 2009). IV methods can help to overcome the attenuation of the treatment effect involved in ITT analysis by using assigned treatment (ITT) as an instrument for predicting actual treatment take-up. Predicted take-up, rather than assigned treatment or treatment delivered, is then used as the experimental status variable in the crime outcome model. This is different from – and avoids the bias of – estimating the effect of treatment delivered. In more formal terms, it reflects the local average treatment effect (LATE), which is the treatment effect for "compliers" – those who receive the treatment to which they were randomly assigned – rather than the average effect on all treated individuals (ATET), who may be compliers or

control group members who received the treatment anyway (Angrist, 2006, pp. 30-31). Thus, we get as close as possible to estimating the actual effect of the treatment. The risk of bias in the basic treatment as delivered approach is minimized in the IV model because treatment as assigned (ITT) is incorporated in the prediction of treatment take-up. Due to the random assignment, treatment status is independent of observable features of the cases. Furthermore, by including interaction terms between assigned treatment and selected offender characteristics (subgroups) as instruments in the IV model, we can also account for their potential indirect impact on treatment take-up and use the resulting estimates to examine differential effects of subgroup membership on recidivism for LIS compliers.

The subgroups we explore include age, sex, race, socioeconomic status (based on neighborhood-level data for the offender's recorded address), prior offending record, and probation region (West or Northeast). We include region as a subgroup because only two probation officers, one from each regional unit, handled low-risk cases, whereas a much broader range of officers was represented in the control group. With such a limited number of low-intensity supervision officers, it is likely that each officer's personality and willingness or ability to follow the experimental protocol could have affected the operation of the low-intensity model. Unfortunately, the quality of additional data on important crime risk and protective factors like marital status, employment, and drug, alcohol, medical, and psychological issues is poor. Reporting by probation officers of these details is inconsistent. Thus, we are unable to explore these additional factors.

IV methods are applied using two-stage least-squares regression (2SLS), which is related to ordinary least-squares regression (OLS) and incorporates simultaneous

equations. The first stage equation uses the instrument variables to predict treatment take-up (the endogenous variable):

$$\hat{T} = \beta_0 + \beta_{TA} + \beta_{REGION} + \beta_{GENDER} + \beta_{RACE} + \beta_{AGE} + \beta_{SES} + \beta_{PRIORS} + \beta_{TAR}$$
$$+\beta_{TA} * \beta_{REGION} + \beta_{TA} * \beta_{GENDER} + \beta_{TA} * \beta_{RACE} + \beta_{TA} * \beta_{AGE} + \beta_{TA} * \beta_{SES} + \beta_{TA} * \beta_{PRIORS}$$

where $\hat{T}$ is the endogenous variable (predicted treatment take-up), $\beta_0$ is the intercept, $\beta_{TA}$ is the instrument for assigned treatment, the interactions between $\beta_{TA}$ and the subgroup variables are the additional instruments, and $\beta_{TAR}$ is a control (exogenous) variable for post-RA time at risk. Note that only the interactions and not the main effects variables for our subgroups are used as instruments. We hypothesize that race, gender, etc. predict treatment take-up through their association with treatment assignment. In the second stage equation, we replace the instruments with the predicted treatment take-up from the first stage ($\beta_{\hat{T}}$) to predict the crime outcome *Y*:

$$Y = \beta_0 + \beta_{\hat{T}} + \beta_{TAR} + \beta_{REGION} + \beta_{GENDER} + \beta_{RACE} + \beta_{AGE} + \beta_{SES} + \beta_{PRIORS}$$

The main effects for the subgroups remain in the second stage model as controls for any direct variation in outcomes by subgroup. Another important part of the 2SLS approach is the estimation of the 'reduced form,' which is simply the OLS estimate of the ITT effect of the instrument and exogenous covariates on crime. Angrist (2006) notes that it is acceptable to use OLS even if the outcome variable is dichotomous.

The coefficient for $\beta_{\hat{T}}$ tells us the actual effect of treatment received, or local average treatment effect (LATE) on the probability of reoffending. It can be compared with the coefficient for assigned treatment in the reduced form model (or the outcomes from our logistic participation model, as described above) to assess whether the estimated

effect of LIS on recidivism is different when treatment actually received is predicted rather than treatment assigned. We can then use the predicted values from the 2SLS model to assign an individual probability of reoffending for each subject, and compare the mean probability of reoffending at different levels of each subgroup for those who received LIS.

Although 17.8 per cent (N = 142) of the 800 offenders assigned to LIS are known to have been excluded from the treatment, only 16.8 per cent (N = 134) are recorded in our data as such. Since we cannot tell which offenders constitute the remaining eight non-treated cases, we simply analyze them as if they received the treatment. We do not expect this to be a substantial limitation of this analysis, since cases with missing data represent just 1 per cent of the treatment group and 0.5 per cent of the entire study sample. Another limitation of our treatment take-up prediction is that it does not account for the actual dosage of contacts received, which varied from the experimental protocol in some cases. However, we are able to estimate outcomes for those offenders who were assigned to the LIS caseload and were likely to have seen the LIS officer at least once during the course of the experiment.

Model construction

All the models we estimate include the same covariates and controls for time at risk, both pre- and post-random assignment. The covariates we include are gender (male = 1), race (white vs. non-white),[11] the offender's age on the date of random assignment, a basic indicator of socioeconomic status (SES),[12] probation region (West = 1), and the

offender's monthly offending rate in the year pre-random assignment. The monthly offending rate was calculated by dividing the number of offenses for which the offender was charged that took place in the year prior to random assignment by the number of months during that year that the offender was able to offend (i.e., was not in jail). Our dataset contained dummy variables showing whether or not the offender was in jail during each month in that time period. We checked for multicollinearity between the race and SES variables by obtaining the correlation coefficient for the two variables, which was 0.44. Although this is a fairly large coefficient, we also obtained the variance inflation factors, which were all between 1 and 1.5 – well within the conventional threshold for assessing multicollinearity.

We account for time at risk post-random assignment slightly differently in each model. In count models, the logged number of months at risk post-random assignment is included as the exposure or offset variable, allowing us to estimate the incidence rate ratios for person-months of follow-up time for the LIS versus SAU groups. In the binary and two-stage least squares models, we include the number of months at risk as a control variable.[13] We lacked detailed information about time at risk, which is an important limitation of our analysis.[14]

The format of our jail time data causes the most problems for assessing time at risk in the survival analysis model. As previously explained, survival analysis techniques allow us to assess whether experimental participants were offenders or non-offenders on a daily basis. Because we are only interested in the time to first offense, offenders who fail are removed from the risk set because they are no longer at risk of that first failure. However, offenders who are in jail cannot offend, so the days on which they are

incarcerated must also be removed from the risk set – their risk of failure on those days is zero. This would be straightforward if we knew the exact dates of entry and exit from jail as well as the exact offense dates, but our dataset only includes monthly jail status indicators. When the jail variable is measured at less frequent time intervals than the failure event, a possible solution is to treat jail as a time-varying covariate (TVC) and interpolate jail data for each day. For example, if the monthly jail indicator for January 2008 shows that the offender was in jail during that month, we code each day from January 1 to January 31 as a day in jail. Cox regression allows for this approach using 'episode splitting,' which involves creating a separate observation for each person-day up to the day of failure or censoring. It is also possible to work around the potential problem of overlapping jail and failure dates created by the interpolation (e.g., when the original dataset indicates that the offender was in jail in February 2008, but he offends on February 23) by simply dropping the month of jail time in which the offense took place. Thus, the offender in this example is coded as being out of jail during February.

An obstacle to using the episode-splitting approach in the present application is that the method is usually used in cases in which it is possible to observe either level of the TVC on the failure date. For example, if our TVC were employment status, the offender could be either employed or unemployed on the offense date. However, we forced the offender to be out of jail on the failure day because it did not make sense theoretically to allow for the overlap. Thus, our jail indicator variable is always coded 0 when our failure variable is coded 1. This results in perfect collinearity between the covariate and the failure event, which prevents us from estimating parameters for the jail variable using the Cox model. We could still have dropped jail days out of the risk set

(albeit based on our interpolated dates rather than accurate ones), but to maintain consistency this would require us to ignore failures that overlapped with jail time. Since this would have meant ignoring one-third of our failure outcomes (111 of 335 failures overlapped with interpolated jail days), we compromised by including the twelve original monthly indicator variables in our model to control for each post-RA month in jail. This is another important caveat in interpreting the results.

Outcome data and measures

Our main outcomes of interest in the present analyses are the prevalence and frequency of any offense committed within two years after the date of random assignment and resulting in a formal charge. Our pre-RA measures of crime are also based on charged offenses occurring in the year before the RA date. To answer our final research question, we examine these outcomes using specific data on drug and violent offenses. We select violent and drug offending as secondary outcomes of interest because they may be most interesting to policymakers[15].

One considerable limitation of our crime outcomes is that they only reflect charges as an adult for offenses in Philadelphia, as we only had access to local adult criminal justice system databases. While almost all of the participants reside in Philadelphia, the city's proximity and ease of access to surrounding counties and state lines mean that the local data almost certainly underestimate the number of charged offenses recorded for these offenders. In addition, we do not have juvenile data available to give a full picture of offenders' lifetime criminal involvement. While most of the

offenders in our sample are older (see Table 2.1), we would expect to see the majority of their criminal offending taking place during their teenage years, so the lack of data on charges filed under the age of eighteen is a substantial omission. However, data are available where the offender was charged as an adult, even if the offense was committed while s/he was under eighteen.

Some of the covariates and control variables we include in our model are likely to be confounded with crime outcomes. For example, it is well-established that younger offenders commit more crime than older offenders. Months in jail post-random assignment may also be associated with crime outcomes, as those who spend more time in jail are likely to be more serious offenders, so will be at greater risk of reoffending while at liberty. For reference we present the conditional distributions of these two variables with the likelihood of committing a new charged offense in Appendix G. Neither of these issues poses a substantial threat to the validity of the data because the covariates are not additionally confounded with the treatment instrument.

Sample characteristics

We assess our four research questions using the full experimental sample of 1,559 offenders (800 LIS treatment and 759 SAU control). They were followed up from the date of random assignment, October 1, 2007, for two years to September 30, 2009. Of the 800 offenders randomly assigned to LIS, 94.5 per cent (N = 756) actually received the treatment and 5.5 per cent (N = 44) did not. Of the 759 SAU offenders, 3.2 per cent (N = 24) were inadvertently placed on low-intensity supervision. Offenders are analyzed according to assigned treatment except in our instrumental variables model, as discussed

above. The random assignment sequence was designed so that the West and Northeast LIS groups each contained 400 offenders. The control group contained slightly more cases from the West than the Northeast (52.8 per cent of the control group was from the West regional unit).

Table 2.1 shows basic demographic and offending history characteristics for the two groups. There are no statistically significant differences between the treatment and control groups on any measure, indicating successful random assignment. The sample is predominantly male (67.0%), nonwhite (60.1%), and on average 41 years old at the time of random assignment. The majority (91.8%) of offenders lived in zip code areas with an average household income of $20,000 or more. As we might expect from a sample already predicted to be low-risk, prior offending rates are low. Approximately 12 per cent (N = 193) members of the sample were incarcerated at any time post-random assignment, for an average of 1.1 months. Members of both groups committed 1.3 offenses per month at risk on average in the year prior to random assignment, with much lower rates for violent and drug offending. Post-RA, the marginal first year difference between the treatment and control groups reported above disappears by year two. Control group members tended to engage in violent recidivism more than the treatment group in the first year post-RA (4.1% vs. 2.9%), but by the second year the gap has begun to close (5.4% vs. 4.5%). Treatment and control group participants committed new drug offenses in similar proportions in the first year post-RA (treatment: 6.4% vs. control: 6.5%), but slightly more of the control group had failed by the second year (control: 10.1% vs. treatment: 8.9%). In all, 21.5 per cent (N = 335) of the 1,559 offenders in the sample committed a new offense of any kind two years post-RA. Among the 335

recidivists in the sample, 77 committed violent offenses and 148 committed drug offenses. The violent and drug offender samples are not independent: some engaged in both types of offending during the follow-up period.

## Results

Table 2.2 shows the odds ratios (OR) from the logistic regression model for the effect of assigned treatment on offending participation two years post-random assignment. The results indicate that when other offender characteristics are controlled, there is no notable difference in the odds of recidivism between the LIS and SAU groups (OR = 1.05, $p \leq .707$). Several other offender characteristics appear to have a greater impact on the odds of a new offense regardless of treatment status. Each additional year of age at random assignment is associated with a 2 per cent decline in the odds of recidivism (OR = .98, $p \leq .002$). SES is consistently associated with significantly reduced odds of reoffending: between 40 and 75 per cent compared to the lowest SES group. Each additional month in jail post-RA is associated with a 29 per cent increase in the odds of a new offense (OR = 1.29, $p < .001$). This could suggest that offenders who spent more time in jail were likely to be at higher risk of reoffending while on the streets.[16] We tested an additional model that included a squared term for post-RA jail time, to account for nonlinearity, such that increased jail time could eventually lead to a decline in reoffending by curtailing time at risk. This term was dropped as it did not reach statistical significance,[17] perhaps because we did not have jail data for the second year. Probationers in the West probation region had 40 per cent lower odds of

reoffending than probationers in the Northeast (OR = .60, $p \leq .002$). This is most likely due to some significant demographic differences between the two regional samples, rather than any effect of the treatment.[18]

Table 2.3 shows the count model outcomes for frequency of reoffending. We used a zero-inflated negative binomial model to produce the coefficients. We display the incidence rate ratios (IRR) for the number of new offenses in the sample across the total time at risk. Following the strategy explained above, the zero-inflated negative binomial model was selected following tests of its fit against the actual observed criminal offending frequencies versus those fitted from the Poisson and negative binomial regression models.[19]

The full model estimates presented in Table 2.3 indicate that assignment to LIS supervision, controlling for other factors, is associated with a small, non-significant reduction in the number of offenses committed post-random assignment (IRR = .89, $p \leq .489$). Other offender characteristics had a greater effect on offending frequency, regardless of treatment assignment. Gender, which had no effect in the participation model, appeared to be an important factor in explaining offending frequency. The rate of offending for men was twice that of women (IRR = 2.03, $p < .001$). Increased age was again associated with declining offending rates (IRR = .98, $p \leq .009$), and was also an important predictor of non-offending in the inflated model. Interestingly, although membership of the West region group was associated with a lower odds of committing any new offense in the logistic model, and also predicts non-offending in the inflated model, those probationers in the West who did offend committed considerably more offenses than recidivists in the Northeast, although this was not statistically significant

(IRR = 1.41, $p \leq .107$). SES did not predict frequency of recidivism, but increasing status was significantly associated with non-offending.

Figure 2.3 shows the Kaplan-Meier survival estimates for the treatment and control groups. The Kaplan-Meier survival estimate is the probability of *not* offending on a daily basis. The most striking feature of the graph is that the probability of failure is very low for all low-risk offenders. Figure 2.3 is a visual representation of the figures reported above: approximately 21 per cent of the sample had failed after two years. The survivor functions for the treatment and control groups look extremely similar. This is confirmed by the log-rank test for equality of survivor functions, which indicates that the probability of survival in the two groups is identical ($\chi^2$ (1 d.f.) = .00, $p \leq .996$).

We extend this basic comparison by modeling time to first failure with controls for additional covariates using Cox proportional hazards regression.[20] Table 2.4 presents the results of the regression model in terms of hazard ratios (HR). The hazard is the risk of failure over time, and thus is equivalent to the incidence rate ratio. The model confirms that there is no difference in hazards between the treatment and control groups (HR = 1.03, $p \leq .777$). As we saw in the results for prevalence and frequency outcomes, other factors appear to have a greater influence on the risk of failure regardless of treatment assignment. Being in the West probation region is associated with a 36 per cent lower risk of failure over time than being in the Northeast, as we would expect from the result of our logistic model (HR = .64, $p \leq .001$). Older and higher SES offenders were also at a significantly lower risk of failure over time.

As we have seen, Cox regression does not require us to make assumptions about the shape of the hazard or survivor functions. However, it is possible to graph the

'typical' survivor function for the model. Figure 2.4 shows the covariate-adjusted comparative survivor functions for the treatment and control groups. In this graph, the continuous covariates are held at their means and categorical covariates are set to the modal category. Thus, the 'typical' offender is male, nonwhite, about 41 years old, supervised in the West region, lives in a zip code area with an average household income between $20,000 and $29,999, and was not incarcerated post-random assignment. The graph shows no difference in the comparative survivor functions for the average treatment group offender compared to the average control group offender, confirming earlier findings. The 'typical' offender has a slightly lower risk of failure over time than the uncontrolled sample average. Only around 15 per cent had failed after two years.

The results of the instrumental variables model used to explore whether differential treatment take-up impacts the effect of LIS are presented in Table 2.5. The first part of the table shows the results of the reduced form OLS regression, from which we obtain the probability of offending by assigned treatment, and the first stage of the 2SLS regression, where assigned treatment and its interaction with offender characteristics are used to predict treatment take-up. The second part of the table shows the outcomes from the second stage regression, which shows the actual effect of treatment on those who comply with random assignment.

The first result to note is that the reduced form (ITT) model is largely consistent with our earlier findings about the impact of assigned treatment and other covariates. There are slight differences because we include interaction terms in this model to adjust for heterogeneity in treatment assignment, but we still see little impact of LIS on the probability of recidivism. SES, region, and months in jail were again related to the

probability of failure regardless of group assignment. The second stage model shows no difference in recidivism for those who actually received the treatment, controlling for time at risk ($b$ = .009, $p \leq .686$). The results from the reduced form also show no significant interaction effects between observable case features and treatment assignment, suggesting that differences in the main null finding are not driven by differences in who was assigned treatment.

We used the fitted values obtained from the IV model presented in Table 2.5 to assess subgroup differences in outcomes among those actually receiving the assigned LIS treatment (N = 690). Table 2.6 shows the mean probability of failure for each level of each subgroup. We observe significant differences in the probability of failure across all the subgroups except race. Consistent with the findings from the ITT models, the probability of offending in the West was lower than that in the Northeast for LIS compliers (17.8% vs. 22.3%; $p < .001$). We included region as a subgroup because only one officer from each regional unit supervised all the low-risk offenders. Thus, we hypothesized that the officers' personalities and ability to follow the experimental protocol might affect outcomes. Although we cannot rule out the possibility that the difference is due to other unobserved factors, as discussed above, this finding provides some evidence in support of that hypothesis. The regional difference continues to appear when we examine only those offenders who were actually exposed to the LIS officers, controlling for several other factors that may have affected treatment take-up and recidivism. Gender had a significant impact on the probability of failure, with males more likely to reoffend than females (20.7% vs. 18.3%, $p \leq .036$). We see a significant decreasing probability of recidivism with increasing age (probability range: 16.2% -

24.1%; $p \leq .001$). Offenders in the lowest SES category were most likely to reoffend, and those in the highest category least likely; there was little difference between the central groups (probability range: 10.8% - 25.1%; overall $p < .001$). The offender's rate of offending in the year before random assignment also predicts treatment failure. Interestingly, those who had not offended in the preceding year were more likely to fail than those with up to or more than one offense per month at risk (20.9% vs. 16.8% and 17.3% respectively; $p \leq .026$). We explore this finding in the discussion section.

We now examine the offense-specific models for violent and drug recidivism. Table 2.7 shows the logistic regression model for the prevalence of new charged violent offenses in the two-year follow-up period. Assignment to the treatment group was associated with a non-significant reduction in the odds of violent reoffending compared with the treatment group (OR = .89, $p \leq .644$). As was the case in the full reoffending model, West probation region, increased age, and increased SES were associated with significant reductions in the odds of recidivism, while post-RA jail time was associated with increased reoffending.[21] In addition, we see substantial effects of gender and prior offending history on the prevalence of violent recidivism, which was not apparent in the full offending model. The odds of a new violent offense were 3.5 times higher for males than females (OR = 3.53, $p < .001$), and offenders who had committed at least one violent offense during the months they were at risk one year pre-RA had 2.5 times the odds of a violent offense as those who had not. Recall that the sample probability of committing a violent offense was very small (less than 5%; N = 77), which may affect the size of the test statistics for these estimates.

The count model outcomes for violent reoffending are presented in Table 2.8. Again, a zero-inflated negative binomial model proved to be the best fit for our data.[22] Assignment to LIS supervision, controlling for other factors, is associated with a small, non-significant decline in offending over the two-year follow up period than assignment to the control group (IRR = .83, $p \leq .453$). The only significant predictor in the full model is the pre-RA violent offending rate, which is associated with a large decline in the rate of violent reoffending (IRR = .34, $p \leq .047$). This is surprising because violent offending history was strongly associated with an increase in the *prevalence* of violent recidivism. Similarly, gender, which was also associated with increased prevalence of reoffending, is associated with a decline in offending frequency. Income and probation region, which were associated with reductions in overall offending, are associated with increased frequency of violent offending However, it is likely that these results are skewed by the very small number of offenders charged with violent offenses post-RA, and the wide variation in the number of charged offenses. Most of the 77 violent offenders were charged with between one and three crimes, but the count ranges up to 52.

Figure 2.5 shows the comparative Kaplan-Meier survival estimates for violent offenses in the treatment and control groups. The graph supports the estimates from the prevalence and count models. It appears that the probability of failure is slightly higher in the control group, and that they fail more quickly than the treatment group (although note that the scale on the y-axis magnifies the size of the gap between the two lines). The log-rank test indicates no significant difference between the time to failure across the two groups ($\chi^2$ (1 d.f.) = .69, $p \leq .405$). Similarly, when controlling for other covariates in a Cox regression model, we see no difference between the treatment and control groups

(HR = .91, $p \leq .686$: Table 2.9 and Figure 2.6[23]). As we saw in the logistic regression model, males had a significantly higher risk of failure than females (HR = 3.24, $p \leq .001$), and offenders with prior violent offenses in the year pre-RA were at greater risk of failure than those without (HR = 2.35, $p \leq .02$). SES and age were also associated with declines in the risk of failure.

Note that we only use one variable to control for time at risk post-RA in this model, compared with the twelve monthly indicator variables included in the full offending model. The dummy indicators were dropped from this model because of problems that were likely caused by the very small proportion of failures in this sample. In the first few months of the first year follow-up period, the numbers of eventually failing offenders who were in jail did not vary at all. The resulting multicollinearity between each of these monthly variables prevented us from being able to estimate their parameters. However, the single jail indicator violates the proportional hazards assumption for this model (see Appendix H for the results of the diagnostic tests). This is to be expected, because we are treating jail stays as time-constant in our model, when they clearly vary with time. We account for this nonproportionality by constructing a new model that allows jail to vary with time by including a jail*time interaction term. The results of this model are presented in Table 2.10. We present the unexponentiated coefficients for this model because the interaction term makes more sense on that scale. However, if we obtained the hazard ratios for the other covariates we would see that the addition of the interaction term barely makes a difference to our original estimates overall. For example, the coefficient of -.092 for treatment group assignment converts to a hazard ratio of .91 – identical to the estimate from the first model. The *p*-value for this

estimate is also practically identical: .691 compared to .686. Thus, our two alternative controls for time at risk do not appear to impact our estimates of the risk of committing a violent offense over time.

Table 2.11 shows the results of the IV regression for violent offending. Consistent with earlier findings, the reduced form (ITT) model shows a slight, non-significant reduction in the probability of violent recidivism for offenders assigned to LIS ($b$ = -.073, $p \leq .239$). Gender, income, and months in jail are also significantly associated with the outcome. However, as in the model for all offense types, the small effect of treatment disappears completely for those who actually received it ($b$ = - .007, $p \leq .586$).

The subgroup effects for violent offending (Table 2.12) are similar to those for all offenses. The overall probability of offending is small across all the subgroups. Region, gender, age, and SES are all significantly associated with outcomes for LIS compliers, and race is not. The prior offending rate (which is dichotomized into no offending vs. any offending for violent and drug offending due to very low offending rates) does not quite reach statistical significance, but the results are in the opposite direction from those observed in the all offenses model. Those with no prior violent offenses were less likely to reoffend compared to those with one or more charge (3.7% vs. 6.1%, $p \leq .073$).

The outcomes of the logistic regression model for prevalence of drug offending are presented in Table 2.13. The results are very similar to the preceding analyses. Treatment group assignment has almost no effect on reoffending (OR = .92, $p \leq .644$). West region, age, and increasing SES are again associated with decreased odds of reoffending, and males had twice the odds of females of committing a new drug offense. On this occasion, we also found that an additional squared jail time term was significant

($p \leq .008$),[24] and it was retained in this model. It would appear that for drug offending, more time in jail may lead to a reduced likelihood of reoffending to some degree. Since we only have one year of jail data, it could be the case that the odds of drug offending decline as jail time increases, but increase again when all offenders are 'returned to the risk set' in the second year. However, because our outcome data extend beyond the range of the jail data, these coefficients should be interpreted with caution.

We used a zero-inflated negative binomial model to assess the frequency of drug offending (Table 2.14).[25] Although it does not reach statistical significance, there is a notable 22 per cent decline in the drug reoffending rate for the treatment group compared to the control group (IRR = .78, $p \leq .169$). Increased age was also associated with a decline in the rate of drug offending of 2 per cent per additional year (IRR = .98, $p \leq .045$). As before, males reoffended at a higher rate than females and increased SES was associated with reduced reoffending, but these relationships were not as strong as in previous models.

The Kaplan-Meier survival estimates for drug offending are shown in Figure 2.7. The pattern of survival probability is very similar to the patterns for overall and violent offending, with the control group at slightly greater risk of failure by the end of the two-year follow-up period (log-rank test for equality: $\chi^2$ (1 d.f.) = .72, $p \leq .395$). Table 2.15 and Figure 2.8 present the hazard ratios and graphical representation of the estimated survivor functions from the Cox regression model.[26] Again, no difference is evident in the risk of failure over time between the treatment and control groups (HR = .93, $p \leq .673$). Gender is again associated with a significantly higher risk of failure, with males at

twice the risk of females (HR = 1.96, $p \leq .001$), and age and increasing SES is associated with a lower risk over time.

The IV regression results for drug offending are presented in Table 2.16. Again, there is a slight reduction in the probability of drug offending associated with LIS assignment and take-up, but its impact is extremely small and becomes even smaller among those who actually receive the treatment, controlling for time at risk (ITT: 3.5% reduction, $p \leq .680$; second stage: 1% reduction, $p \leq .550$). As before, age and increased SES were associated with a reduced risk of reoffending in the ITT model. Those in the highest SES category were 10 per cent less likely to commit a new drugs offense than those in the lowest category, regardless of treatment assignment.

Contrary to other models, we see no effect of probation region or prior offending history on recidivism for drug offenses among LIS compliers (Table 2.17). The only significant differences we observe are by gender, age, and SES. Males had a 9.5 per cent probability of reoffending compared to 4.5 per cent for females ($p < .001$) when actually supervised in the LIS caseload. There is a linear relationship between increased age and SES and a reduced probability of recidivism.

## Discussion

Our first research question assessed whether the impact of LIS changes if we consider offending participation (the proportion of offenders with a new offense) compared to frequency (the number of offenses committed by the offenders who fail). Overall, there was no difference at all in the prevalence of new charges between the

treatment and control groups after two years. When we controlled for other factors that could affect reoffending, such as offenders' demographic characteristics, criminal histories, and time at risk for reoffending, assignment to LIS was associated with no differences in the odds of recidivism. On the other hand, age, SES, and gender were strongly associated with both the probability and number of new offenses.

Although no difference in the prevalence of new offenses was observed, it is still possible that the frequency of offending could be affected by the experimental supervision strategy. In particular, in an analysis where our controls for post-random assignment time at risk are relatively weak (see above), any impact on offending frequency might be telling because the proportion of offenders failing will be attenuated by the fact that some did not have the opportunity to do so because of incarceration. Of course, time in jail restricts the number of offenses that could be committed too, but it allows us to examine whether those who had the opportunity to offend and received LIS did so at a higher rate than those assigned to regular supervision. Further, offending frequency is more informative about the longer-term effects of treatment than prevalence, which measures the more immediate impact of LIS. However, in our analysis we found little difference in offending frequency between groups. Some of the effect sizes were moderate, but did not reach statistical significance.

Our third alternative outcome measure was time to failure. We examined whether assignment to LIS might have caused offenders to commit a new offense more or less quickly than their control group counterparts. A potential danger of reducing supervision is that offenders may then be on a 'free rein' to engage in offending, whereas those under closer scrutiny may have an incentive to wait until their period of supervision comes to

an end. On the other hand, research on intensive supervision programs has indicated that increased supervision may result in increased detection of new offenses or violations. We found no differences in the time to failure for LIS and control offenders, regardless of whether or not other covariates were controlled. In reality, the probability of offending in both groups of low-risk offenders was so low that any differences that might have existed are probably too minor to detect. Our survival analyses confirmed that the average low-risk offender has a very low probability of failure over time. A limitation of our entire analysis that is particularly important here and has already been discussed at length is that we were not able to account for offenders' time on the streets and ability to reoffend on a daily basis.

We found no evidence that the substantial non-delivery of treatment affected the results we find elsewhere in our analysis. We predicted actual treatment take-up based on assigned treatment and its interactions with offender characteristics that might predict non-delivery. The effect of the treatment for 'compliers' who were randomly assigned to and actually received LIS was even closer to zero than it was in our ITT-based analyses. This provides further support for LIS as an appropriate strategy for dealing with low-risk offenders. Many of those who did not receive the treatment as assigned were likely to have been higher risk offenders. The majority were excluded because of factors that occurred before random assignment but were not discovered until afterwards, such as noncompliance, absconding, or placement in intensive treatment-based caseloads before random assignment. These offenders may have been more likely to offend regardless of the type of supervision they received, and their inclusion in the ITT analyses may have led to the slightly higher prevalence of offending we saw in the treatment group.

Our final analysis, which examined subgroup effects for those actually receiving the treatment, did show some substantial differences between LIS and SAU. Age, gender, SES, and to a lesser extent prior offending history were significantly associated with outcomes for general and specific crime types. Overall, it appeared that low-risk offenders, despite considerable heterogeneity in characteristics compared to the traditional image of the young, male offender that appears in the criminological literature, were homogeneous in their propensity to reoffend. However, when we predicted outcomes according to offender characteristics, we saw that the offenders with the highest probability of recidivism were those who look more similar to that 'traditional' offender. Young males from low-income neighborhoods were close to or above the sample average in their likelihood of reoffending, while the less traditional offenders were generally well below it. Our sample, on average, was 'non-traditional,' being older and containing a broad mix of gender, SES and other characteristics. Thus, we may conclude that there is such a thing as a typical *low-risk* offender, who looks different from the norm. The propensity to reoffend is homogeneous within this offender subpopulation. The low-risk prediction model may also identify offenders who have had little contact with the criminal justice system, or younger offenders who have so far only engaged in low-level offending, but exhibit some of the risk factors usually associated with more extensive criminal careers. Procedures could be built into the low-intensity supervision model to subject these offenders to somewhat more monitoring than 'typical' low-risk cases (perhaps more frequent check-ins by telephone than the experimental protocol requires). Focusing some more attention on the more traditional offenders in

low-intensity supervision would improve the likelihood of picking up the false positives, but resources would still have been conserved if the risk prediction proved accurate.

Two other notable findings from this analysis were the differential effects by probation region, and the finding that offenders without any offending history in the previous year were more likely to fail than those without. The latter can be easily explained. This variable was a rate calculated by dividing the number of offenses by the number of months in the year in which the offender was out of jail and able to reoffend. Thus, those who spent the entire year in prison have a rate of zero, as do offenders who were at liberty the whole time but did not engage in criminal behavior. Offenders who spent more time in jail are more likely to be serious offenders with a higher risk of failure once free. It is difficult to tell whether the difference in supervision styles between the two probation officers really drove the difference in recidivism outcomes between the two regions, but reduced recidivism in the West was consistently observed. Since the officers were required to have minimal contact with their clients, it seems unlikely that their personalities would have made a great deal of difference to offenders' behavior. Thus, unobserved factors most likely operated here. We discussed above some of the differences in offender characteristics between the two regions. It is also possible that more offenders from the West region were incarcerated (or incarcerated for longer periods) during the follow-up period than those in the Northeast.

A major limitation of our subgroup analysis was a lack of available information about offenders. The subgroups we studied – gender, race, SES, age, and offending history – are theoretically some of the most important covariates with offending, but there is no reason why they would specifically impact the performance of low-risk offenders

on low-intensity supervision. There are many other factors that could be more relevant to this association, especially given the possible theoretical foundations for low-intensity probation set out above. Do low-risk offenders perform better under a lack of supervision if they have the support of a spouse, or the structure of employment? How do drug, alcohol, or mental health issues interact with reduced supervision and more limited opportunities for intervention? We can hypothesize that the lives of low-risk probationers may be less chaotic than those of more hardened offenders, such that factors related to stability like marriage and employment could have a more profound impact on their success or failure. On the other hand, we would expect to see differences by gender and past behavior in almost any sample subject to any intervention.

We also examined whether our results held for more specific offense types. Although the offenders in this sample were predicted to be low risk and reoffended at a very low rate, the range of offenses they committed ranged from bad checks to homicide. We selected two fairly common offense types that may be of interest to policymakers, especially in the context of a supposedly low-risk caseload: violence and drugs. Largely, our results did not diverge from those found for all offenses (although these offenses are a subset of the latter outcome measure, so this is to be expected to some extent). There were no differences between groups based on time to failure or treatment delivered for either offense type, and the same subgroup differences were observed, although there was slightly more homogeneity in recidivism propensity across drug offender subgroups.

## Conclusion

The aim of this paper was to examine the viability of giving low-risk probationers less supervision compared to usual probation standards, in order to reallocate resources toward the highest risk offenders. We explored whether the characteristics of low-level offenders and the types of outcome measures selected in the analysis of the first randomized controlled trial of low-intensity supervision (LIS) affected the impact of LIS on recidivism. Overall, we found no evidence against the hypothesis that reducing supervision for the lowest-risk offenders is an efficient strategy.

The credibility of probation and parole supervision as a safe and effective strategy for dealing with offenders in the community is under threat from a poor image driven by a severe lack of resources and some high-profile failures. Although most probation agencies do use some form of risk assessment or triage process to steer higher-risk offenders into smaller, more intensive caseloads, real-world constraints may mean that even these caseloads operate at full capacity. This leaves probation officers with little time or ability to provide appropriate services. The Philadelphia Adult Probation and Parole Department (APPD) was one agency that experienced these difficulties first-hand, and worked with the University of Pennsylvania to conduct an experimental analysis of a new strategy for assessing risk as the basis for channeling offenders into appropriate supervision. The first stage of the restructuring of supervision, on which the present study is focused, involved directing offenders at the lowest risk of serious reoffending into large caseloads, in which they received few probation contacts. This strategy was intended to free scarce staffing resources to be used for reducing caseloads and increasing APPD's ability to provide suitable surveillance and services at the highest end of the risk

spectrum. However, this strategy is politically risky. Arguably, even the lowest risk offenders are still adjudicated criminals who pose a threat to society and should not be left effectively unsupervised. One could also claim that the needs of probationers should not be ignored by an agency supposedly put in place to serve them, simply on the basis of a decision that those probationers are unlikely to reoffend.

The consideration of a policy of LIS also raises the broader theoretical question: what is the nature of low-risk offending? Little research has been produced on this question to date. Philadelphia's statistical prediction model proceeded from the basic argument that any offender unlikely to commit a serious crime within two years of their probation start date should be considered 'low-risk.' The present analyses use the model-generated sample to explore further the general characteristics of low-risk offenders, and whether heterogeneity among the sample affects LIS take-up and outcomes.

We were unable to find any evidence that reducing probation supervision for low-risk offenders causes harm. After two years, the probability of failure was identical in both the LIS and regular supervision groups. There was no indication that LIS clients failed more quickly than their counterparts on regular supervision. We used a rigorous analytic strategy to examine the impact of treatment on those who actually received it, avoiding the bias associated with simpler methods by accounting for factors potentially related to non-delivery and maintaining the integrity of random assignment. Although we found significant differences in the likelihood of reoffending by subgroups, overall reoffending rates were low and there was no indication that the main effects analysis masked any major backfire effect of LIS. Our findings held up across specific offense types as well as for all charged offenses.

We found that low-risk offenders are a relatively heterogeneous group compared to the 'typical' image of the young male offender. Low-risk offenders appear to encompass a much broader spectrum of society in terms of gender, race, age, social circumstances and offending profiles. We found no evidence that this heterogeneity has any impact on which offenders actually received the treatment. The homogeneity in low-risk offenders' propensity to reoffend is interesting from both theoretical and policy perspectives, because it suggests that the relationship of crime risk factors to the likelihood of reoffending may be considerably weakened when the risk of a future serious offense is low. This may be a much more important theoretical basis for the effectiveness of LIS than any of the possible mechanisms we considered at the outset of this paper. The finding is also important for the operation of a probation agency because it implies that specific case attributes do not affect probation performance for a large proportion of the population. Differential attributes only appear to come into play for the majority of this group at the risk prediction stage. After that, the propensity of low-risk probationers to offend is going to be the same regardless of the level of supervision they receive, with just a few modifications required for early-career or more 'traditional' clients who may still be on an upward trajectory of offending. Overall, this seems to be a very powerful justification for the use of LIS instead of 'supervision as usual.'

Thus, we conclude that low-intensity probation supervision, coupled with a rigorous method for predicting the risk of serious offending, is a defensible model for effective probation operations. It remains to be seen whether the other arm of APPD's restructuring strategy – providing increased supervision and services to the most serious offenders in small caseloads – proves successful, but we have ascertained that it can be

done within existing departmental constraints by reducing supervision for those already deemed unlikely to fail. This experiment clearly demonstrates that there is no need to distribute valuable resources equally to all types of offender. Struggling probation agencies should ask whether it remains necessary to provide more supervision when it makes so little difference to the offending outcomes of what will likely be a large proportion of their total caseload. Whichever way one looks at it, probation supervision for low-risk offenders is clearly one area where 'more,' in the usual care sense, does not inevitably mean 'better.'

## Notes

[1] All operational information about the Philadelphia APPD presented in this paper was gathered through conversations with APPD staff and University of Pennsylvania research staff, and in-person visits to the APPD offices.

[2] In this experiment, murder, attempted murder, aggravated assault, robbery, and sexual offenses were deemed 'serious.'

[3] Intake information includes the offender's personal and residential characteristics, and information about the instant offense and prior criminal history.

[4] Prior to the implementation of the Low Risk Experiment, Philadelphia APPD already had several low-risk caseloads within the regional units. Offenders were assigned to them based on the judgment of their probation officer rather than a standardized prediction model. However, because they had already experienced a systematic lower-intensity model of probation compared to standard practice that was too similar to the experimental design, they were excluded from the Low Risk Experiment.

[5] Offenders with potential direct violations were excluded because the workload of preparing new cases for court was deemed too onerous for a probation officer with such a large caseload. All offenders assigned to low-intensity probation were returned to standard supervision if they were arrested for a new offense during the experimental period.

[6] A major exclusion criterion that had not been considered at the time of random assignment was the FIR (Forensic Intensive Recovery) condition. FIR offenders are supervised in regional caseloads but are required to attend an intensive drug treatment program. The supervision of their participation in the program was too involved for low-intensity probation officers to handle in their large caseloads. Pre-screening also revealed that a considerable number of offenders were either on absconder warrants or had not been in contact with their probation officer for more than 90 days (which is grounds for obtaining a warrant). Again, because these offenders were in violation of their probation and more work would be required to process them, they were not transferred to low-intensity supervision.

[7] LIS participants who were transferred back to standard supervision as a result of a violation were analyzed as randomly assigned.

[8] There was a transitional period after random assignment during which some treatment group participants were still attending appointments that were scheduled before random assignment. In addition, one of the low-intensity probation officers was somewhat resistant to the idea of reducing supervision and continued to schedule monthly visits. This was discovered about two months into the experiment, and with further training the officer began to schedule visits according to the protocol.

[9] The 'power few,' as described by popular author Malcolm Gladwell (see Sherman, 2007) is a phenomenon found throughout social research. It is the small fraction of a population to which a disproportionate amount of a certain resource or condition may be attributed. In criminological research it is often noted that a small proportion of offenders or places produce a substantial amount of the total crime (e.g., Sherman, Gartin, & Buerger, 1989; Weisburd et al, 2004). Within a probation agency, a small proportion of probationers are at the greatest risk for committing most of the serious offending among the agency's clients (see Fig. 2.1).

[10] Regular Poisson or negative binomial models may underpredict zeros and overpredict larger numbers, which is problematic when the majority of the data are zeros.

[11] Problems in the coding of the race variable in our dataset forced us to use this dichotomy rather than a more detailed categorical variable for race. The race indicator variable was populated with data from two different sources, with one source selected as the default. However, serious discrepancies arose because the categories of race in the two original sources were substantially different.

[12] Information about SES at the individual offender level was not available. However, 2000 Census data were obtained for each offender's recorded zip code. We used the Census measure of average household income for the offender's zip code as an estimate of SES. This was coded as a categorical variable with four levels: less than $20,000 (used as the reference category in our models); $20,000-$29,999; $30,000-$39,999; and $40,000 or more.

[13] We recognize that the jail time variables included in our models may be endogenous; that is, the effect of jail time on the odds of recidivism may in fact represent a causal effect of the recidivism outcome on the jail variable. We are unable to separate post-random assignment jail time resulting from pre- and post-RA offending. Thus, while we present the models with jail time controls included, we also ran each model without those variables and include the results in Appendix F. Appendix F shows that the inclusion of the terms did not substantially bias our findings.

[14] Our only data on the timing of jail stays are contained in monthly dummies for whether or not the offender was in jail in that month. A further limitation is that these variables are only available for the first year post-random assignment. Thus, while we control for post-RA time at risk as far as possible, it is important to remember in the analysis that the second year of follow-up data is analyzed as if none of the sample spent time in jail. While this does not greatly affect the participation-based outcome measures, it does mean that our post-random assignment offending frequency estimates may be overstated and the survival analysis models overstate the number of person-days at risk (some offenders who would have been incarcerated in the second year are treated as if they had a nonzero probability of offending for the entire year).

[15] Where these offense-specific outcomes are used, the covariate for monthly pre-RA offending rate used in our models is also based on these specific offense types, rather than all offending.

[16] On the other hand, given the possibility of endogeneity, it could also suggest that offenders who commit more than one offense post-random assignment spend more time in jail.

[17] OR = .98, $p \leq .08$. A likelihood ratio test comparing the models with and without the squared terms also indicated that the inclusion of the squared term did not improve model fit: LR $\chi^2$ (1 d.f.) = 3.16, $p \leq .08$.

[18] Probationers from the West were significantly older than those in the Northeast, and significantly more likely to be nonwhite and of lower SES. We examined several combinations of interaction terms between region and these offender characteristics, and found that West region offenders at the $20,000-$29,999 SES level were significantly less likely to reoffend ($b = -1.84$, $p \leq .023$). When this interaction is controlled, the probability of recidivism is higher in the West than the Northeast, but the association is non-significant (OR = 2.91, $p \leq .172$).

[19] The likelihood ratio test comparing the negative binomial and Poisson models was highly significant, suggesting that the negative binomial model fits the data better (LR $\chi^2$ (1 d.f.) = 6913.3, $p < .001$). The likelihood ratio test comparing zero-inflated negative binomial versus zero-inflated Poisson also supports the use of the former model (LR $\chi^2$ (1 d.f.) = 1930.81, $p < .0001$). The Vuong test statistic comparing the zero-inflated and standard negative binomial models is positive and highly significant ($z = 7.04$, $p < .0001$), again suggesting that the zero-inflated negative binomial model is the most appropriate.

[20] The key assumption of Cox regression is that the hazard function for each individual follows the same form, although we do not impose any shape for the form. We used scaled Schoenfeld residuals to examine proportionality. The detailed results of this test are presented in Appendix G. Non-significant coefficients indicate that the proportional hazards assumption is satisfied. For this model, our assumption appears to be justified. One of the control variables for jail time appeared to be nonproportional, but this is not a substantial cause for concern.

[21] We again tested a squared jail time term in this model, which was not statistically significant and did not improve model fit (OR = .99, $p \leq .630$; LR $\chi^2$ (1 d.f.) = .23, $p \leq .634$.

[22] Likelihood ratio test for Poisson vs. negative binomial: LR $\chi^2$ (1 d.f.) = 2946.78, $p < .001$. Likelihood ratio test for zero-inflated Poisson vs. zero-inflated negative binomial: LR $\chi^2$ (1 d.f.) = 304.47, $p < .0001$. Vuong test for zero-inflated vs. standard negative binomial: $z = 5.68$, $p < .0001$.

[23] Covariates in Figures 2.6 and 2.8 are held at the same values as they were for Figure 2.4.

[24] Likelihood ratio test comparing models with and without squared term: LR $\chi^2$ (1 d.f.) = 6.62, $p \leq .010$.

[25] Likelihood ratio test for Poisson vs. negative binomial: LR $\chi^2$ (1 d.f.) = 1041.73, $p < .001$. Likelihood ratio test for zero-inflated Poisson vs. zero-inflated negative binomial: LR $\chi^2$ (1 d.f.) = 90.93, $p < .0001$. Vuong test for zero-inflated vs. standard negative binomial: $z = 5.62$, $p < .0001$.

[26] See Appendix G for test of the assumptions of proportional hazards. We proceeded with the proportional hazards model despite some evidence for nonproportionality in one of the monthly jail indicator variables.

## Tables

Table 2.1: Sample Characteristics

| | Treatment (N=800) | Control (N=759) |
|---|---|---|
| % West region | 50.0 | 52.8 |
| % Male | 66.5 | 67.6 |
| % White | 41.8 | 37.9 |
| Mean age at RA date | 40.78 | 40.58 |
| Average household income in ZIP | | |
| % less than $20,000 | 7.6 | 8.8 |
| % $20,000 - $29,999 | 37.9 | 41.2 |
| % $30,000 - $39,999 | 33.9 | 31.6 |
| % $40,000 or more | 20.6 | 18.3 |
| Mean monthly offending rate 1 year pre-RA (any charged offense) | .13 | .13 |
| Mean monthly offending rate 1 year pre-RA (charged violent offenses) | .02 | .02 |
| Mean monthly offending rate 1 year pre-RA (charged drug offenses) | .03 | .04 |
| Mean number of months in jail 1 year post-RA | 1.1 | 1.1 |
| % with any charged offense 1 year post-RA | 16.0 | 15.0 |
| % with any charged offense 2 years post-RA | 21.5 | 21.5 |
| % with charged violent offense 1 year post-RA | 2.9 | 4.1 |
| % with charged violent offense 2 years post-RA | 4.5 | 5.4 |
| % with charged drug offense 1 year post-RA | 6.4 | 6.5 |
| % with charged drug offense 2 years post-RA | 8.9 | 10.1 |

No significant differences between groups at $p \leq .05$ ($\chi^2$ for proportions & 2-tailed $t$ for means).

Table 2.2: Prevalence of Recidivism (All Offenses, 2-Year Follow Up)

| Logistic Regression | | | Number of observations = 1,559 | |
|---|---|---|---|---|
| | | | Likelihood Ratio $\chi^2$ (10 d.f.) = 252.43 | |
| | | | Pr > $\chi^2$ = .000 | |
| Log likelihood = -685.013 | | | Pseudo $R^2$ = .156 | |
| **Term** | **Odds Ratio** | **S.E.** | **z** | ***p*** |
| Treatment group | 1.05 | .145 | .38 | .707 |
| West probation region | .60 | .098 | -3.15 | .002 |
| Male | 1.24 | .185 | 1.44 | .151 |
| White | 1.00 | .171 | -.03 | .978 |
| Age at RA | .98 | .007 | -3.03 | .002 |
| Income $20,000-$29,999 | .60 | .147 | -2.08 | .037 |
| Income $30,000-$39,999 | .52 | .135 | -2.54 | .011 |
| Income $40,000 or more | .25 | .076 | -4.56 | .000 |
| Monthly offending rate 1 year pre-RA (any charged offense) | .98 | .132 | -.15 | .882 |
| Months in jail post-RA | 1.29 | .025 | 13.06 | .000 |

Hosmer-Lemeshow $\chi^2$ for goodness-of-fit (8 d.f., 10 groups) = 9.84, $p \leq .276$

Table 2.3: Frequency of Recidivism (All Offenses, 2-Year Follow-Up)

| Zero-Inflated Negative Binomial Regression | | | Number of observations = 1,559 | |
|---|---|---|---|---|
| | | | Nonzero observations = 335 | |
| | | | Zero observations = 1,224 | |
| Inflation model = logit | | | Likelihood Ratio $\chi^2$ (9 d.f.) = 29.77 | |
| Log likelihood= -1700.398 | | | Pr > $\chi^2$ = .001 | |
| **Full Model** | **Incidence Rate Ratio** | **S.E.** | **z** | ***p*** |
| Treatment group | .89 | .148 | -.69 | .489 |
| West probation region | 1.41 | .296 | 1.61 | .107 |
| Male | 2.03 | .365 | 3.93 | .000 |
| White | 1.14 | .227 | .64 | .521 |
| Age at RA | .98 | .009 | -2.62 | .009 |
| Income $20,000-$29,999 | 1.53 | .464 | 1.41 | .160 |
| Income $30,000-$39,999 | 1.35 | .464 | .86 | .388 |
| Income $40,000 or more | 1.11 | .435 | .26 | .798 |
| Monthly offending rate 1 year pre-RA (any charged offense) | .99 | .196 | -.06 | .951 |
| Log of post-RA months at risk [offset] | | | | |
| **Inflated Model** | **b** | **S.E.** | **z** | ***p*** |
| Treatment group | -.067 | .152 | -.44 | .658 |
| West probation region | .494 | .187 | 2.64 | .008 |
| Male | -.104 | .171 | -.61 | .542 |
| White | .003 | .191 | .01 | .989 |
| Age at RA | .020 | .008 | 2.62 | .009 |
| Income $20,000-$29,999 | .569 | .307 | 1.85 | .064 |
| Income $30,000-$39,999 | .795 | .331 | 2.38 | .018 |
| Income $40,000 or more | 1.478 | .371 | 3.98 | .000 |
| Monthly offending rate 1 year pre-RA (any charged offense) | .199 | .186 | 1.07 | .287 |
| Constant | -4.027 | .510 | -7.90 | .000 |
| Log of post-RA months at risk [offset] | | | | |
| Ln(Alpha) | .779 | .162 | 4.82 | .000 |
| Alpha | 2.179 | .352 | | |

Vuong test of zero-inflated vs. standard negative binomial: z = 7.04, $p < .001$

Table 2.4: Time to Failure (All Offenses, 2-Year Follow-Up)

| Cox Regression (Breslow Method for Ties) | Number of subjects = 1,559 | | | |
|---|---|---|---|---|
| | Number of failures = 335 | | | |
| | Time at risk (person-days) = 976,440 | | | |
| | Likelihood Ratio $\chi^2$ (21 d.f.) = 298.04 | | | |
| Log likelihood = -2275.151 | Pr > $\chi^2$ = .000 | | | |
| **Term** | **Hazard Ratio** | **S. E.** | **z** | ***p*** |
| Treatment group | 1.03 | .116 | .28 | .777 |
| West probation region | .64 | .087 | -3.30 | .001 |
| Male | 1.24 | .152 | 1.76 | .078 |
| White | 1.04 | .146 | .27 | .790 |
| Age at RA | .98 | .006 | -3.03 | .002 |
| Income $20,000-$29,999 | .60 | .118 | -2.59 | .010 |
| Income $30,000-$39,999 | .49 | .103 | -3.42 | .001 |
| Income $40,000 or more | .29 | .071 | -5.01 | .000 |
| Monthly offending rate 1 year pre-RA (any charged offense) | .99 | .119 | -.09 | .929 |
| In jail Oct 2007 | 3.76 | 1.915 | 2.61 | .009 |
| In jail Nov 2007 | .60 | .522 | -.59 | .556 |
| In jail Dec 2007 | .77 | .947 | -.21 | .833 |
| In jail Jan 2008 | 1.76 | 1.938 | .52 | .605 |
| In jail Feb 2008 | .35 | .390 | -.94 | .346 |
| In jail Mar 2008 | 3.79 | 4.263 | 1.18 | .236 |
| In jail Apr 2008 | 2.37 | 1.381 | 1.49 | .137 |
| In jail May 2008 | .83 | .397 | -.38 | .701 |
| In jail Jun 2008 | 1.04 | .559 | .07 | .945 |
| In jail Jul 2008 | 1.46 | .686 | .80 | .422 |
| In jail Aug 2008 | .68 | .273 | -.95 | .340 |
| In jail Sep 2008 | 1.08 | .348 | .24 | .812 |

Table 2.5: Instrumental Variables Model: Prevalence of Recidivism (All Offenses, 2-Year Follow-Up)

| | **First Stage** Treatment Take-Up | **Reduced Form (ITT)** Post-RA Any Off. |
|---|---|---|
| | Observations = 1,559 | Observations = 1,559 |
| | $R^2$ = .663 | $R^2$ = .193 |
| | Adjusted $R^2$ = .659 | Adjusted $R^2$ = .183 |
| **Instruments** | **b (S.E.)** | **b (S.E.)** |
| Assigned LIS | .812 (.088)*** | -.047 (.113) |
| Assigned LIS*West | -.075 (.035)* | -.007 (.045) |
| Assigned LIS*Male | .014 (.032) | -.004 (.040) |
| Assigned LIS*White | -.024 (.037) | .029 (.048) |
| Assigned LIS*Age | -.000 (.001) | -.001 (.002) |
| Assigned LIS*Income20 | .026 (.057) | .111 (.073) |
| Assigned LIS*Income30 | .086 (.060) | .095 (.077) |
| Assigned LIS*Income40 | .028 (.066) | .122 (.085) |
| Assigned LIS*Prior offending | -.030 (.027)* | .032 (.035) |
| **Exogenous** Months in jail post-RA | -.014 (.002)*** | . 052 (.003)*** |
| West probation region | .055 (.025)* | -.068 (.032)* |
| Male | .009 (.023) | .031 (.029) |
| White | .009 (.027) | -.017 (.034) |
| Age at RA | .001 (.001) | -.002 (.001) |
| Income $20,000-$29,999 | -.002 (.039) | -.126 (.050)* |
| Income $30,000-$39,999 | -.013 (.042) | -.139 (.054)** |
| Income $40,000 or more | -.002 (.047) | -.245 (.060)*** |
| Monthly offending rate 1 year pre-RA (any offense) | -.012 (.022) | -.024 (.028) |
| Constant | -.025 (.064) | .408 (.081)*** |

* $p \leq .05$; ** $p \leq .01$; *** $p \leq .001$.

| Second-Stage Instrumental Variables Regression | | | Number of observations = 1,559 | |
|---|---|---|---|---|
| | | | Wald $\chi^2$ (10 d.f.) = 364.12 | |
| | | | Pr > $\chi^2$ = .000 | |
| Any New Charged Offense | | | $R^2$ = .189 | |
| **Term** | **b** | **S. E.** | **Z** | ***p*** |
| Predicted Treatment Take-up | .009 | .023 | .40 | .686 |
| Months in jail post-RA | .052 | .003 | 17.50 | .000 |
| Constant | .383 | .057 | 6.76 | .000 |

Controlling for region, gender, race, age, SES and offending history.

Table 2.6: Treatment Effects by Subgroup (All Offenses, 2-Year Follow Up)

| Subgroup | | N | Mean Pr. of Failure | *t/F*† | *p* |
|---|---|---|---|---|---|
| **Region** | Northeast | 339 | .223 | 3.85 | .000 |
| | West | 351 | .178 | | |
| **Gender** | Male | 464 | .214 | -3.29 | .001 |
| | Female | 226 | .172 | | |
| **Race** | White | 283 | .203 | -.47 | .641 |
| | Nonwhite | 407 | .198 | | |
| **Age at RA** | Under 25 | 34 | .241 | 4.92 | .001 |
| | 25-34 | 182 | .237 | | |
| | 35-44 | 176 | .187 | | |
| | 45-54 | 236 | .186 | | |
| | 55 + | 62 | .162 | | |
| **Income** | < $20,000 | 52 | .251 | 22.91 | .000 |
| | $20,000-$29,999 | 261 | .218 | | |
| | $30,000-$39,999 | 239 | .223 | | |
| | $40,000 + | 138 | .108 | | |
| **Offending rate per month at risk 1 year pre-RA** | No offending | 550 | .209 | 3.67 | .026 |
| | Up to 1 | 125 | .168 | | |
| | More than 1 | 15 | .173 | | |

† 2-sample *t*-test or one-way ANOVA (*F*-test).

Table 2.7: Prevalence of Recidivism (Violent Offenses, 2-Year Follow Up)

| Logistic Regression | | | Number of observations = 1,559 | |
|---|---|---|---|---|
| | | | Likelihood Ratio $\chi^2$ (10 d.f.) = 118.40 | |
| | | | Pr > $\chi^2$ = .000 | |
| Log likelihood = -247.480 | | | Pseudo $R^2$ = .193 | |
| **Term** | **Odds Ratio** | **S.E.** | **z** | ***p*** |
| Treatment group | .89 | .225 | -.46 | .644 |
| West probation region | .53 | .167 | -2.00 | .045 |
| Male | 3.53 | 1.263 | 3.52 | .000 |
| White | .93 | .297 | -.23 | .820 |
| Age at RA | .97 | .013 | -2.13 | .033 |
| Income $20,000-$29,999 | .46 | .195 | -1.83 | .067 |
| Income $30,000-$39,999 | .32 | .152 | -2.41 | .016 |
| Income $40,000 or more | .18 | .099 | -3.10 | .002 |
| Monthly offending rate 1 year pre-RA (charged violent off.) | 2.61 | 1.139 | 2.20 | .028 |
| Months in jail post-RA | 1.26 | .030 | 9.47 | .000 |

Hosmer-Lemeshow $\chi^2$ for goodness-of-fit (8 d.f., 10 groups) = 5.82, $p \leq .668$

Table 2.8: Frequency of Recidivism (Violent Offenses, 2-Year Follow-Up)

| Zero-Inflated Negative Binomial Regression<br><br>Inflation model = logit<br>Log likelihood = -499.228 | | | Number of observations = 1,559<br>Nonzero observations = 77<br>Zero observations = 1,482<br>Likelihood Ratio $\chi^2$ (9 d.f.) = 7.08<br>Pr > $\chi^2$ = .629 | |
|---|---|---|---|---|
| **Full Model** | **Incidence Rate Ratio** | **S.E.** | **z** | ***p*** |
| Treatment group | .83 | .205 | -.75 | .453 |
| West probation region | 1.18 | .389 | .49 | .624 |
| Male | .87 | .336 | -.35 | .724 |
| White | .99 | .331 | -.04 | .966 |
| Age at RA | .98 | .016 | -1.08 | .280 |
| Income $20,000-$29,999 | 1.53 | .717 | .90 | .366 |
| Income $30,000-$39,999 | 1.24 | .767 | .34 | .731 |
| Income $40,000 or more | 1.13 | .767 | .18 | .857 |
| Monthly offending rate 1 year pre-RA (charged violent off.) | .34 | .185 | -1.98 | .047 |
| Log of post-RA months at risk [offset] | | | | |
| **Inflated Model** | **b** | **S.E.** | **z** | ***p*** |
| Treatment group | .109 | .243 | .45 | .654 |
| West probation region | .444 | .293 | 1.52 | .130 |
| Male | -1.335 | .350 | -3.81 | .000 |
| White | -.031 | .301 | -.10 | .917 |
| Age at RA | .032 | .013 | 2.53 | .011 |
| Income $20,000-$29,999 | .522 | .415 | 1.26 | .209 |
| Income $30,000-$39,999 | .923 | .459 | 2.01 | .044 |
| Income $40,000 or more | 1.530 | .547 | 2.80 | .005 |
| Monthly offending rate 1 year pre-RA (charged violent off.) | -1.469 | .814 | -1.80 | .071 |
| Constant | -1.441 | .732 | -1.97 | .049 |
| Log of post-RA months at risk [offset] | | | | |
| Ln(Alpha) | -.089 | .268 | -.33 | .740 |
| Alpha | .915 | .245 | | |

Vuong test of zero-inflated vs. standard negative binomial: z = 5.68, $p < .001$.

Table 2.9: Time to Failure (Violent Offenses, 2-Year Follow-Up)

| Cox Regression (Breslow Method for Ties) | | | Number of subjects = 1,559 | |
|---|---|---|---|---|
| | | | Number of failures = 77 | |
| | | | Time at risk (person-days) = 1,103,248 | |
| | | | Likelihood Ratio $\chi^2$ (10 d.f.) = 105.11 | |
| Log likelihood = -511.625 | | | Pr > $\chi^2$ = .000 | |
| **Term** | **Hazard Ratio** | **S. E.** | **z** | ***p*** |
| Treatment group | .91 | .211 | -.40 | .686 |
| West probation region | .65 | .180 | -1.55 | .122 |
| Male | 3.24 | 1.108 | 3.44 | .001 |
| White | .97 | .279 | -.09 | .927 |
| Age at RA | .97 | .012 | -2.25 | .024 |
| Income $20,000-$29,999 | .45 | .175 | -2.05 | .040 |
| Income $30,000-$39,999 | .36 | .153 | -2.41 | .016 |
| Income $40,000 or more | .22 | .113 | -2.94 | .003 |
| Monthly offending rate 1 year pre-RA (charged violent off.) | 2.35 | .864 | 2.33 | .020 |
| In jail 1 year post-RA | 7.81 | 1.839 | 8.73 | .000 |

Table 2.10: Time to Failure with Jail-Time Interaction (Violent Offenses, 2-Year Follow-Up)

| Cox Regression (Breslow Method for Ties) | | | Number of subjects = 1,559 | |
|---|---|---|---|---|
| | | | Number of failures = 77 | |
| | | | Time at risk (person-days) = 1,103,248 | |
| | | | Likelihood Ratio $\chi^2$ (11 d.f.) = 112.10 | |
| Log likelihood = -508.130 | | | Pr > $\chi^2$ = .000 | |
| **Term** | **b** | **S. E.** | **z** | ***p*** |
| Treatment group | -.092 | .231 | -.40 | .691 |
| West probation region | -.425 | .277 | -1.53 | .126 |
| Male | 1.173 | .342 | 3.43 | .001 |
| White | -.025 | .287 | -.09 | .932 |
| Age at RA | -.027 | .012 | -2.23 | .025 |
| Income $20,000-$29,999 | -.791 | .391 | -2.02 | .043 |
| Income $30,000-$39,999 | -1.013 | .425 | -2.38 | .017 |
| Income $40,000 or more | -1.513 | .518 | -2.92 | .003 |
| Monthly offending rate 1 year pre-RA (charged violent off.) | .859 | .366 | 2.35 | .019 |
| In jail 1 year post-RA | 3.017 | .455 | 6.64 | .000 |
| Jail*time | -.004 | .001 | -2.51 | .012 |

Table 2.11: Instrumental Variables Model: Prevalence of Recidivism (Violent Offenses, 2-Year Follow-Up)

| | **First Stage** Treatment Take-Up | **Reduced Form (ITT)** Post-RA Violent Off. |
|---|---|---|
| | Observations = 1,559 | Observations = 1,559 |
| | $R^2$ = .661 | $R^2$ = .117 |
| | Adjusted $R^2$ = .658 | Adjusted $R^2$ = .107 |
| **Instruments** | **b (S.E.)** | **b (S.E.)** |
| Assigned LIS | .808 (.088)*** | -.073 (.062) |
| Assigned LIS*West | -.075 (.035)* | .005 (.025) |
| Assigned LIS*Male | .013 (.032) | .009 (.022) |
| Assigned LIS*White | -.021 (.037) | .048 (.026) |
| Assigned LIS*Age | -.000 (.001) | .000 (.001) |
| Assigned LIS*Income20 | .025 (.057) | .078 (.040) |
| Assigned LIS*Income30 | .083 (.060) | .060 (.043) |
| Assigned LIS*Income40 | .021 (.066) | .053 (.047) |
| Assigned LIS*Prior offending | -.004 (.099) | -.104 (.070) |
| **Exogenous** Months in jail post-RA | -.013 (.002)*** | .021 (.002)*** |
| West probation region | .056 (.025)* | -.028 (.017) |
| Male | .010 (.023) | .037 (.016)* |
| White | .009 (.027) | -.028 (.019) |
| Age at RA | .001 (.001) | -.001 (.001) |
| Income $20,000-$29,999 | -.001 (.039) | -.066 (.028)* |
| Income $30,000-$39,999 | -.014 (.042) | -.070 (.030)* |
| Income $40,000 or more | -.002 (.047) | -.085 (.033)** |
| Monthly offending rate 1 year pre-RA (violent offense) | -.069 (.082) | .133 (.058)* |
| Constant | -.029 (.063) | .131 (.045)** |

* $p \leq .05$; ** $p \leq .01$; *** $p \leq .001$.

| Second-Stage Instrumental Variables Regression | | | Number of observations = 1,559 | |
|---|---|---|---|---|
| | | | Wald $\chi^2$ (10 d.f.) = 192.81 | |
| | | | Pr > $\chi^2$ = .000 | |
| New Charged Violent Off. | | | $R^2$ = .110 | |
| **Term** | **b** | **S. E** | **z** | ***p*** |
| Predicted treatment take-up | -.007 | .013 | -.54 | .586 |
| Months in jail post-RA | .020 | .002 | 12.27 | .000 |
| Constant | .010 | .031 | 3.19 | .001 |

Controlling for region, gender, race, age, SES and offending history.

Table 2.12: Treatment Effect by Subgroups (Violent Offenses, 2-Year Follow Up)

| Subgroup | | N | Mean Pr. of Failure | *t/F*† | *p* |
|---|---|---|---|---|---|
| **Region** | Northeast | 339 | .045 | 3.09 | .002 |
| | West | 351 | .030 | | |
| **Gender** | Male | 464 | .053 | -9.76 | .000 |
| | Female | 226 | .006 | | |
| **Race** | White | 283 | .038 | .01 | .989 |
| | Nonwhite | 407 | .038 | | |
| **Age at RA** | Under 25 | 34 | .048 | 3.38 | .009 |
| | 25-34 | 182 | .050 | | |
| | 35-44 | 176 | .032 | | |
| | 45-54 | 236 | .034 | | |
| | 55 + | 62 | .025 | | |
| **Income** | < $20,000 | 52 | .059 | 12.13 | .000 |
| | $20,000-$29,999 | 261 | .044 | | |
| | $30,000-$39,999 | 239 | .042 | | |
| | $40,000 + | 138 | .011 | | |
| **Offending rate per month at risk 1 year pre-RA** | No violent offending | 667 | .037 | -1.79 | .073 |
| | Violent offending | 23 | .061 | | |

† 2-sample *t*-test or one-way ANOVA (*F*-test).

Table 2.13: Prevalence of Recidivism (Drug Offenses, 2-Year Follow Up)

| Logistic Regression | | | Number of observations = 1,559 | |
|---|---|---|---|---|
| | | | Likelihood Ratio $\chi^2$ (11 d.f.) = 125.97 | |
| | | | Pr > $\chi^2$ = .000 | |
| Log likelihood = -426.233 | | | Pseudo $R^2$ = .129 | |
| **Term** | **Odds Ratio** | **S.E.** | **z** | ***p*** |
| Treatment group | .92 | .169 | -.46 | .644 |
| West probation region | .88 | .195 | -.59 | .554 |
| Male | 1.93 | .420 | 3.00 | .003 |
| White | 1.18 | .275 | .72 | .473 |
| Age at RA | .97 | .009 | -3.13 | .002 |
| Income $20,000-$29,999 | .49 | .154 | -2.27 | .023 |
| Income $30,000-$39,999 | .43 | .147 | -2.47 | .013 |
| Income $40,000 or more | .27 | .108 | -3.29 | .001 |
| Monthly offending rate 1 year pre-RA (charged drug off.) | 1.55 | .694 | .99 | .324 |
| Months in jail post-RA | 1.58 | .165 | 4.42 | .000 |
| Months in jail post-RA(squared) | .98 | .009 | -2.65 | .008 |

Hosmer-Lemeshow $\chi^2$ for goodness-of-fit (8 d.f., 10 groups) = 4.96, $p \leq .762$

Table 2.14: Frequency of Recidivism (Drug Offenses, 2-Year Follow-Up)

| Zero-Inflated Negative Binomial Regression<br><br>Inflation model = logit<br>Log likelihood = -736.235 | | | Number of observations = 1,559<br>Nonzero observations = 148<br>Zero observations = 1,411<br>Likelihood Ratio $\chi^2$ (9 d.f.) = 15.07<br>Pr > $\chi^2$ = .089 | |
|---|---|---|---|---|
| **Full Model** | **Incidence Rate Ratio** | **S.E.** | **z** | ***p*** |
| Treatment group | .78 | .143 | -1.38 | .169 |
| West probation region | 1.42 | .311 | 1.60 | .111 |
| Male | 1.53 | .379 | 1.72 | .086 |
| White | .90 | .212 | -.45 | .655 |
| Age at RA | .98 | .009 | -2.01 | .045 |
| Income $20,000-$29,999 | .99 | .315 | -.02 | .981 |
| Income $30,000-$39,999 | .91 | .325 | -.27 | .785 |
| Income $40,000 or more | .57 | .244 | -1.31 | .189 |
| Monthly offending rate 1 year pre-RA (charged drug off.) | 1.10 | .668 | .16 | .873 |
| Log of post-RA months at risk [offset] | | | | |
| **Inflated Model** | **b** | **S.E.** | **z** | ***p*** |
| Treatment group | .032 | .192 | .16 | .870 |
| West probation region | .148 | .230 | .64 | .520 |
| Male | -.589 | .236 | -2.49 | .013 |
| White | -.242 | .243 | -.99 | .320 |
| Age at RA | .029 | .010 | 2.92 | .004 |
| Income $20,000-$29,999 | .512 | .326 | 1.57 | .116 |
| Income $30,000-$39,999 | .716 | .357 | 2.01 | .045 |
| Income $40,000 or more | 1.071 | .427 | 2.51 | .012 |
| Monthly offending rate 1 year pre-RA (charged drug off.) | -.009 | .545 | -.02 | .987 |
| Constant | -2.477 | .563 | -4.40 | .000 |
| Log of post-RA months at risk [offset] | | | | |
| Ln(Alpha) | -.473 | .331 | -1.43 | .153 |
| Alpha | .623 | .206 | | |

Vuong test of zero-inflated vs. standard negative binomial: z = 5.62, $p < .001$

Table 2.15: Time to Failure (Drug Offenses, 2-Year Follow-Up)

| Cox Regression (Breslow Method for Ties) | | | Number of subjects = 1,559 | |
|---|---|---|---|---|
| | | | Number of failures = 148 | |
| | | | Time at risk (person-days) = 1,071,848 | |
| | | | Likelihood Ratio $\chi^2$ (21 d.f.) = 142.68 | |
| Log likelihood = -1009.539 | | | Pr > $\chi^2$ = .000 | |
| **Term** | **Hazard Ratio** | **S. E.** | **z** | ***p*** |
| Treatment group | .93 | .157 | -.42 | .673 |
| West probation region | .87 | .178 | -.70 | .483 |
| Male | 1.96 | .399 | 3.32 | .001 |
| White | 1.14 | .243 | .63 | .529 |
| Age at RA | .97 | .009 | -3.40 | .001 |
| Income $20,000-$29,999 | .44 | .127 | -2.85 | .004 |
| Income $30,000-$39,999 | .40 | .122 | -3.01 | .003 |
| Income $40,000 or more | .26 | .093 | -3.77 | .000 |
| Monthly offending rate 1 year pre-RA (charged drug off.) | 1.46 | .595 | .93 | .353 |
| In jail Oct 2007 | 2.42 | 2.445 | .88 | .382 |
| In jail Nov 2007 | 1.13 | 1.599 | .08 | .934 |
| In jail Dec 2007 | 1.94 | 2.750 | .47 | .641 |
| In jail Jan 2008 | .22 | .308 | -1.08 | .282 |
| In jail Feb 2008 | 2.33 | 3.319 | .60 | .552 |
| In jail Mar 2008 | 1.72 | 2.130 | .44 | .661 |
| In jail Apr 2008 | 2.03 | 1.640 | .88 | .378 |
| In jail May 2008 | .99 | .628 | -.01 | .991 |
| In jail Jun 2008 | .55 | .428 | -.76 | .445 |
| In jail Jul 2008 | 2.84 | 1.930 | 1.54 | .124 |
| In jail Aug 2008 | .48 | .257 | -1.37 | .171 |
| In jail Sep 2008 | .86 | .395 | -.33 | .744 |

Table 2.16: Instrumental Variables Model: Prevalence of Recidivism (Drug Offenses, 2-Year Follow-Up)

| | **First Stage** Treatment Take-Up | **Reduced Form (ITT)** Post-RA Drug Off. |
|---|---|---|
| | Observations = 1,559 | Observations = 1,559 |
| | $R^2$ = .662 | $R^2$ = .103 |
| | Adjusted $R^2$ = .658 | Adjusted $R^2$ = .093 |
| **Instruments** | **b (S.E.)** | **b (S.E.)** |
| Assigned LIS | .803 (.088)*** | -.035 (.085) |
| Assigned LIS*West | -.077 (.035)* | .009 (.034) |
| Assigned LIS*Male | .011 (.032) | .037 (.030) |
| Assigned LIS*White | -.021 (.037) | .015 (.036) |
| Assigned LIS*Age | -.000 (.001) | .000 (.001) |
| Assigned LIS*Income20 | .024 (.057) | .018 (.055) |
| Assigned LIS*Income30 | .082 (.060) | -.008 (.058) |
| Assigned LIS*Income40 | .021 (.066) | .013 (.064) |
| Assigned LIS*Prior offending | .193 (.092)* | .022 (.089) |
| **Exogenous** Months in jail post-RA | -.013 (.002)*** | .026 (.002)*** |
| West probation region | .056 (.025)* | -.017 (.024) |
| Male | .009 (.023) | .028 (.022) |
| White | .010 (.027) | .004 (.026) |
| Age at RA | .001 (.001) | -.002 (.001)* |
| Income $20,000-$29,999 | -.001 (.040) | -.063 (.038) |
| Income $30,000-$39,999 | -.013 (.042) | -.060 (.040) |
| Income $40,000 or more | -.002 (.047) | -.104 (.045)* |
| Monthly offending rate 1 year pre-RA (drug offense) | -.025 (.054) | .023 (.052) |
| Constant | -.028 (.063) | .205 (.061)*** |

| Second-Stage Instrumental Variables Regression | | | Number of observations = 1,559 | |
|---|---|---|---|---|
| | | | Wald $\chi^2$ (10 d.f.) = 176.88 | |
| | | | Pr > $\chi^2$ = .000 | |
| New Charged Drug Offense | | | $R^2$ = .102 | |
| **Term** | **b** | **S. E** | **z** | ***p*** |
| Predicted treatment take-up | -.010 | .018 | -.60 | .550 |
| Months in jail post-RA | .026 | .002 | 11.67 | .000 |
| Constant | .191 | .043 | 4.50 | .000 |

Controlling for region, gender, race, age, SES and offending history.

Table 2.17: Treatment Effects by Subgroups (Drug Offenses, 2-Year Follow Up)

| Subgroup | | N | Mean Pr. of Failure | *t/F*† | *p* |
|---|---|---|---|---|---|
| **Region** | Northeast | 339 | .083 | 1.42 | .156 |
| | West | 351 | .074 | | |
| **Gender** | Male | 464 | .095 | -7.97 | .000 |
| | Female | 226 | .045 | | |
| **Race** | White | 283 | .083 | -1.11 | .268 |
| | Nonwhite | 407 | .076 | | |
| **Age at RA** | Under 25 | 34 | .110 | 9.86 | .000 |
| | 25-34 | 182 | .104 | | |
| | 35-44 | 176 | .071 | | |
| | 45-54 | 236 | .068 | | |
| | 55 + | 62 | .048 | | |
| **Income** | < $20,000 | 52 | .116 | 12.87 | .000 |
| | $20,000-$29,999 | 261 | .083 | | |
| | $30,000-$39,999 | 239 | .085 | | |
| | $40,000 + | 138 | .045 | | |
| **Offending rate per month at risk 1 year pre-RA** | No drug offending | 607 | .079 | .133 | .894 |
| | Drug offending | 83 | .078 | | |

† 2-sample *t*-test or one-way ANOVA (*F*-test).

## Figures

Figure 2.1: Risk-Based Allocation Strategies in the Philadelphia APPD

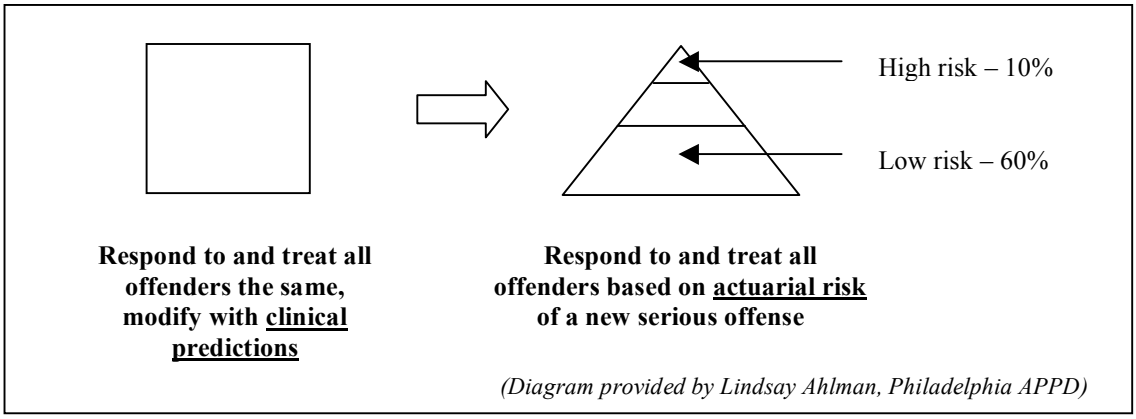



Figure 2.2: Case Flow Chart for the Philadelphia APPD Low Risk Experiment

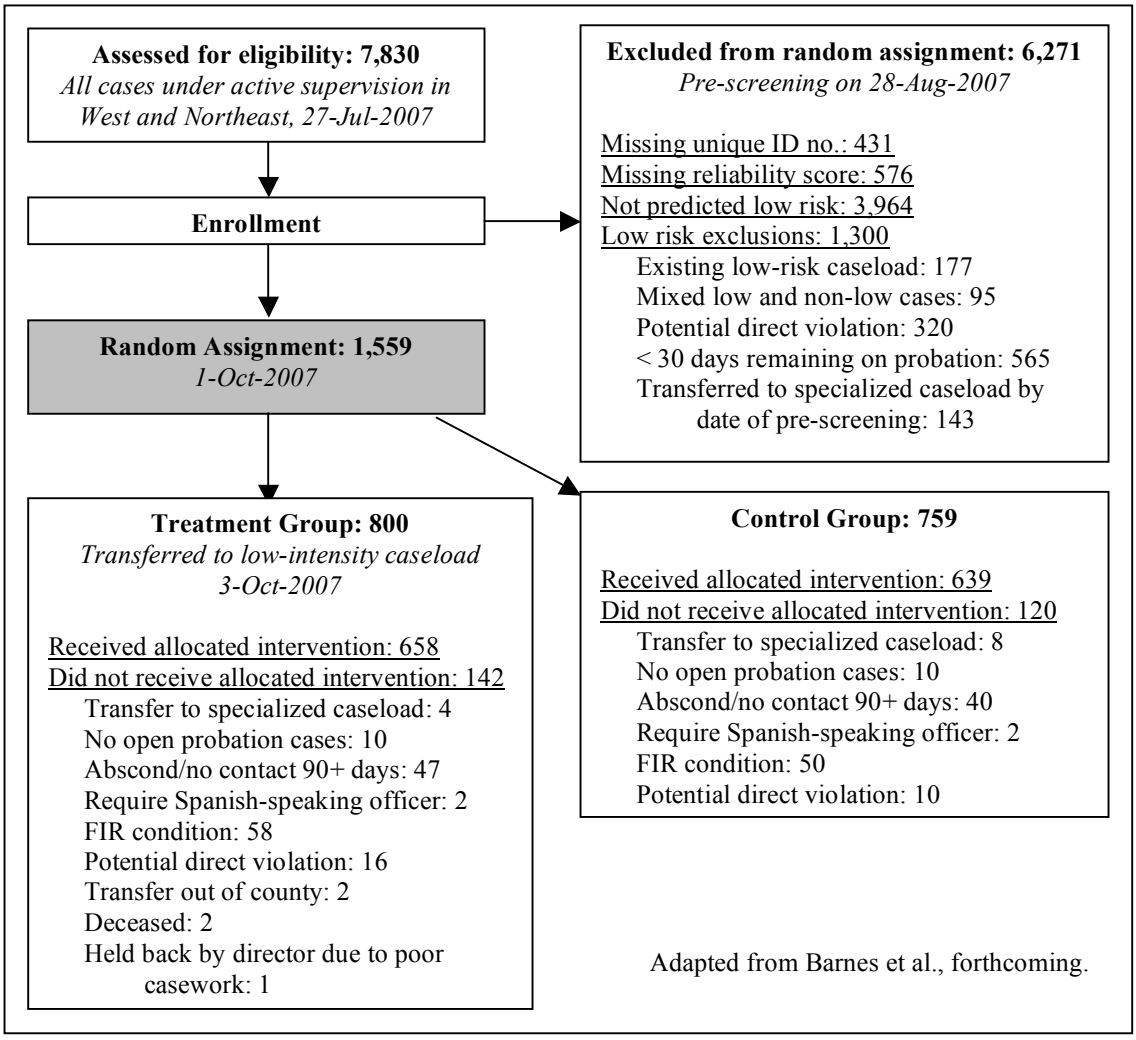


Figure 2.3: Survival Time for LIS Experiment Participants, by Assigned Treatment (All Offenses, 2-Year Follow-Up)

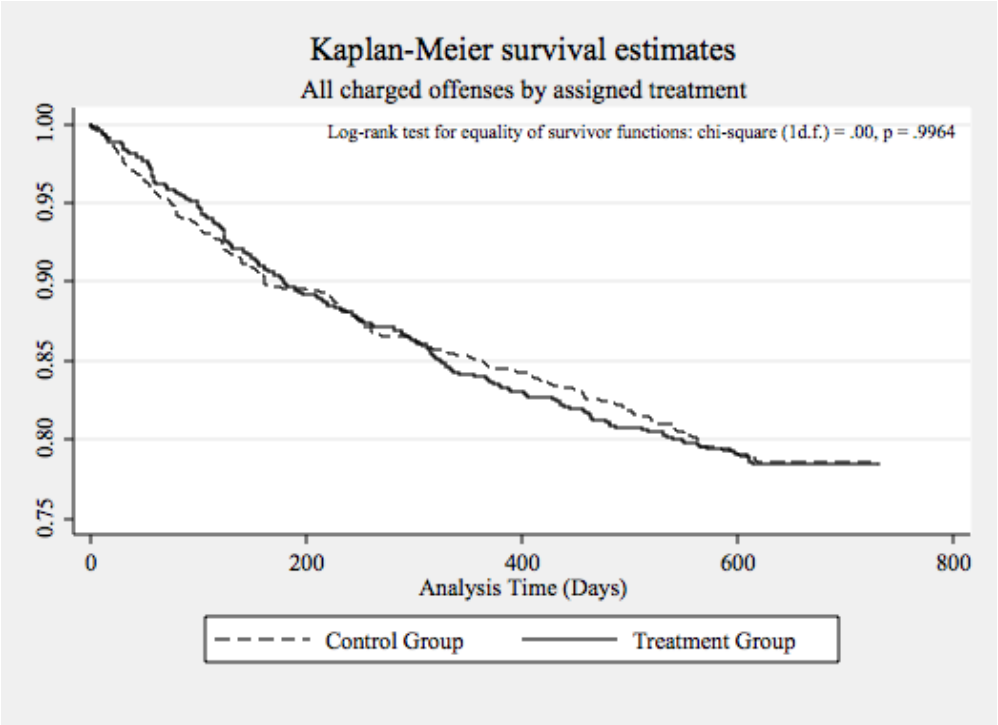



Figure 2.4: Cox Proportional Hazards Survivor Function by Assigned Treatment (All Offenses, 2-Year Follow-Up)

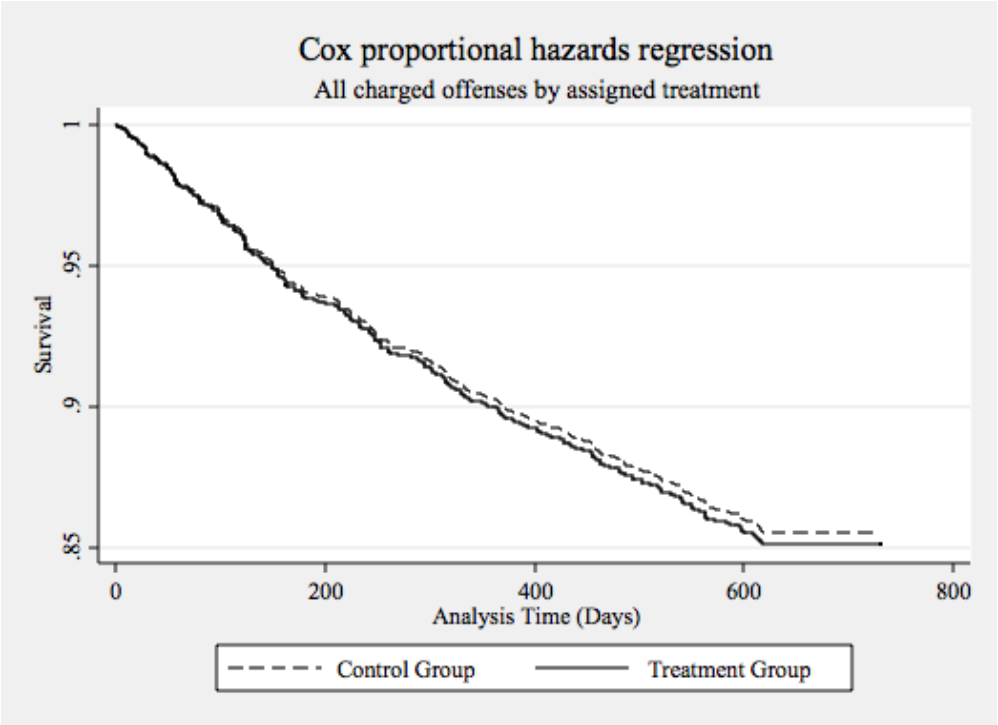


Figure 2.5: Survival Time for LIS Experiment Participants, by Assigned Treatment (Violent Offenses, 2-Year Follow-Up)

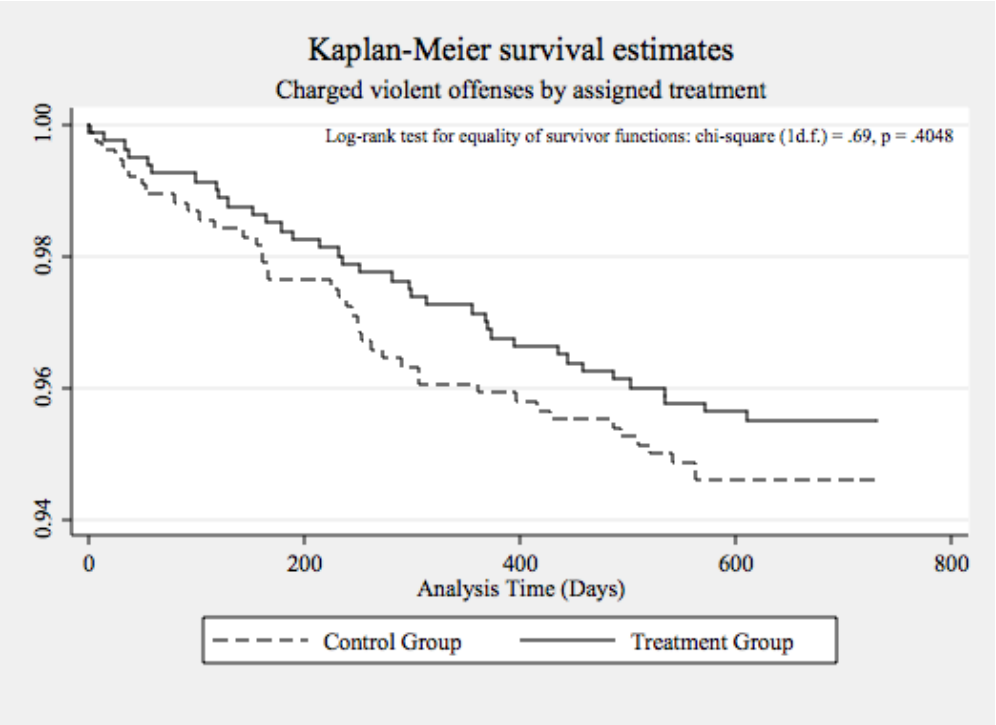



Figure 2.6: Cox Proportional Hazards Survivor Function (Violent Offenses, 2-Year Follow-Up)

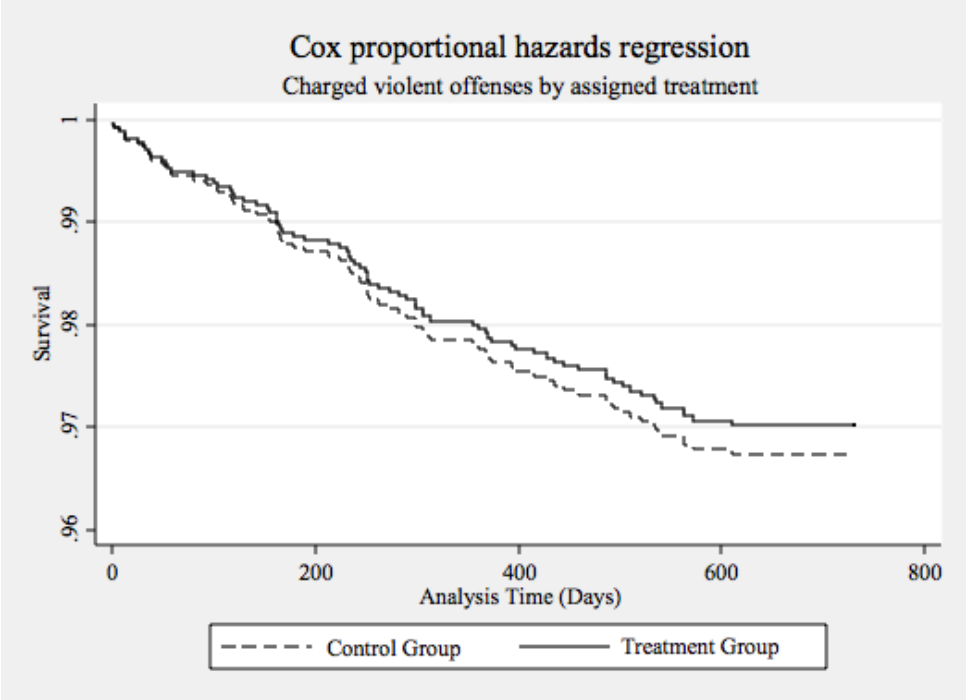


Figure 2.7: Survival Time for LIS Experiment Participants, by Assigned Treatment (Drug Offenses, 2-Year Follow-Up)

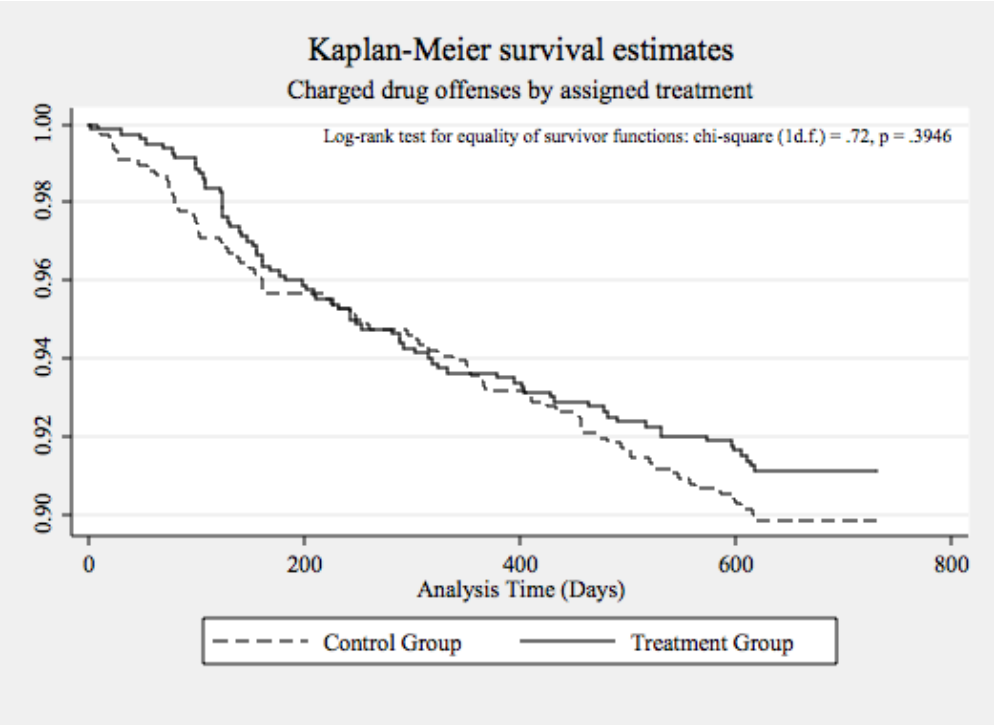
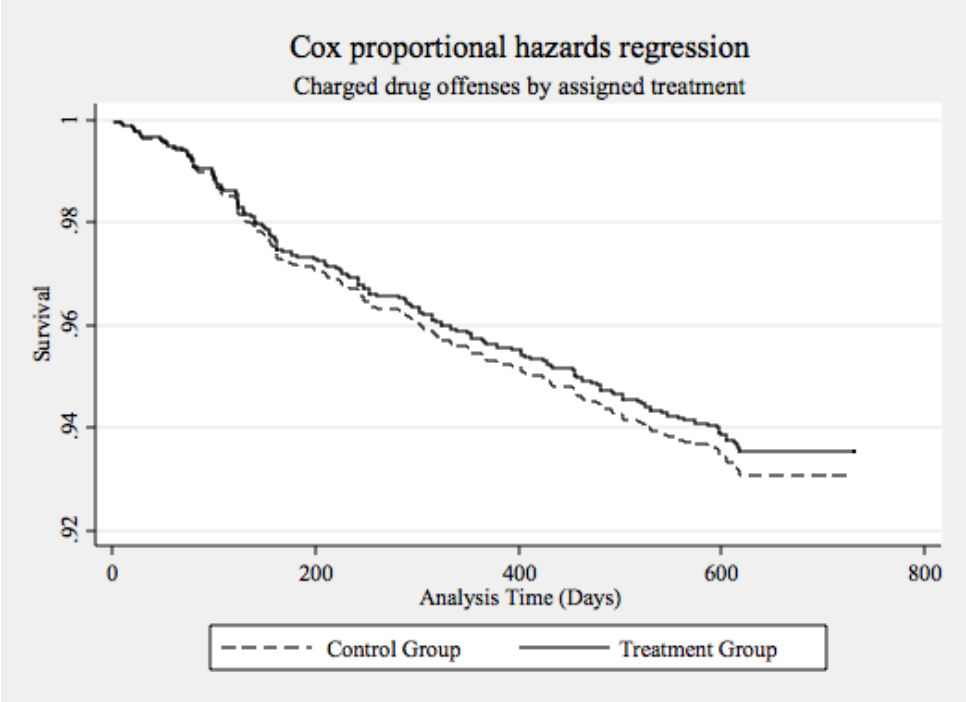



Figure 2.8: Cox Proportional Hazards Survivor Function (Drug Offenses, 2-Year Follow-Up)



# CHAPTER 3. Risk Prediction for Effective Offender Management: Patterns of Offending Severity among Probationers.

## Introduction

Probation in the early twentieth century was generally used as a disposition for first-time or minor offenders, while felons and recidivists tended to receive parole after incarceration. By mid-century, the use of probation expanded as interest in rehabilitation and community corrections increased (Clear & Braga, 1995). The backlash against community sentences and the pervasive retributive attitude to punishment since the 1970s has continued to pressurize probation agencies, as many struggle to deal with supervising parolees on their release from prison alongside offenders on more intensive probation programs. Furthermore, probation is now frequently used in addition to, rather than in place of, a jail or prison sentence (Ruth & Reitz, 2003). The complementary use of probation and prison may increase the number of serious offenders who come under the supervision of probation agencies, placing a strain on their limited resources. Figures from Philadelphia's Adult Probation and Parole Department (APPD) indicate the extent of the problem: in 2006, over 22 per cent of murder arrestees and 16 per cent of homicide victims in the city were under community supervision (Berk et al., 2009, p. 192).

Sherman (2007, p. 843) has argued that in the absence of short-term solutions to funding problems, probation agencies should focus their efforts on these most serious cases, "perform[ing] triage on their caseload to concentrate scarce resources on homicide prevention" at the expense of closely supervising offenders who pose little threat of harm

to society. This is the ultimate purpose behind the Philadelphia APPD Low Risk Experiment, which aims to identify and divert the lowest-level offenders in the agency into a large caseload with reduced ('low-intensity') supervision to allow probation officers to work more closely with serious offenders.

The implementation of a low-intensity model of probation supervision for low-risk offenders is dependent on the development of a reliable method of predicting the risk of serious recidivism. First and foremost, such a method must ensure that individuals at the *highest* risk of committing a serious crime do not receive less supervision than they need. However, if less severe offenders are included, probation officers will continue to be overwhelmed. Furthermore, research has suggested that probation supervision that is too 'intense' (in terms of frequency of contact and time spent with probation officers) may increase recidivism among offenders at the lowest risk level (e.g., Erwin, 1986; Hanley, 2006). Assigning low-risk offenders to excessively intensive programs might provoke defiant reactions (Sherman, 1993), reinforce delinquent attitudes and behavior, and disrupt pro-social networks and opportunities such as family ties and employment by requiring too much intervention from the criminal justice system. All of these factors may have unfavorable effects on future offending (Lowenkamp & Latessa, 2004). Following Sherman's (2007) proposition, then, the definition of 'serious' must incorporate only those crimes that represent the greatest threat to public safety.

The aim of this paper is to provide an in-depth analysis of the risk prediction and supervision strategies described above. We examine actual serious recidivism outcomes for a sample of probationers predicted by the statistical model used in the Philadelphia experiment to be at low risk of committing the most serious crimes. We examine the

sensitivity of the selected threshold for determining 'low risk' according to the model by contrasting the serious reoffending outcomes of offenders predicted to be low risk to those of offenders not predicted to be low risk. We also examine the model's sensitivity to different definitions of offending severity beyond the original substantive definitions it was designed to predict. Finally, we examine whether the intensity of supervision affects the relationship between past and future serious offending by comparing low risk experimental participants randomly assigned to low intensity supervision with those subject to standard reporting requirements.

## Risk Prediction in Offender Management

The prediction and assessment of risk is a long-standing concern of criminological theory and research. While it may not be possible (or indeed ethical) to precisely predict and act against those who will commit crimes in the future, the policy and practice implications of understanding the risk factors of crime and how they relate to offender management and crime prevention are obvious.

The history of risk prediction in criminal justice dates back at least to the first half of the twentieth century, when criminologists such as Burgess (1928) and the Gluecks (1950) developed simple predictive models based on checklists of risk and protective factors. Burgess, for example, combined unweighted predictors considered by expert opinion to be related to parole outcomes. More recently, risk prediction has been refined by extensive research on the factors that contribute favorably or unfavorably to offending behavior, many of which are now strongly confirmed by numerous studies and meta-

analyses (see e.g., Andrews, 1989; Farrington, 1998; Lipsey & Derzon, 1998). Risk factors for crime may be classified as static (characteristics that do not change or change in only one direction, such as age or age at first arrest), or dynamic (measurements of change in the offender, such as attitudes and employment) (Bonta, 2002).

The prediction of risk based on such variables falls into two distinct categories: clinical and actuarial/statistical. Clinical prediction is based on subjective human judgment and experience, informal consideration, and discussions with others. Statistical risk prediction is formal, objective, structured, quantitative, and grounded in theory, research and empiricism (Grove & Meehl, 1996; Bonta, 2002; Gottfredson & Moriarty, 2006). Clinical prediction is frequently used in criminal justice agencies, although statistical methods have also been introduced, largely from the psychology/psychometric disciplines. The statistical methods used by probation and other criminal justice agencies are generally inventories based on cognitive, emotional, and social development. Some widely-used and well-validated examples are the Psychopathy Checklist-Revised (PCL-R); the Level of Service Inventory-Revised (LSI-R), which is frequently used for risk management in probation settings; and the Violence Risk Appraisal Guide (VRAG) (Bonta, 2002); and the Offender Group Reconviction Scale (OGRS), which was developed for probation risk management in the United Kingdom (Howard et al., 2009).

Although the measurement and prediction of risk has always been a crucial part of criminological theory and practice, risk has been brought to the forefront of crime prevention programming with the development of the 'principles of effective intervention' (PEI: Andrews, Bonta, & Hoge, 1990). The PEI set out the optimal circumstances and considerations for providing effective correctional treatment.

Programs should be designed to adhere to three core principles: risk, need, and responsivity. The principles are interlinked: programs must be *responsive* to offenders' specific *risk* factors and *needs*. The risk principle is the most widely researched and validated of the PEI. Its key implication is that high-intensity interventions should be reserved for high risk, high need offenders. Thus, the risk principle corresponds to an earlier paradigm for correctional treatment set out in the criminological literature: the principle of the 'least restrictive alternative' (Rubin, 1975), which posits that punishment should be as unobtrusive as possible – no more than the minimum level needed to manage offenders' behavior.

The risk principle clearly highlights a need to develop effective and reliable risk prediction instruments in order to identify low- and high-risk offenders and direct them to appropriate sentences, supervision, and treatment. This is important not only in ensuring that offenders receive suitable services, but also in developing effective resource management in criminal justice agencies. Probation and parole agencies are a classic example of an environment in which good risk assessment is crucial. The growing use of probation and parole in the last two decades (e.g., Glaze & Bonczar, 2009), coupled with a crisis in funding (Petersilia, 1997), has led to large caseloads and limited ability of probation officers to supervise offenders appropriately. In Philadelphia, average caseloads can be as large as 150 to 200 clients to one officer (Berk et al., 2009). This is not unusual, or particularly new: similar standard caseload sizes were reported in the California probation agencies included in the RAND Corporation's intensive probation experiments in the 1980s (Petersilia & Turner, 1990). In the absence of short-term solutions to funding difficulties, risk assessment is needed to identify the most serious

offenders and focus the most intensive supervision on them, rather than treating them in the same way as low-risk offenders (Sherman, 2007). Berk et al. (2009) argue that both false positives (offenders incorrectly predicted to be low-risk) and false negatives (offenders incorrectly predicted to be high-risk) are detrimental to the effective operation of the criminal justice system, as well as public safety. False negatives drive up prison populations, leading to overcrowding and financial pressures as well as the social consequences for offenders, their families and communities. False positives lead to tragic crimes (such as the fatal shootings of several Philadelphia police officers in 2008 by paroled felons that lead to a moratorium on parole releases in Pennsylvania: Pennsylvania Fraternal Order of Police, 2008) and undermine public and political confidence in the criminal justice system. Thus, reliable risk prediction is vital for ensuring that as many offenders as possible are correctly classified and managed.

There is considerable evidence that statistical risk prediction is superior to clinical methods in criminal justice (see Grove & Meehl, 1996 for a detailed review; also Van Voorhis & Brown, 1997; Lowenkamp, Holsinger, & Latessa, 2001). For example, one study of violent recidivism among mentally ill offenders indicated predictive correlation coefficients of .09 for clinical predictions and .30 for statistical predictions. Similarly, for sex offender recidivism, the correlation coefficient was .10 for clinical predictions and .46 for statistical instruments (reported in Bonta, 2002). Gottfredson and Moriarty (2006) cite numerous studies supporting their contention that statistical predictions outperform clinical predictions in almost all situations involving human decision-making. They argue that humans do not use information reliably: in particular, we are poor at considering base rates, easily influenced by spurious causation, and do not systematically

weight information in an appropriate manner. Furthermore, clinical prediction is not standardized. Bonta (1996) notes that clinical decision rules are not easily observable or replicable.

This is not to say that clinical predictions are not useful, nor that statistical prediction methods are without flaws. Statistical predictions by definition provide average results, and the experience of criminal justice professionals plays an important role in highlighting deviations from the mean. In fact, professional override is the little-discussed fourth 'principle of effective intervention' (Andrews, Bonta, & Hoge, 1990). Some critics of statistical models have also pointed out the ethical concerns about imposing such impersonal judgments on offenders with individualized needs (see Gottfredson & Jarjoura, 1996, for a review of the criticisms of statistical models). One important and much-discussed ethical concern about risk prediction highlighted by Gottfredson and Jarjoura is that many risk factors of crime are highly correlated with race. This makes agency staff fearful of taking them into account in decision making. However, Gottfredson and Moriarty (2006) defend statistical risk assessment, noting its importance at the "nexus of research and practice," and pointing out that: "Properly developed and implemented, risk assessment devices can impose criminal justice decision making, properly target and potentially save resources, and potentially increase the public safety" (p. 195). Gottfredson and Jarjoura (ibid.) also set out a solution for reducing the bias in risk assessment. Although the role of professional override cannot be discounted, both papers argue that ignoring important predictive variables because of ethical concerns, rather than investigating how prediction instruments can empirically deal with these difficult issues, severely limits the utility of predictive devices and thus

does not contribute to the development of good practice in offender and resource management. Furthermore, clinical decision-making is not immune to these problems either. Bridges and Steen (1998) showed that probation officers' clinical judgments about the causes of offending can be influenced by the client's race, and may subsequently factor into case planning and sentencing recommendations.

The Philadelphia APPD and the University of Pennsylvania have developed a new risk prediction model that is the focus of the present study. The following section describes the model, its contribution to the existing body of literature on risk assessment, and its use in a practical setting.

## The Philadelphia APPD Low Risk Model and Supervision Experiment

The foundation for the Philadelphia APPD low risk prediction model and supervision experiment was laid in 2005, when APPD and the University of Pennsylvania began working together to restructure the agency's probation supervision practice according to predicted risk of serious crime. This approach represented a change from the existing standard supervision model for all offenders that was modified on an ad hoc basis largely according to officer discretion. In accordance with the risk principle, APPD's eventual goal was to reallocate the highest risk offenders to more intensive supervision, with a small ratio of clients to officers so that more time could be put into assessing and addressing those clients' needs. In order to do this without spending scarce resources on new staff, the lowest risk offenders in the agency needed to be assigned to large caseloads with minimal supervision. The first step in this process was to create a

statistical model that would predict the risk of serious reoffending[1] so that the whole APPD population could be stratified by risk level.

The risk prediction model

The statistical model used to forecast the risk of serious offending is described in full in Berk et al. (2009). Random forests methods were applied to a dataset of all probation and parole cases in Philadelphia between 2002 and 2004, containing only the data available to probation officers at intake,[2] to predict the risk of being charged with a new serious crime within two years of the probation or parole case start date. Random forests is a statistical learning procedure that forecasts outcomes by aggregating results from multiple classification and regression trees. The model was designed to stratify the population according to APPD's operational needs, with the assumption that the majority of the caseload was at low risk of serious recidivism and thus appropriate for low-intensity supervision. At the agency's request, 61 per cent of cases were to be deemed low risk, with the remainder either high risk (approximately 10 per cent) or neither low nor high (approximately 30 per cent) (Fig. 3.1). APPD also deemed the proportions of false positives and false negatives expected in the final model to be operationally acceptable. The proportion of false positives (offenders erroneously identified as low-risk) was set at 5 per cent, and the proportion of false negatives (offenders erroneously identified as high risk) was 20 per cent. A higher false negative rate was accepted given the lesser public safety concerns around this type of error. The initial 2002-2004 probation dataset is described as a "training sample," which is used to ensure the

independence of data (i.e., the model parameters were not derived from the same sample for which predictions would then be made). This helps to determine whether the relationships found in the initial sample are generalizable to other members of the same population. Once the model has been specified, it may then be used to derive risk predictions for probationers in current caseloads.

The model assigns each probation case (not each offender) a 'reliability score' to indicate its risk level. The reliability score is a value between 0 and 1. The selected threshold for low-risk cases was 0.5, so that cases with a reliability score greater than 0.5 were designated as low risk and scores equal to or less than 0.5 were not low risk. A specific offender's risk score is based on the average reliability score across all his or her active probation cases. However, even if the average reliability score exceeded 0.5, an offender could not be designated low risk if any one of his or her active cases scored 0.5 or below.

The Philadelphia model meets many of the recommendations set out by Bonta (2002) and Gottfredson and Moriarty (2006) for optimal risk prediction. Bonta suggests that risk assessments require predictive validity, direct relevance to criminal behavior and the correctional setting, and should adhere to the principle of the least restrictive alternative. The Philadelphia model is validated by its development on a training sample of cases for which outcomes were already known and its application to other members of the same population. The present paper attempts to further validate its ability to predict who will be low risk. The model's focus on serious offending is directly relevant to correctional priorities. Finally, the express purpose of the model is to ensure that the most intensive supervision and treatment is only reserved for those who need it most.

The Philadelphia model differs from other risk assessment and prediction instruments in several ways. Its focus on only the most serious offenses is its primary distinguishing feature. Berk et al. (2009) argue that it is not helpful for the purposes of effective resource allocation to predict any type of reoffending, as many existing prediction instruments do. There are both operational and political advantages to focusing only on the most extreme cases. Furthermore, the model uses charges rather than convictions as the outcome measure. While all recidivism outcome measures have well-documented advantages and disadvantages, charges are appropriate in this context because serious crimes are more likely to be pursued, but they do not all result in a conviction, often because of issues such as witness intimidation, which is a significant problem in Philadelphia (Berk et al., 2009, p. 194).

Although the Philadelphia model is clearly statistical, it also respects the clinical decision making processes that are used by probation intake officers. The model uses only information routinely available to intake officers (demographic characteristics and criminal history) and already used by probation officers when making clinical judgments. While this may not fully assuage the ethical concerns about statistical prediction, it cannot be said that the model imposes constraints on decision making beyond standard practice. It simply makes these processes more transparent and replicable.

The Low-Intensity Supervision Experiment

The Philadelphia APPD Low Risk Experiment ran from October 2007 to October 2008. It tested the hypothesis that low-intensity supervision (LIS) would not cause a harmful increase in recidivism for low-risk offenders compared to APPD's existing

supervision model, or 'supervision as usual' (SAU). Under SAU, offenders were supervised in regional units based on their residence, unless they were ordered by the court or APPD to be supervised in a specialist unit (e.g., sex offender or mental health units). Offenders in APPD's active caseload who were previously assigned to the West or Northeast regional units and were predicted to be low risk according to the prediction model described above were randomly assigned to LIS or SAU. In total, 1,559 offenders were randomly assigned: 800 to the LIS (treatment) group (400 from each region), and 759 to the control group (401 in the West and 358 in the Northeast).[3]

Probation clients assigned to the treatment group were placed in a caseload of four hundred. Two probation officers handled the entire low-intensity caseload. Probationers received only one office visit every six months, with telephone reporting appointments every six months and approximately halfway between office visits (see Appendix E for full details of the LIS model). They were returned to standard supervision if they were arrested for a new crime, because the LIS probation officers' caseloads were too large to handle the extra work required to process these cases. The experiment followed an 'intention-to-treat' analysis (Montori & Guyatt, 2001), so those offenders who were randomly assigned but could not be supervised in the low-intensity caseload due to failure or other operational issues were analyzed in their assigned groups rather than according to the type of supervision they actually received. In order to maintain the integrity of the LIS model, LIS officers' caseloads were kept at 400 by topping them up with so-called 'backfill' cases: offenders from the existing APPD caseload who were also predicted to be low-risk but were not part of the random

assignment pool. Backfill probationers were not included in the analysis of the main results of the experiment.

SAU for the control group usually consisted of monthly office visits, although the frequency could be increased or decreased at the probation officer's discretion for reasons relating to compliance or time left on the probation term. They continued regular appointments with their usual probation officer and no part of their supervision changed as a result of their experimental status. Probationers and probation officers were not informed of their status. Caseloads in this group were still large enough (approximately 145 clients per officer) that the content of meetings was essentially the same in both the treatment and control groups. However, control group offenders saw their probation officers more frequently.

Treatment group cases received approximately 45 per cent fewer contacts than they had in the year prior to random assignment, while the amount of contact in the control group did not change. Control group offenders received approximately twice as many contacts as treatment group offenders. The experimental protocol called for three control group contacts to every one in the treatment group, or six to one in terms of face-to-face contacts (assuming monthly office-based contacts in the control group), so although this standard was not quite achieved, the treatment group still received lower-intensity supervision. No significant differences in recidivism were found between the treatment and control groups after one year. Sixteen per cent of the treatment group and 15 per cent of the control group were charged with a new offense of any type ($p \leq .593$). Thus, it appeared that LIS did not lead to more crime compared to SAU, and was

therefore a safe strategy for restructuring probation supervision according to APPD's plans (Barnes et al., forthcoming).

## The Present Study

While the results of the Low Risk Experiment are promising, this summary does not provide a full picture of the predictive power of the model or the severity of offending in our sample. Despite the compelling evidence in favor of statistical risk prediction, it can never be an error-free endeavor. As Grove and Meehl (1996) note: "[T]he statistics furnish us with the probabilities so far as anything can" (p. 306). Actuarial prediction for correctional policy provides neither individualized predictions nor actual outcomes for a given person. The Philadelphia model identifies low-risk offenders in part based on their prior history, but it is entirely possible that offenders with a criminal history considered to be 'serious' on some basis could have been assigned a low-risk prediction, based either on the balance of other factors or varying definitions of offending severity. It is also possible that low-intensity supervision could lead to an escalation in offending severity as it becomes less likely that probation officers will pick up and act on violations and transgressions (and as offenders begin to realize this). These are the cases that get picked up by the public, politicians, and the media, and serve to undermine an otherwise rational policy. To this end, it is important to ensure that the low-risk prediction model and low-intensity supervision do not have hidden harmful effects.

The APPD experiment also raises a broader theoretical question – what is the nature of 'non-serious' offending, and how does it differ from higher-level offending?

Little research has been done on the characteristics of low-level offenders. While focusing on the more serious offenders is logical from a public safety perspective, the premise of the Philadelphia model is that the majority of offenders are likely to be low-risk. Thus, the population is worth considering for its size alone.

The objectives of this paper are to investigate the sensitivity of Philadelphia's prediction model to serious offending at various risk levels, and to examine whether low-intensity supervision could have the unintended consequence of increasing offending severity in a sample that should indicate little history of serious criminal behavior. In doing so, we learn more about the nature and degree of 'serious' crime among the majority of lower-level offenders. Our specific research questions are:

1. How successfully does the model categorize offenders as low or non-low risk?
2. Does the sensitivity of the model change under different definitions of offending severity?
3. How does the selected threshold for determining low/non-low risk (reliability score > 0.5) compare in its predictive ability to alternative cut-points?
4. Does low-intensity probation supervision affect offenders' propensity for serious offending?

## Methodology

Outcome data

A wide range of data collected as part of the experiment and model development are available for the present analysis. In addition to data on charges for offenses pre- and post-random assignment for the low-risk offenders who participated in the experiment,

we also have crime outcomes for the non-randomized low-risk backfill cases and for a group of offenders predicted to be non-low risk by the model. Data for the latter group of probationers were collected for comparison with low-risk cases in a regression discontinuity analysis, the outcomes of which were contrasted with the experimental results (Berk et al., forthcoming). The full dataset contains information on 93,540 charges for 3,207 offenders (2,207 of whom were predicted to be low risk) covering a period of 42 years (1967-2009).

One significant limitation of the charges database is that it only includes charges as an adult for offenses in Philadelphia, as we only had access to local adult criminal justice system databases. While almost all of the participants reside in Philadelphia, the city's proximity and ease of access to surrounding counties and state lines mean that the local data almost certainly underestimate the number of charges recorded for these offenders. In addition, we do not have juvenile data available to give a full picture of offenders' lifetime criminal involvement. While most of the offenders in our sample are older (see Table 3.1), we would expect to see the majority of their criminal offending taking place during their teenage years, so the lack of data on charges filed under the age of eighteen is a substantial omission. However, data are available where the offender was charged as an adult, even if the offense was committed while s/he was under eighteen.

Outcome measures

As we noted above, the definition of offending severity may extend beyond the substantive nature of the offense itself. This is particularly true in a sample of offenders

already predicted to be at low risk of serious reoffending. A successful model of low-risk prediction should minimize the possibility that these offenders pose a serious threat to society by any measure. This section considers the various ways in which offending severity may be conceptualized, and how we operationalize some of these ideas as outcome measures for the present study.

The assessment of offending severity has been a long-standing concern of criminological research. Blumstein et al. (1986), for example, describe offending severity as a "key dimension of individual criminal careers" (p. 76) and note the crucial policy interest in focusing on understanding and identifying serious offenders. Despite this interest, there remains little consensus on how best to measure severity. The conventional approach (Ramchand et al., 2009, p. 130) has been to weight different crime types based on the perceptual method established by Sellin and Wolfgang (1978). Sellin and Wolfgang asked panels of university students, juvenile court judges, and police officers to rate the severity of a range of crimes compared to a trivial baseline offense of the theft of $1. Based on these ratings, they assigned a weight to each offense relative to the baseline. This type of crime severity rating appears to hold across different samples and contexts (although some race-based differences have been noted), and remains popular despite criticisms that no context is provided to panel participants, leaving them free to speculate about unreported details of the offenses (Ramchand et al., 2009).

The U.S. Federal Bureau of Investigation (FBI) also provides an offense severity classification that is used by police departments when they report crime data through the Uniform Crime Reporting (UCR) program. The classification of offenses into Part I and Part II offenses was introduced in 1929 and now contains eight offenses in the Part I

category and the remainder in Part II (Federal Bureau of Investigation, 2004). The classification is based on offense seriousness, frequency of occurrence, nationwide pervasiveness, and likelihood of being reported. The eight Part I offenses are criminal homicide, forcible rape, robbery, aggravated assault, burglary, larceny-theft, motor vehicle theft, and arson. Thus, they are broader than the offenses considered as 'serious' in the development of Philadelphia's low-risk prediction model (although the Philadelphia model classifies a broader range of sexual offenses as 'serious').

An alternative to rating severity according to the substantive offense is to assess the economic cost of crime. Interest in cost-benefit analysis as a part of criminal justice program evaluation is beginning to grow (e.g., Marsh, Chalfin, & Roman, 2008), and estimates of the cost of each type of crime to society is a key part of the methodology entailed by this approach. The severity of crime is ranked by the extent of its cost to society in terms of victimization costs (e.g., stolen property, loss of earnings, medical expenses, trauma, and suffering: Cohen, 1988), and criminal justice system costs (police investigation, court processing, cost of commensurate sentence: Marsh & Fox, 2008). Cohen (2000) and colleagues (2004) also propose that the cost of crime can be quantified in terms of the public's theoretical willingness to pay for crime prevention programs that would reduce the prevalence of a particular offense type by 10 per cent.

More recently, some sophisticated statistical approaches to assessing offending severity have emerged. One recent example is Ramchand et al.'s (2009) developmental model of crime severity. They hypothesize that offenders will progress to more serious crime after engaging in low-level offending. This approach is appealing because, unlike previous classification models, it accounts for offender preference and is culture-specific.

It assumes that if offenders consider a particular crime type to be serious, they will only be drawn to it as their offending career escalates. Thus, the sequencing of crime types over the life course provides insight into which offenses offenders perceive as more or less serious. We do not examine this approach in our paper because we lack full lifetime offending data for our sample, so the utility of the developmental focus is limited. The quality of the data do not justify the complexity of the methods.

Our operationalization of substantive severity is straightforward. We simply classify each charge as a serious or non-serious offense according to the definition of severity used in the prediction model (see above). We also compare this definition of severity with the standard UCR distinction between serious and less serious offending, which includes more offenses than the Philadelphia model's distinction.

We propose a simpler method for analyzing economic severity. The more complex cost models may not be well suited to our data. Because our sample has already been deemed low-risk, in part because of their non-serious offending backgrounds, serious offenses will be rare events. Thus, it makes sense to simply dichotomize offending history and outcomes into serious/non-serious rather than attempting to create more detailed categories. However, an economic severity rating is more difficult to dichotomize, beyond the rudimentary approach of assigning an arbitrary dollar value as a threshold between serious and non-serious offending. The studies that do provide U.S. dollar estimates for crime types (e.g., Cohen, 1988; Miller, Cohen, & Wiersema, 1996) provide various formulas for assigning costs, and largely focus on the most serious crimes, so it is difficult to obtain estimates for the lower-level offenses that are more prevalent in our sample. For example, cost estimates set out in Cohen (2000) contain no

specific information for drug offenses, which constitute 16 per cent of all the charges in our sample. Furthermore, the age of these studies and changes in prices and monetary value may limit the usefulness of their estimates (although procedures are available for converting them into current values).

One proxy for dollar value that would also allow a basic distinction to be drawn between more or less 'expensive' crimes is victim status (crimes with victims, such as assault, versus 'victimless' crimes like drug and weapon possession). This approach assumes that crimes with victims cost society more because the victims themselves suffer both tangible (e.g., loss of earnings) and non-tangible (e.g. fear) costs, in addition to the criminal justice system costs of processing the offender, while non-victim crimes (crudely) only involve offender-related costs. Of course, the reality is less clear-cut, but victim status remains a useful proxy for cost, and does not fully overlap with substantive severity.[4]

We define 'victim' crimes as those offenses that were most likely to have involved injury or death, psychological distress, loss of earnings, or other costs such as loss of or damage to property of a personal (not corporate) victim. The process of applying the definition was somewhat crude, because only limited information about the offense was available. Our criminal history database contained a free-text description of each offense and a statute section and subsection reference.[5] The offense type was initially determined by reference to the relevant statute, as we believed that variable would be more accurate than the free-text description. The description was used for confirmation and additional information about the offense. However, without full crime reports for each offense, it was not possible to know for certain whether a victim was

involved. Crimes that obviously met our definition included homicide, rape, assaults, and residential burglaries. Arson was also included due to the high probability of costly damage, although it was not always possible to determine whether the offense was committed against personal or commercial property. Some other offenses in Pennsylvania (e.g., criminal mischief) have specific subsections relating to different types of victims and these were classified as victim crimes where it was clear that personal victims were involved. Retail theft and some other acquisitive offenses like theft of services, which were most likely to involve corporate victims, were excluded. We also excluded 'non-permanent theft' (unauthorized use offenses) against any victim type.

Analytic strategy

We use straightforward tests to compare the prevalence (proportion of offenders involved in serious offending) and frequency of serious offending across groups. Frequencies are compared using a two-sample *t*-test for the difference between means. To assess prevalence, we examine the relative risk (risk ratio) of serious offending between groups. The risk ratio is not often used in criminological research, but it is common in epidemiological research for analyzing dichotomous outcomes in cohort studies (in which known exposure/non-exposure to a risk factor is cross-tabulated with disease/non-disease status). It is simply a ratio of the probability (risk) of disease given exposure status, calculated by dividing the proportion of subjects with the disease at one level of exposure by the proportion at the other level. As such, it is somewhat similar to an odds ratio, but has a considerably more intuitive interpretation.[6] Like the odds ratio,

the risk ratio is bounded by zero at the lower end and has no upper bound. A risk ratio of 1 indicates no difference in risk between the groups. In the present study, a risk ratio of 2 would indicate that one group ('exposure': risk level or treatment status) is twice as likely to have a serious offense ('disease') than the other.

We use another epidemiological tool, sensitivity/specificity analysis, to assess the effect of changing the model's cut-off point for classifying risk. We examine how many offenders were correctly classified as low-risk (having no serious offense two years post-risk assessment date) at each cut point. Sensitivity is defined as the proportion of 'true positives' correctly identified by the model, or the proportion of offenders without a future serious offense who had been predicted low risk. Specificity is defined as the proportion of 'true negatives' identified: the proportion of serious recidivists who were classified as non-low risk. Sensitivity or specificity of 100 per cent indicate that the classification tool is able to identify all the true positives or all the true negatives, respectively. In practice, most classification models require a trade-off between one or the other: no model will perfectly classify every case, so users must decide whether it is more important to identify mostly true positives, or mostly true negatives. We also present the positive and negative predictive values of the model. The positive predictive value is the proportion of offenders predicted to be low risk who are actually low risk, and the negative predictive value is the proportion of offenders predicted non-low risk who go on to commit a serious offense. Formulas for calculating each of these measures are presented in Appendix J. We also define false positives as the proportion of predicted low-risk offenders committing serious offenses, and false negatives as the proportion of predicted non-low risk offenders not committing serious offenses.

We examine the potential interaction of low-intensity supervision and pre-random assignment (RA) serious offending on post-RA serious offending for participants enrolled in the Low Risk Experiment using Mantel-Haenszel methods for calculating an adjusted risk ratio across different levels of a covariate. This is also a commonly-used approach in epidemiological research. First, we calculate the unadjusted risk ratio for the prevalence of post-RA offending by assigned treatment. We then stratify by presence or absence of pre-RA serious offending, calculating two stratum-specific risk ratios. The Mantel-Haenszel method assigns a weight to each stratum and produces an adjusted overall risk ratio based on the weighted stratum-specific values. The accompanying Mantel-Haenszel chi-square test of homogeneity is used to consider whether an interaction effect may be present. If $\chi^2$ is statistically significant, we reject the null hypothesis of homogeneity of the stratum-specific risk ratios. That is, we consider them sufficiently different to constitute evidence that the stratifying variable (serious offending history) interacts with the independent variable (assigned treatment) to affect post-RA serious offending outcomes.[7] All analyses are conducted using the epidemiological methods suite in STATA 10.

Sample characteristics

We assess each of our four research questions using one of two separate samples ('full sample' and 'experimental sample') drawn from the complete set of 3,207 probationers described above. That dataset comprised 1,559 predicted low-risk experimental participants (800 LIS treatment and 759 SAU control), 648 predicted low-

risk backfill cases who were not randomly assigned but received LIS, and 1,000 predicted non-low risk offenders selected as part of a separate study.

Our 'full sample' is a subset of all the groups that make up the 3,207-offender dataset, divided into low and non-low risk cases. Most of the backfill cases (N = 588) did not have a recorded reliability score from the prediction model. Although we could assume that they were low risk because they were in the backfill group, they could not be used to examine the questions relating to the cut-off point for a low risk prediction, and we decided to exclude them from the analysis completely. In addition, one treatment group case had no recorded reliability score and was excluded from the full sample, but is included in the experimental sample. Finally, fifteen backfill cases had a reliability score between 0.49 and 0.5. We strictly followed the requirement for classifying cases with a reliability score of over 0.5 as low risk in this analysis, so those fifteen cases are classified as non-low risk in the full sample. Thus, the full sample comprises 2,618 offenders in total: 1,603 predicted by the model to be low risk, and 1,015 predicted to be non-low risk. Our 'experimental sample' consists of the 1,559 experimental participants, analyzed on an intention-to-treat basis (i.e., they remain in their assigned groups regardless of whether they actually received the assigned treatment).

Tables 3.1 and 3.2 show basic demographic and offending history characteristics for the two samples. Race is presented only as the proportion of white offenders because of problems in the recording of race in the original dataset,[8] which meant that it was only possible to reliably say whether the offender was white or nonwhite. The mean age is calculated according to the offender's age (based on recorded date of birth) on October 1, 2007. This was the date on which experimental participants were randomly assigned, and

age is based on this date regardless of whether or not the offender participated in the experiment. The proportions and means of charged offenses are based on the full range of offending data from the offender's first recorded charge until September 30, 2009. For the full sample, data were available on a total of 81,643 charged offenses committed between 1967 and 2009. Our analyses are based only on those offenses committed after the date of the risk assessment (July 27, 2007; N = 6,808), because prior offending history variables were used in the predictive model. For the experimental sample we had data on 34,777 charged offenses over the same timeframe. Of this number, around 6 per cent (N = 2,147) were committed post-random assignment.

As we might expect, the low risk and non-low risk groups in the full sample look very different (Table 3.1). The non-low risk group is much more likely to be male (87.4% vs. 66.9%, $p < .001$), nonwhite (75.6% vs. 60.4%, $p < .001$), and younger (31 years old vs. 40.7, $p < .001$). The non-low risk group has also been charged with more than twice as many offenses overall as the low risk group (45 vs. 22.4, $p < .001$).

The characteristics of the treatment and control groups in the experimental sample are very similar, indicating successful random assignment. 66.5 per cent of the treatment group and 67.6 per cent of the control group are male. Slightly more treatment group members than control group members are white (41.8% vs. 38.0%, $p \leq .125$). Participants in both groups were, on average, just under 41 years old on the date of random assignment, and members of both groups have been charged with an offense on average 22.3 times as adults up to two years post-random assignment.

## Results

How successfully does the model categorize offenders as low or non-low risk?

The first three rows of Table 3.3 show the prevalence and frequency of post-risk assessment serious offending (as defined in the Philadelphia APPD prediction model) for probationers predicted to be low or non-low risk. The differences between the two groups are substantial and highly statistically significant on both measures. Among offenders predicted to be low risk, 3.4 per cent were charged with a serious offense such as murder, aggravated assault, or a sexual offense over the course of their available offending histories. In the non-low risk group, 10.2 per cent were so charged. Thus, offenders receiving a low risk prediction were 67 per cent less likely to have ever been charged with a serious offense than their non-low risk counterparts (risk ratio RR = .33, $p < .001$). Similarly, the mean number of serious offenses committed by the non-low risk group in the available records was almost three times greater than the mean for the low risk group (.38 vs. .13, $p < .001$). It appears, therefore, that the predictive model was successful in classifying offenders into low and non-low risk groups at the 0.5 reliability score threshold, when applied to a new set of current probationers.

Table 3.4 shows more detail about the types of serious offenses committed by the two groups, and provides compelling evidence that the low risk group poses a substantially smaller threat to public safety. Offenders predicted to be non-low risk were nearly eight times more likely than those predicted low risk to be charged with homicide or attempted homicide after the risk assessment date (1.5% vs. .2%, $p < .001$). We see similar, highly significant differences for sexual offenses, aggravated assaults, and

robberies. The only crime on which the two groups did not differ statistically is forcible rape (a subset of all sexual offenses), but this is most likely due to the very small number of events. Fewer than 1 per cent of each group were charged with rape, but the probability is still more than three times greater in the non-low risk group (.39% vs. .12%, $p \leq .160$).

Does the model's sensitivity change under different definitions of serious offending?

We conducted similar analyses with the low and non-low risk groups using two alternative definitions of offending severity, based on UCR Part I offenses and offenses deemed to be more likely to involve a victim or serious damage (i.e., involving a greater economic cost). Of course, the Philadelphia model was not designed to predict such offenses, so the purpose of this question is not to validate the model, but rather to examine the types of offenses committed by probationers deemed to be at low risk of serious harm, and whether they could be considered serious under alternative definitions. The results of these analyses are presented in the remaining parts of Table 3.3.

As we would expect from the preceding analysis, the low and non-low risk groups also differ substantially and significantly on these alternative measures of severity. However, our alternative definitions slightly inflate the proportion of both groups that would be identified as 'serious' offenders. 9.1 per cent of low risk offenders had been charged post-risk assessment with a UCR Part I offense, which include murder and rape, and also burglary and motor vehicle theft. The proportion of non-low risk offenders charged with a UCR Part I offense is 15.7 per cent. Low risk offenders are 42 per cent

less likely to have been charged with a Part I offense than non-low risk offenders (RR = .58, $p < .001$). Again, the frequency of serious offending is much greater for the non-low risk group, with a mean of .62 Part I offenses compared to .34 in the low risk group ($p < .001$). The proportion of offenses involving a victim or serious damage is comparable to the UCR Part I results for both groups (low: 8.2%, non-low: 17.5%; RR = .47, $p < .001$). Again, victim/damage charges appeared much more frequently in the histories of non-low risk offenders (.91 vs. .41 on average, $p < .001$). Thus, the predictive model is still able to distinguish low and non-low risk participants based on either UCR Part I or victim/damage offending, but considerably more probationers in both groups would now be said to have been involved in 'serious' offending.

How does the model's threshold for determining low risk compare to alternative cut-offs?

Tables 3.5 to 3.7 show the sensitivity, specificity, and positive and negative predictive values for the model when different thresholds of the reliability score are used for classifying offenders as low or non-low risk. In the present study, offenders with an average reliability score of above 0.5 were classified as low risk. We compare the model's predictive ability at this threshold with alternative cut-off points ranging from 0.05 (at which all offenders were classified as low risk) to 0.95 (at which all offenders were non-low risk).

Identifying the most suitable threshold necessarily involves balancing the model's ability to predict low risk cases against its ability to identify who will commit a serious offense. We suggest that the latter concern is more important to the viability of a policy

of low-intensity supervision based on risk of serious offending, because there may be public and political anxiety about reducing criminal justice intervention to adjudicated offenders. Thus, for low-intensity probation to maintain credibility, it is arguably more sensible to demonstrate that few serious offenders slipped through the net than to show how many non-serious offenders had their supervision requirements reduced. From the model standpoint, we must ensure a high positive predictive value, which indicates the proportion of predicted low-risk offenders who were actually low risk, and high specificity (proportion of serious offenders who received a non-low risk prediction). Large values for these two measures indicate a low rate of false positives (predicted low-risk offenders who commit serious crimes). Conversely, we expect to see lower sensitivity and lower negative predictive values, because the sample contains many false negatives. Most non-low risk probationers do not go on to be serious offenders in the two-year follow-up. Serious offenses are rare events in our sample, and non-low risk offenders are not necessarily *high* risk.[9] Low values on these two measures are more acceptable because there is no harm when an offense is not committed, regardless of the risk prediction. However, we must also keep the purpose of the model in mind: the diversion of a majority of offenders in APPD's caseload to low-intensity supervision. If sensitivity is too low (too few non-serious offenders received low-risk predictions), that goal will not be fulfilled.

Table 3.5 shows the results of these tests using the model definition of severity. The positive predictive values are high at all thresholds, indicating a low rate of false positives in general. This is promising, but will be driven by the very low sample prevalence of serious offending post-risk assessment. The present cut-off point of 0.5

appears to be a good classification threshold. Here, the model's sensitivity and specificity are most balanced compared to other cut-off points (Sn = 63.0%; Sp = 65.8%). This means that the probability that a non-serious offender received a low risk prediction and the probability that a serious offender received a non-low risk prediction are roughly the same. Of the offenders receiving a low-risk prediction, 96.6 per cent were in fact low-risk. The 'worst case scenario' false positive rate (low-risk offenders who committed serious offenses) is very low, at 3.4 per cent.

Table 3.5 suggests that the cut-off point should not be set below 0.5. Although the positive predictive value remains high at thresholds of 0.45 and below, there is a considerable loss of specificity at the expense of the less important sensitivity (Sn = 71.3%; Sp = 55.1% at 0.45 threshold). The probability of finding a false positive also begins to increase. On the other hand, there may be a case for increasing the cut-off point to 0.55, but no more. At 0.55, the positive predictive value increases to 97.3 per cent, specificity increases to 78.5 per cent, and the likelihood of a false positive drops to 2.7 per cent. However, the sensitivity drops to just over 50 per cent, which starts to raise questions about the model's ability to meet its purpose. Thus, a threshold between 0.5 and 0.55 appears to provide the best trade-off between all the factors discussed above.

Tables 3.6 and 3.7 show the same analyses repeated for UCR Part I and victim/damage offenses. Note that the model is not designed to predict these offense types (as the slightly lower positive predictive values in these two tables suggest), so the results from Table 3.5 should be taken as the definitive examination of the threshold. However, we use these additional outcomes to examine whether the cut-off point allows too many offenders who might be considered 'serious' by alternative standards to be

classified as low risk. The 0.5 threshold again performed reasonably well. This cut-point gave the best balance of sensitivity and specificity for both measures, but specificity was lower than sensitivity (UCR: Sn = 63.0%, Sp = 52.1%; victim/damage: Sn = 63.7%, Sp = 57.4%). Again, increasing the cut point to 0.55 improved specificity for both measures, at the expense of a reasonable degree of sensitivity (UCR: Sn = 50.6%, Sp = 63.9%; victim/damage: Sn = 51.5%, Sp = 70.3%). At the 0.55 threshold the positive predictive values increase slightly and the proportion of false positives is reduced. We see higher rates of false positives in these analyses compared to Table 3.5 because there is a higher prevalence of offending in these categories.

Does low-intensity probation supervision alter the propensity for serious offending?

This analysis focuses only on the experimental sample: 1,559 predicted low risk offenders who participated in the Low Risk Experiment and were randomly assigned to low-intensity supervision (LIS) or supervision as usual (SAU). Regardless of their low-risk status, we would expect that those offenders who had been involved in serious offending prior to random assignment (RA) would be more likely to continue to do so.[10] However, because the ultimate practical purpose of the predictive model is to identify low-risk offenders *so that they can be diverted to LIS*, it is very important to ensure that low-risk supervision itself does not increase the likelihood that offenders will engage in serious recidivism during or after supervision, over and above the extent to which we would expect given their past behavior. Table 3.8 shows the proportion of the sample that were charged with a serious offense post-RA. Because of the small number of

serious offenses in the low risk sample as a whole, this analysis focuses only on prevalence, not frequency. However, we do consider the two alternative definitions of severity along with model-defined severity. Table 3.9 presents a stratified analysis according to whether or not the offender had committed a serious offense pre-RA.

Table 3.8 shows that control group members were slightly more likely to have committed a serious offense post-RA, regardless of the definition, although none of the results reaches statistical significance. The treatment group was 39 per cent less likely than the control group to have committed a serious offense as defined by the model (RR = .61, $p \leq .079$); 17 per cent less likely to have committed a UCR Part I offense (RR = .83, $p \leq .259$); and 21 per cent less likely to have committed an offense involving a victim or damage (RR = .79, $p \leq .179$). It appears, then, that the low-intensity supervision model did not lead to any increase in the propensity for serious offending. However, we cannot say this with certainty, nor suggest that LIS helps to reduce serious offending, because the analysis does not account for past behavior.

We examined whether prior offending interacts with treatment assignment to affect future offending by stratifying our analysis according to the prevalence of pre-RA serious offending. Table 3.9 sets out the stratum-specific and Mantel-Haenszel adjusted risk ratios for each definition of severity. The stratum-specific risk ratios tell us if there is any difference in the effect of LIS on post-RA offending depending on whether or not the offender had previously committed a serious offenses. We then compare the overall Mantel-Haenszel adjusted risk ratios to the unadjusted risk ratios from Table 3.8 to assess whether the evidence for an interaction effect is sufficient.

All the risk ratios in Table 3.9 indicate that treatment group participants were less likely to commit a serious offense post-RA than the control group, although only one is statistically significant. There is also a notable difference between the risk ratios of probationers who had and had not committed a serious offense pre-RA. For model-defined severity, treatment group participants who had not committed a prior offense were 51 per cent less likely to have been charged post-RA than similarly-situated control group members (2.0% vs. 4.0%, stratum-specific RR = .49, $p \leq .042$). However, there was little difference between treatment and control group participants who had committed a prior serious offense (3.8% vs. 4.2%, stratum-specific RR = .90, $p \leq .816$). We see the same pattern with UCR and victim/damage crimes. For UCR offenses, treatment group participants without a prior offense were 44 per cent less likely to be charged than control group participants without a prior offense, but there was no difference between treatment and control group participants with a prior offense (no prior: 3.7% vs. 6.5%, RR = .56, $p \leq .136$; prior: 10.2% vs. 11.1%, RR = .92, $p \leq .660$). For victim/damage crimes, the risk ratio was .58 for non-serious prior offenders compared to .87 for serious prior offenders (no prior: 3.9% vs. 6.8%, $p \leq ..134$; prior: 8.4% vs. 9.7%, $p \leq .494$).

Despite the magnitude of some of these results, the Mantel-Haenszel adjusted risk ratios show no evidence of an interaction effect between prior serious offending and treatment. For all outcome measures they are identical to the unadjusted risk ratios, and we were unable to reject the null hypothesis of homogeneity in any case (model defined: $p \leq .294$; UCR: $p \leq .248$; victim/damage: $p \leq .326$), meaning that the stratum-specific risk ratios do not differ enough statistically to suggest a strong interaction. Nonetheless,

although we should be cautious about reading too much into the stratum-specific risk ratios due to the small number of events, it is clear that low-intensity supervision was more effective than treatment as usual for offenders without a prior history of serious offending, than it was for offenders who had committed a serious offense.

## Discussion

Our first three research questions examined the sensitivity of the prediction model used by Philadelphia APPD in classifying offenders by risk across several different definitions of severity. The model appears to successfully categorize probationers into low and non-low risk. Overall, the probability that an offender in our sample had been charged with a serious offense (according to the model definition: murder, attempted murder, aggravated assault, robbery and sexual offenses) was substantially lower if they received a low risk prediction than if they did not. The average 'reliability score' assigned by the model to each offender across all of his or her probation cases also appeared to be linearly related to the offender's likelihood of serious offending: in general, the higher the score (higher scores represent the lowest risk levels), the less likely an offender was to have been charged with a serious offense.

The threshold used to distinguish predicted low risk offenders from predicted non-low risk offenders in the model, an average reliability score of 0.5, largely appears to be an appropriate cut-off point in the present sample. However, the model performs slightly better in terms of avoiding the most serious errors – offenders who were predicted to be low risk but committed serious offenses – if the threshold is raised to

0.55. Raising the threshold results in a slight loss in the ability to predict who will *not* commit a serious offense, but the trade-off is small and favors increased public safety. This alternative threshold is close enough to the original 0.5 cut-off that the results of the present study are unlikely to be greatly affected, and as a matter of policy any change would depend on the extent to which the probation agency was willing to trade considerable resource savings for a small potential decrease in false positives. Twelve per cent of the full sample (323 offenders) had risk scores between 0.5 and 0.55, and would not have been eligible for low-intensity supervision at the higher classification threshold.

One limitation of these diagnostic tests is the very low prevalence of serious recidivism post-risk assessment in the sample as a whole, which led to a large number of false negatives. This in turn reduces the ability of the model to predict true positives (true low-risk cases), which is the ultimate goal of the risk assessment process. This is an issue that is unlikely to be easily overcome, because serious offenses like homicide and rape will always be relatively rare events. However, the prediction model can be applied to any new probationer entering the Philadelphia APPD, so the potential exists for more data to be collected on low and non-low risk clients, their serious offending, and the performance of low risk offenders under low-intensity supervision. These data could be added to an analysis like those presented here, in order to conduct continuous validation and refinement of the model.

Our analysis shows that the model is also somewhat successful in ensuring that offenders who might not have committed one of the most serious offenses, but could be considered serious offenders by other standards, are not channeled into receiving less

supervision than they might need. When we repeated our analyses of the model using UCR Part I offenses (those offenses considered by the FBI to be of greater concern to the authorities, based on severity among other factors), and offenses with victims that were likely to involve a greater economic cost, we saw similar patterns of offending as for model-defined severity. The low risk group were still at lower risk of committing these offenses than the non-low risk group, and our findings about the threshold held relatively constant. However, using these alternative definitions may defeat the key object of Philadelphia's prediction model: to better distribute agency resources according to the risk of the type of serious offending that poses the greatest threat to public safety and fear of crime. Only 31 per cent of low risk offenders had ever been charged with such an offense (3 per cent post-risk assessment), but nearly 70 per cent on average had been charged with a UCR Part I or victim/damage offense (nearly 9 per cent post-risk assessment). While these are undoubtedly offenses that contribute substantially to the crime problem, their high prevalence in a sample of offenders known not to be causing the worst kinds of harm to society suggests that it is less crucial to focus on these crime types than homicide, robbery, serious assaults, and sexual crimes.

Our choice of alternative definitions of severity was the main limitation of this part of the exploration because it was not possible with the available data to create a more detailed ranking of severity based on different factors. Our two proxy measures were necessarily too broad because of the difficulties (discussed above) in analyzing limited data on rare events. They tended to overstate serious offending by including offenses that might not be considered serious at all when deciding which offenders require more intensive criminal justice system intervention. UCR Part I offenses, as discussed above,

are partly selected on the basis of substantive severity, but also on other factors such as frequency of commission and likelihood of detection. It is clear, given these additional qualities of the UCR Part I offenses, that any offender, regardless of risk, would be much more likely to commit some of these offense types. Since there is some overlap of the most severe crimes in the UCR list with the model-defined serious offenses, it is likely that the majority of additional UCR Part I offenses committed by the low risk group were the less serious, more nationally prevalent crimes like motor vehicle theft. Similarly, our victim proxy for cost likely captured some less serious crimes (substantively or economically) simply because we selected them based on victim status only. The 1993 dollar estimates provided in Cohen (2000) show that the tangible and quality of life costs of victim crimes vary widely, not to mention the costs of some non-victim crimes. For example, the cost per victimization for homicide was thought to be nearly $3 million, compared to just $2,000 for an assault without serious injury, but our measure included both. A more refined analysis of economic severity in comparison to the model's definition would require more data, updated cost estimates for a wide range of victim and non-victim crimes, and more detail about each charge in order to make more accurate judgments about injury and damage, beyond the simple victim/non-victim distinction. Additional data on lifetime offending would also allow for some more sophisticated developmental analyses of severity and offending escalation like the model proposed in Ramchand et al. (2009).

In all, it would appear that the Philadelphia model has achieved its purpose with this sample of offenders. Returning to the earlier discussion of the qualities that make the model a valuable contribution to the risk prediction literature, we can confirm that the

model provides predictive validity; is directly relevant to correctional priorities in its focus on the rare but highly harmful crime events of concern to Philadelphia APPD, and its role in facilitating the most effective allocation of operational resources; and adheres to the principle of the least restrictive alternative by using a measure of offending severity that does not overestimate the number of offenders requiring more intensive supervision.

The results of our investigation into a possible interaction between supervision intensity and serious offending showed no evidence that reducing supervision intensity for predicted low-risk offenders might increase the risk of serious offending. However, although the analysis did not indicate a statistically significant interaction effect, we found that probationers assigned to low-intensity supervision only *reduced* their offending compared to controls when they had no prior serious offending history. In such cases, the probability of a new offense was halved. Low-risk probationers with a history of serious offending performed no better than their counterparts on traditional supervision. It is possible that this interaction did not reach significance because of the very small number of post-RA serious charges during the two-year follow-up period of the Low Risk Experiment.

This finding provides further evidence that low-intensity supervision can be a safe, effective probation strategy. The idea that some offenders receiving a low-risk prediction might have a history of serious offending could be objectionable to policymakers and the public. However, this analysis shows that reduced supervision is, at *worst*, no different from the status quo. At best, it may reduce reoffending for probationers with the lowest-level criminal careers. This is also an important discovery about the nature of low risk offenders in general. It appears that low-level offenders

make up the majority of APPD's caseload. Their overall propensity to reoffend at the time of risk assessment is extremely low. Regardless of their past history, offenders predicted to be at low-risk of committing a serious crime respond just as well to a less restrictive intervention as they do to a more intensive one. The less severe their history, the more likely they are to improve their outcomes, even with minimal involvement from the criminal justice system. Building such knowledge about the cases that make up the majority of a probation agency's caseload could be vital for the planning and allocation of resources, and the ability to tailor supervision to clients' needs and requirements.

Our analysis is somewhat limited because of the low prevalence of serious offending in the sample. With more data, it might be possible to shed further light on the characteristics of low-level offending to improve probation agency decision-making. One useful line of inquiry would be to look at the escalation in serious offending up to the point of random assignment, and whether the timing of the serious priors has any bearing on their interaction with supervision intensity. Our basic approach of examining the presence or absence of serious prior offending may mask differences between, for example, an older offender who was charged with a serious offense twenty-five years ago and has only committed a few minor offenses since, and a younger offender whose earliest offenses were trivial but whose career had started to escalate shortly before the current probation term. In addition, our measure of serious offending is based on charges, and as with any crime outcome measure there are numerous factors in the decisions to charge an offender, and to drop a charge or fail to convict, that are unrelated to whether or not the offender actually committed the crime. This is a crucial consideration in any discussion of the relationship between past and future behavior.

Although it is unlikely that we would be able to learn the full details of the offense and subsequent criminal justice decision-making processes, it should be possible to take into account the ultimate disposition of the charge in future analyses.

## Conclusion

This paper examined patterns of offending severity in a sample of probationers in order to assess how risk can best be predicted and managed for the effective operation of a probation agency. As the use of probation continues to grow, and especially as offenders who pose a significant threat of harm to society are placed under community supervision, creative resource allocation is required to manage these offenders effectively.

Risk prediction techniques have been used in probation and other criminal justice agencies for one-hundred years, with varying success. The Philadelphia Adult Probation and Parole Department and the University of Pennsylvania developed a new statistical risk prediction model that differs from other instruments in two main ways: it attempts to operationalize the clinical, informal decision rules already used by probation officers in the department, and it focuses only on predicting a handful of crimes considered to be the most detrimental to public safety and confidence in the criminal justice system. The ultimate goal in creating the model was to provide a tool for classifying the entire APPD caseload along risk-based lines, and channeling a majority of offenders who posed little risk of serious harm into a large caseload receiving reduced intensity supervision. This

strategy allows probation officers to focus their time and resources on the highest-risk clients.

We explored the ability of the model to correctly classify offenders as low or non-low risk, using the sample of offenders who participated in the trial of low intensity supervision and an additional group of APPD clients who also received low or non-low risk predictions according to the model. We also examined the possibility that assignment to low-intensity supervision could interact with prior serious offending to increase serious recidivism compared to regular probation. Our analyses revealed that the model is largely successful, perhaps needing just a slight adjustment the threshold for defining low-risk. In the context of the APPD's goal of classifying the majority of its caseload as low risk, we also found that the crimes defined in the model as 'serious' provided a better indication of who the higher-risk offenders were than did UCR Part I offenses or a simple victim/non-victim crime status indicator. The majority of offenders predicted by the model to be low risk had committed a 'serious' offense by those definitions at some point in their careers, although their risk of an offense in these categories remained lower than that of predicted non-low risk offenders. Thus, we conclude that a statistical model of the type developed in Philadelphia would appear to be a useful offender and resource management tool.

Proceeding from the assumption that a majority of a probation agency's caseload could be classified as low-risk for the most serious recidivism, we discovered that such offenders respond just as well, if not better, to reduced supervision as they do to traditional probation. Offenders with no prior history of serious offending appear to improve their outcomes regardless of probation's input. This is a strong justification for

the use of low-intensity supervision with the lowest-level offenders. These clients need no more than the minimum input of resources necessary to ensure that they are not 'false positives' and have the tools needed to rebuild their lives. The strength of the Philadelphia prediction model allows us to say with some confidence that we can identify a large proportion of offenders who fall into this category. This leaves the agency much better equipped to deal with the 'power few' highest-risk offenders.

Of course, no assessment or validation of a statistical prediction model can bring complete peace of mind in terms of guaranteeing the offender management approach that best assures public safety, just as the model itself cannot indicate exactly who will turn out to be low or high risk. We conclude that the Philadelphia model is successful in classifying offenders by risk, but 30 per cent of the predicted low risk offenders in our sample had committed at least one of the most serious offenses at some point in their adult offending careers. Some of that group do not react as well to low-intensity supervision as their counterparts with no history of serious offending, and we do not yet know why, or if their performance will worsen as more data are collected. We return to consideration of Grove and Meehl's (1996) comment: "The statistics furnish us with the probabilities so far as anything can" (p. 306). Any attempt to routinize criminal justice decision making will necessarily be concerned with averages, but it seems that Philadelphia's probation agency has developed a successful model that can help to allocate resources where they are needed most, and can easily be adapted for use elsewhere based on the information available to the specific agency.

# Notes

[1] In this study, murder, attempted murder, aggravated assault, robbery, and sexual offenses were deemed 'serious' offenses.

[2] Intake information includes the offender's personal and residential characteristics, and information about the instant offense and prior criminal history.

[3] Full details about the experimental design and how the sample was selected, assessed for eligibility, and randomly assigned may be found in Barnes et al. (forthcoming).

[4] An alternative approach could be to use incarceration status as a proxy for cost. The disadvantage is that we must use either actual incarceration data for our sample, or assume the types of offenses that might result in a sentence of imprisonment. Assumptions may be too subjective given the discretion involved in sentencing (although state sentencing guidelines could assist), and full incarceration data are not available for our sample. In particular, it is possible that some post-random assignment sentencing decisions are still pending given the relatively short period of time since the experiment ended. While assigning victim status to each offense type also involves assumptions, the level of subjectivity is likely considerably lower than it would be for incarceration.

[5] Most offenses in the dataset are derived from the Pennsylvania Consolidated Statutes, Section 18 (Crimes and Offenses).

[6] The risk ratio is simply $p_{E=1}/p_{E=0}$ (where $p$ = probability and $E$ = dichotomous exposure status), whereas the odds ratio is $(p/(1-p)_{E=1})/(p/(1-p)_{E=0})$. The odds ratio tends to overstate our 'natural' interpretation of relative outcomes: if the exposed group has a 50% risk of disease and the unexposed group has a 25% risk, the risk ratio is clearly 2 (the exposed group is twice as likely to get the disease than the unexposed group), but the odds ratio is 3 (the odds of disease in the exposed group are three times those of disease in the unexposed group), which seems greater. The risk ratio also remains stable regardless of the size of the risk; the magnitude of the odds ratio is closer to the risk ratio when the probability of disease in each group is small, and further away when it is large. Following the example above, if the risks were reduced to 20% and 10% respectively, the risk ratio would still be 2 but the odds ratio would fall to 2.25.

[7] In addition, an adjusted risk ratio that is substantially different from the unadjusted risk ratio (a 10-15% difference is a commonly-used rule of thumb) suggests that the stratifying variable is a confounder.

[8] The race indicator variable was populated with data from two different sources, with one source selected as the default. However, serious discrepancies arose because the categories of race in the two original sources were substantially different.

[9] On the other hand, it is also possible that non-low risk probationers are in fact more serious offenders, and are more likely to be incarcerated as a result.

[10] That past behavior is one of the strongest predictors of future behavior is one of the best documented findings in criminological research (e.g., Nagin & Paternoster, 1991; Nagin & Farrington, 1992).

## Tables

Table 3.1: Sample Characteristics (Full Sample)

| | | | Low Risk Group (N=1,603) | Non-Low Risk Group (N=1,015) |
|---|---|---|---|---|
| **Offender Characteristics** | | | | |
| % Male | | | 66.9 | 87.4*** |
| % White | | | 39.6 | 24.4*** |
| Mean age | | | 40.71 | 31.00*** |
| **All Charges** | | | | |
| Lifetime | | % | 99.9 | 99.9*** |
| | | Mean | 22.43 | 45.00*** |
| Post-Risk Assessment | | % | 22.7 | 37.5*** |
| | | Mean | 1.46 | 3.03*** |
| **Serious Charges** | | | | |
| **Model-Defined** | Lifetime | % | 30.8 | 78.4*** |
| | | Mean | 1.08 | 5.00*** |
| | Post-Risk Assessment | % | 3.4 | 10.2*** |
| | | Mean | .13 | .38*** |
| **UCR Part 1** | Lifetime | % | 68.5 | 89.7*** |
| | | Mean | 5.84 | 11.69*** |
| | Post-Risk Assessment | % | 9.1 | 15.7*** |
| | | Mean | .34 | .62*** |
| **Victim/ Damage** | Lifetime | % | 67.4 | 90.1*** |
| | | Mean | 5.87 | 14.34*** |
| | Post-Risk Assessment | % | 8.2 | 17.5*** |
| | | Mean | .41 | .91*** |

*** $p < .001$, 2-tailed $z$ (proportion) & 2-tailed $t$ (mean).

Table 3.2: Sample Characteristics (Experimental Sample)

| | | | Treatment Group (N=800) | Control Group (N=759) |
|---|---|---|---|---|
| **Offender Characteristics** | | | | |
| % Male | | | 66.5 | 67.6 |
| % White | | | 41.8 | 38.0 |
| Mean age | | | 40.78 | 40.58 |
| **All Charges** | | | | |
| Pre-RA | | % | 99.6 | 99.9 |
| | | Mean | 20.99 | 20.87 |
| Post-RA | | % | 21.5 | 21.5 |
| | | Mean | 1.31 | 1.44 |
| **Serious Charges** | | | | |
| **Model-Defined** | Pre-RA | % | 29.5 | 27.9 |
| | | Mean | 1.02 | .89 |
| | Post-RA | % | 2.5 | 4.1 |
| | | Mean | .13 | .13 |
| **UCR Part I** | Pre-RA | % | 65.9 | 67.7 |
| | | Mean | 5.19 | 5.73 |
| | Post-RA | % | 8.0 | 9.6 |
| | | Mean | .32 | .34 |
| **Victim/ Damage** | Pre-RA | % | 65.1 | 65.4 |
| | | Mean | 5.41 | 5.51 |
| | Post-RA | % | 6.9 | 8.7 |
| | | Mean | .34 | .42 |

No significant differences. 2-tailed *z* (proportion) & 2-tailed *t* (mean).

Table 3.3: Prevalence and Frequency of Post-Risk Assessment Serious Offending by Risk Level

| | | Low Risk (N=1,603) | Non-Low Risk (N=1,015) |
|---|---|---|---|
| **Model-Defined** | % Serious | 3.4 | 10.2 |
| | Risk Ratio | .33*** | |
| | Mean | .13 | .38*** |
| **UCR Part I** | % Serious | 9.1 | 15.7 |
| | Risk Ratio | .58*** | |
| | Mean | .34 | .62*** |
| **Victim/Damage** | % Serious | 8.2 | 17.5 |
| | Risk Ratio | .47*** | |
| | Mean | .41 | .91*** |

*** $p < .001$, $\chi^2$ (prevalence) & 2-tailed *t* (frequency).

Table 3.4: Types of Post-Risk Assessment Serious Charges by Risk Level

| % Charged | Low Risk Group (N=1,603) | Non-Low Risk Group (N=1,015) |
|---|---|---|
| Homicide | .2 | 1.5*** |
| Murder | .2 | 1.4*** |
| Sexual Offense | .3 | .9** |
| Forcible Rape | .1 | .4 |
| Aggravated Assault | 2.1 | 7.6*** |
| Robbery/Carjacking | 1.6 | 3.5*** |

Includes attempts, except "Murder," which includes only completed of the first to third degrees.
** $p \leq .01$, *** $p < .001$, 2-tailed $z$.

Table 3.5: Predictive Ability at Alternative Thresholds (Model-Defined Severity)

| Cut-point | Positive Predictive Value (%) | Negative Predictive Value (%) | Sensitivity (%) | Specificity (%) | False Positives (%) | False Negatives (%) |
|---|---|---|---|---|---|---|
| **0.05** | 94.0 | - | 100.0 | 0.0 | 6.0 | - |
| **0.1** | 94.0 | - | 100.0 | 0.0 | 6.0 | - |
| **0.15** | 94.0 | - | 100.0 | 0.0 | 6.0 | - |
| **0.2** | 94.1 | 60.0 | 99.9 | 1.9 | 5.9 | 40.0 |
| **0.25** | 94.5 | 27.6 | 98.3 | 10.1 | 5.5 | 72.4 |
| **0.3** | 94.9 | 18.8 | 94.2 | 20.9 | 5.1 | 81.3 |
| **0.35** | 95.1 | 13.2 | 86.9 | 31.0 | 4.9 | 86.8 |
| **0.4** | 95.9 | 13.2 | 79.9 | 47.5 | 4.1 | 86.8 |
| **0.45** | 96.1 | 11.0 | 71.3 | 55.1 | 3.9 | 89.0 |
| **0.5** | **96.6** | **10.2** | **63.0** | **65.8** | **3.4** | **89.8** |
| **0.55** | 97.3 | 9.3 | 50.7 | 78.5 | 2.7 | 90.7 |
| **0.6** | 97.6 | 8.2 | 39.1 | 84.8 | 2.4 | 91.8 |
| **0.65** | 97.4 | 7.4 | 29.4 | 88.0 | 2.6 | 92.6 |
| **0.7** | 97.8 | 7.0 | 20.3 | 93.0 | 2.2 | 93.0 |
| **0.75** | 98.1 | 6.6 | 12.7 | 96.2 | 1.9 | 93.4 |
| **0.8** | 98.9 | 6.4 | 7.4 | 98.7 | 1.1 | 93.6 |
| **0.85** | 98.6 | 6.2 | 2.9 | 99.4 | 1.4 | 93.8 |
| **0.9** | 100.0 | 6.1 | 0.9 | 100.0 | 0.0 | 93.9 |
| **0.95** | 100.0 | 6.0 | 0.0 | 100.0 | 0.0 | 94.0 |

Table 3.6: Predictive Ability at Alternative Thresholds (UCR Part I Offenses)

| Cut-point | Positive Predictive Value (%) | Negative Predictive Value (%) | Sensitivity (%) | Specificity (%) | False Positives (%) | False Negatives (%) |
|---|---|---|---|---|---|---|
| **0.05** | 88.3 | - | 100.0 | 0.0 | 11.7 | - |
| **0.1** | 88.3 | - | 100.0 | 0.0 | 11.7 | - |
| **0.15** | 88.3 | - | 100.0 | 0.0 | 11.7 | - |
| **0.2** | 88.4 | 60.0 | 99.9 | 1.0 | 11.6 | 40.0 |
| **0.25** | 88.8 | 29.3 | 98.2 | 5.6 | 11.3 | 70.7 |
| **0.3** | 89.1 | 22.2 | 94.1 | 12.8 | 10.9 | 77.8 |
| **0.35** | 89.5 | 18.5 | 86.9 | 22.6 | 10.5 | 81.5 |
| **0.4** | 90.1 | 18.1 | 79.9 | 33.8 | 9.9 | 81.9 |
| **0.45** | 90.4 | 16.4 | 71.3 | 42.6 | 9.6 | 83.6 |
| **0.5** | **90.9** | **15.7** | **63.0** | **52.1** | **9.1** | **84.3** |
| **0.55** | 91.4 | 14.6 | 50.6 | 63.9 | 8.6 | 85.4 |
| **0.6** | 92.0 | 13.9 | 39.3 | 74.1 | 8.0 | 86.1 |
| **0.65** | 91.5 | 12.9 | 29.4 | 79.3 | 8.5 | 87.1 |
| **0.7** | 92.7 | 12.7 | 20.4 | 87.9 | 7.3 | 87.3 |
| **0.75** | 93.7 | 12.4 | 12.9 | 93.4 | 6.3 | 87.6 |
| **0.8** | 95.7 | 12.2 | 7.7 | 97.4 | 4.3 | 87.8 |
| **0.85** | 93.2 | 11.8 | 2.9 | 98.4 | 6.8 | 88.2 |
| **0.9** | 100.0 | 11.7 | 1.0 | 100.0 | 0.0 | 88.3 |
| **0.95** | 100.0 | 11.7 | 0.0 | 100.0 | 0.0 | 88.3 |

Table 3.7: Predictive Ability at Alternative Thresholds (Victim/Damage Offenses)

| Cut-point | Positive Predictive Value (%) | Negative Predictive Value (%) | Sensitivity (%) | Specificity (%) | False Positives (%) | False Negatives (%) |
|---|---|---|---|---|---|---|
| **0.05** | 88.2 | - | 100.0 | 0.0 | 11.8 | - |
| **0.1** | 88.2 | - | 100.0 | 0.0 | 11.8 | - |
| **0.15** | 88.2 | - | 100.0 | 0.0 | 11.8 | - |
| **0.2** | 88.3 | 60.0 | 99.9 | 1.0 | 11.7 | 40.0 |
| **0.25** | 88.6 | 29.3 | 98.2 | 5.5 | 11.4 | 70.7 |
| **0.3** | 89.1 | 24.4 | 94.2 | 13.9 | 10.9 | 75.6 |
| **0.35** | 89.7 | 21.2 | 87.3 | 25.5 | 10.3 | 78.8 |
| **0.4** | 90.6 | 20.6 | 80.4 | 37.7 | 9.4 | 79.4 |
| **0.45** | 91.0 | 18.3 | 71.9 | 46.8 | 9.0 | 81.7 |
| **0.5** | **91.8** | **17.5** | **63.7** | **57.4** | **8.2** | **82.5** |
| **0.55** | 92.8 | 16.3 | 51.5 | 70.3 | 7.2 | 83.7 |
| **0.6** | 93.5 | 15.1 | 40.0 | 79.4 | 6.5 | 84.9 |
| **0.65** | 93.5 | 14.0 | 30.1 | 84.5 | 6.5 | 86.0 |
| **0.7** | 94.9 | 13.5 | 21.0 | 91.6 | 5.1 | 86.5 |
| **0.75** | 95.6 | 12.9 | 13.2 | 95.5 | 4.4 | 87.1 |
| **0.8** | 98.4 | 12.6 | 7.9 | 99.0 | 1.6 | 87.4 |
| **0.85** | 97.3 | 12.1 | 3.1 | 99.4 | 2.7 | 87.9 |
| **0.9** | 100.0 | 11.9 | 1.0 | 100.0 | 0.0 | 88.1 |
| **0.95** | 88.2 | 0.0 | 100.0 | 0.0 | 11.8 | 100.0 |

Table 3.8: Post-Random Assignment Serious Offending in Experimental Sample

| | | Treatment Group (N=800) | Control Group (N=759) |
|---|---|---|---|
| **Model-Defined** | % Post | 2.5 | 4.1 |
| | Risk Ratio | .61 | |
| **UCR Part I** | % Post | 8.0 | 9.6 |
| | Risk Ratio | .83 | |
| **Victim/Damage** | % Post | 6.9 | 8.7 |
| | Risk Ratio | .79 | |

No significant differences ($\chi^2$).

Table 3.9: Post-Random Assignment Offending Severity Stratified by Prior History

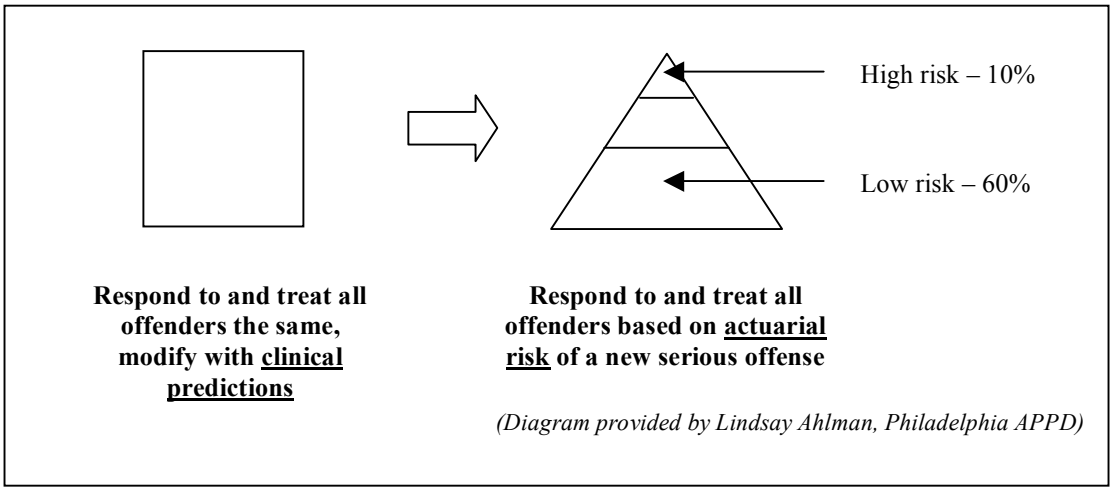
| | | No Serious Offense Pre-RA | | Serious Offense Pre-RA | |
|---|---|---|---|---|---|
| | | **Treatment** | **Control** | **Treatment** | **Control** |
| **Model-Defined** | N | 564 | 547 | 236 | 212 |
| | % Post | 2.0 | 4.0 | 3.8 | 4.2 |
| | Stratum Risk Ratio | .49* | | .90 | |
| | M-H Adjusted RR[a] | .61 | | | |
| **UCR Part I** | N | 273 | 245 | 527 | 514 |
| | % Post | 3.7 | 6.5 | 10.2 | 11.1 |
| | Stratum Risk Ratio | .56 | | .92 | |
| | M-H Adjusted RR | .84 | | | |
| **Victim/ Damage** | N | 279 | 263 | 521 | 496 |
| | % Post | 3.9 | 6.8 | 8.4 | 9.7 |
| | Stratum Risk Ratio | .58 | | .87 | |
| | M-H Adjusted RR | .79 | | | |

* $p \leq .05$, $\chi^2$.

[a] M-H: Mantel-Haenszel; RR: Risk Ratio.

## Figures

Figure 3.1: Philadelphia APPD's Risk-Based Caseload Stratification



# APPENDICES

## Appendix A: Systematic Review Search Strategy

List of Online Databases

1. Australian Criminology Database (CINCH)
2. Campbell Collaboration Social, Psychological, Educational, and Criminological Trials Register (C2-SPECTR)
3. Criminal Justice Abstracts
4. Dissertation Abstracts
5. Google, Google Scholar, Google Books
6. Government Publications Office Monthly Catalog
7. International Bibliography of the Social Sciences
8. ISI Web of Knowledge
9. JSTOR
10. National Criminal Justice Reference Service (NCJRS) Abstracts
11. PsycINFO
12. Sage Full Text Collection: Criminology
13. Sage Full Text Collection: Political Science
14. Sage Full Text Collection: Sociology
15. Social Science Citation Index
16. Social Services Abstracts
17. Sociological Abstracts
18. Worldwide Political Science Abstracts

List of Research Organizations and Government Department Websites

1. American Correctional Association
2. American Probation and Parole Association
3. Home Office Research, Development, and Statistics (U.K.)
4. International Community Corrections Association
5. Ministry of Justice (U.K.)
6. National Association for the Care and Resettlement of Offenders (U.K.)
7. National Institute of Corrections (U.S.A.)
8. National Institute of Justice (U.S.A.)
9. National Offender Management Service (U.K.)
10. National Probation Service (U.K.)
11. Pew Center on the States (U.S.A.)
12. RAND Corporation (chiefly U.S.A.)
13. Swedish National Council on Crime Prevention (BRÅ)
14. Urban Institute (U.S.A.)
15. Vera Institute of Justice (U.S.A.)
16. Washington State Institute of Public Policy (U.S.A.)

Keywords

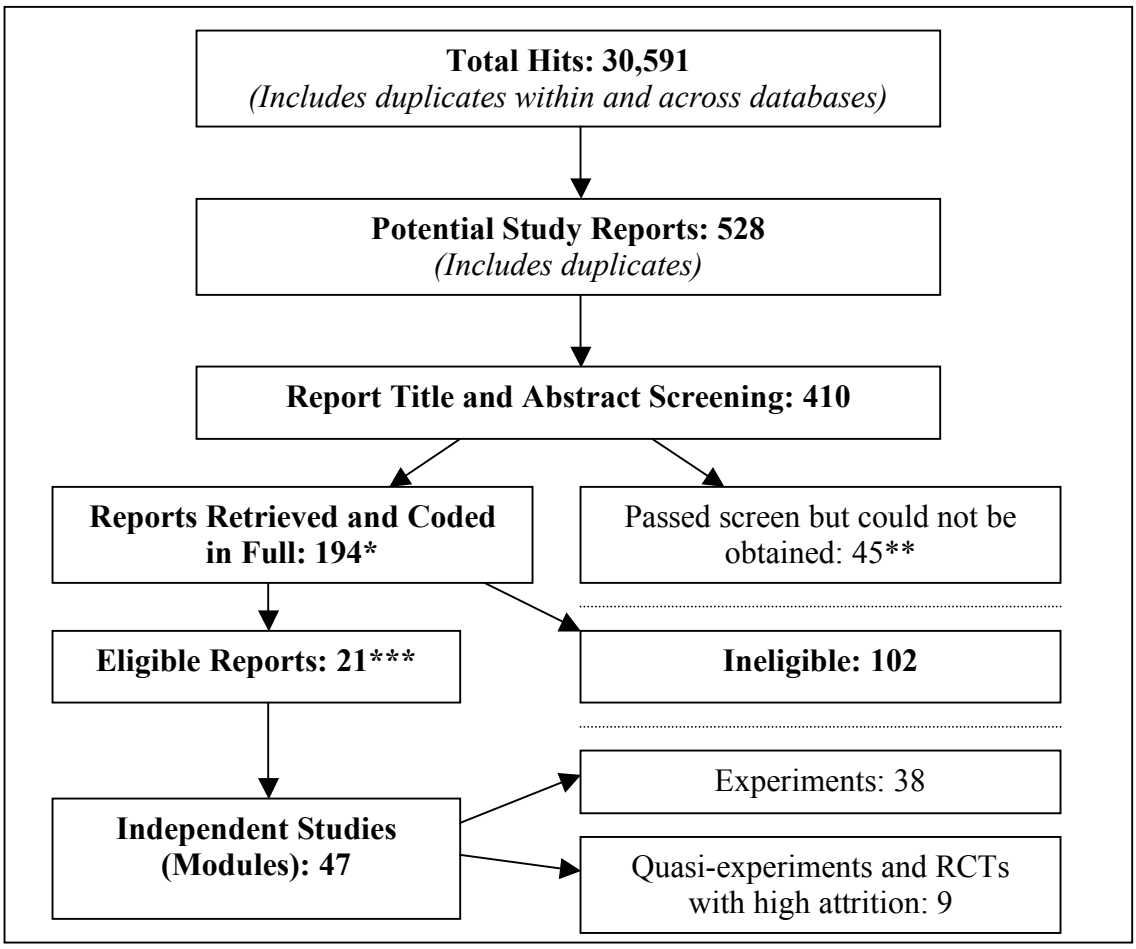
The following search strings of key words were used to search the databases and websites, adapted as necessary to meet the requirements of the different search engines. Where including all the search terms was problematic, we opted for the broadest possible

combination. The search terms were deliberately left broad (they do not include limiting terms such as 'evaluation,' 'experiment,' 'trial') so that relevant background literature could also be systematically obtained through the searches. '*' indicates where terms were truncated to find all possible variants of the word:

*probation* AND supervis* AND case* AND (intens* OR frequen* OR ratio)*

*AND (recidiv* OR *arrest* OR *convict*)*

Electronic Search Results




* Reports not classed 'eligible' or 'ineligible contained supplementary information on eligible studies or relevant background literature.
** Some of these reports may include information that was obtained from other retrieved documents.
*** Includes some studies obtained from sources other than the electronic search.

# Appendix B: Systematic Review Coding Protocol

## A. STUDY LEVEL CODING SHEET

Instructions: One study level coding sheet to be used per study. If the study is reported in multiple documents, use the primary publication as the study identifier and list other document numbers below.

A1. Study ID:_____ studid

A2. Cross-ref document ID:_____ xref1

A3. Cross-ref document ID:_____ xref2

A4. Cross-ref document ID:_____ xref3

A5. Coder initials:_____ coder

A6. Date coded:_____ codate

A7. Title:_____ title

A8. Author(s):_____ author

A9. Publication type: pubtype

1. Book
2. Book chapter
3. Peer-reviewed journal article
4. Government report (federal)
5. Government report (state/local)
6. Unpublished (e.g., dissertation, technical report, conference paper
8. Other: _____

A10. Journal ref. (vol., issue):_____ jref

A11. Publication year:_____ pubyr

A12. Date range of **research**:_____ resdate

A13. Country of publication:_____ publoc

A14. Country of study setting:_____ resloc

A15. Number of treatment-comparison contrasts in report:_____ mods

*Only independent treatment group samples should be counted; see Instructions for Section B. If no comparison group, just complete* ***B. ELIGIBILITY CHECKLIST.***

A16. Is the same comparison group used in each contrast? cxlmod

0. No
1. Yes
8. N/A

**B. ELIGIBILITY CHECKLIST**

B1. First author's last name:_____ elname

B2. Coder initials:_____ coelig

B3. Date eligibility determined:_____ eldate

**To be eligible, a study must meet the following criteria. Answer each question with 1 = Yes, 0 = No.**

B4. The study evaluates an intensive probation or parole program involving increased supervision by probation officers in a reduced caseload, or low-intensity probation (increased caseload, less supervision). 1. Yes 0. No evpro

B5. A difference in probation intensity between the treatment and comparison groups, as evidenced by a change in caseload size, ratio of clients to officers, or other control measures, is a key component of the overall program. 1. Yes 0. No evsep

B6. The study includes a comparison group receiving 'standard probation,' not comprised of dropouts from ISP/low intensity, or other supervision by probation officer (not incarcerated controls). Study design may be experimental or quasi-experimental, but not a one-group research design. 1. Yes 0. No evcomp

B7. The study includes a post-program measure of criminal behavior (arrest, conviction) or technical violation of probation/parole – may be official or self-reported and dichotomous or continuous. 1. Yes 0. No evoutc

**For documents that do not meet the above criteria, answer the following questions:**

B8. Document is not a quantitative evaluation (no data regarding effects of ISP/LIP reported). 1. Yes 0. No evndat

B9. Document is a review article relevant to this project (e.g., references to studies, background information for write-up). 1. Yes 0. No evusef

B10. Document status (circle one): elstat

1. Eligible
0. Not eligible
9. Relevant review

Notes:

## C. TREATMENT-COMPARISON CODING SHEET

Instructions: If the study reports on multiple treatment-comparison contrasts, or multiple treatments compared to a single comparison group, each contrast should be coded on separate Treatment-Comparison Coding Sheets. Only independent evaluations should be included in analyses (i.e., multiple treatment groups should not have overlapping participants).

**Identifying Information**

C1. Study ID:_____ studid

C2. Module ID:_____ modid

C3. Coder initials:_____ comod

**Program Details**

C4. Description of what happens to treatment group:_____ txdesc

_____

_____

C5. Description of what happens to control group:_____ cxldesc

_____

_____

C6. Primary program type: progtype
1. Increase in probation intensity
2. Decrease in probation intensity
8. Other:_____

C6a. If increased intensity, what was the precise nature of the program? progdesc
1. 'Front door' prison diversion (probation instead of prison)
2. 'Backdoor' prison diversion (early release from prison)
3. Enhanced probation
4. Enhanced parole
5. Enhanced probation and parole
8. Other:_____

C6b. Primary program components *(indicate whether present or not)*:

| | | | |
|---|---|---|---|
| Program increases ratio of clients to probation officers | 1. Yes | 0. No | progir |
| Program decreases ratio of clients to probation officers | 1. Yes | 0. No | progdr |
| Program increases frequency of contact with probation officer | 1. Yes | 0. No | progif |
| Program decreases frequency of contact with probation officer | 1. Yes | 0. No | progdf |
| Program increases drug testing requirements | 1. Yes | 0. No | progidt |
| Program decreases drug testing requirements | 1. Yes | 0. No | progddt |
| Other:_____ | 1. Yes | 0. No | progoth |

C6c. If Yes for any of the above, state exact numbers if available *(999 if not)*:

| | |
|---|---|
| Control ratio:_____ / Treatment ratio:_____ | racxl/ratx |
| Control freq:_____ / Treatment freq:_____ | frcxl/frtx |
| Control drug tests:_____ / Treatment drug tests:_____ | drcxl/drtx |

Other:_____ txcxloth

C6d. Additional program components *(indicate whether present or not)*:

| | | |
|---|---|---|
| Curfew | 1. Yes 0. No | addcomp_curf |
| Drug treatment | 1. Yes 0. No | addcomp_drug |
| Electronic monitoring | 1. Yes 0. No | addcomp_em |
| Employment program/assistance | 1. Yes 0. No | addcomp_cmpl |
| Halfway house | 1. Yes 0. No | addcomp_hh |
| Home visits | 1. Yes 0. No | addcomp_hv |
| House arrest | 1. Yes 0. No | addcomp_harr |
| Offense-specific treatment (e.g., sex offender treatment) | 1. Yes 0. No | addcomp_offtx |
| Other treatment | 1. Yes 0. No | addcomp_tx |
| Other:_____ | 1. Yes 0. No | addcomp_oth |

C7. What happened to the comparison group? cxltype
1. 'Supervision as usual'
8. Other:_____

C8. Was supervision for treatment group provided by anyone other than probation officer? posup
0. No
1. Yes (explain):_____
9. Don't know/can't tell

C9. Length of intervention in months (weeks/4.3):
Minimum:_____ txlmin
Maximum:_____ txlmax
Mean:_____ txlmn
Fixed (same for all subjects):_____ txlfix

C10. Did the intervention follow a set protocol? txprot
0. No
1. Yes
9. Don't know/can't tell

C11. What supervision philosophy was stated? txphil
1. Control/surveillance
2. Treatment
3. Hybrid
8. Other:_____
_____
9. Don't know/not stated

C12. Did the intervention remain consistent over time? txcons
0. No
1. Yes
9. Don't know/can't tell

**Methodological Rigor**

C13. Control variables used in statistical analyses to account for initial group differences?
0. No cxlvars
1. Yes

C14. Subject-level matching? matched

0. No
1. Yes

C15. Random assignment to conditions? rassgt
0. No
1. Yes

C16. Measurement of prior criminal involvement? prior
0. No
1. Yes

C17. Rating of initial similarity between treatment and control group: prsim

| 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|

*(1 = Nonrandomized; high likelihood of baseline differences between groups or known differences related to future recidivism)*
*(5 = Nonrandomized design with strong evidence of initial equivalence)*
*(7 = Randomized design with large N or small N design with matching)*

C18. Was attrition discussed in the report? attrep
0. No
1. Yes

C19. Is there a potential threat to generalizability from overall attrition? attgen
0. No
1. Yes

C20. Is there a potential threat to internal validity from differential attrition? attint
0. No
1. Yes

C21. Did the statistical analysis attempt to control for differential attrition effects? attstat
0. No
1. Yes
9. Don't know/can't tell

C22. Statistical significance testing used? sigtest
0. No
1. Yes

C23. Overall methodology rating methrat
1. Comparison group lacks demonstrated comparability to treatment group
2. Comparison between 2+ groups, one with and one without the intervention
3. Comparison between program group and one or more control groups, controlling for other factors, or nonequivalent comparison group is only slightly different from program group, or randomized controlled trial with high attrition
4. Random assignment and analysis of comparable program and comparison groups, including controls for attrition

Notes on methodology:

**D. SAMPLE LEVEL CODING SHEET**

Instructions: A study may report results separately for distinct samples (e.g., persons with/without prior arrests). Each distinct sample must have its own coding sheet. The treatment-comparison contrast is the same for the different samples.
Samples should be independent; i.e., no overlapping participants. Some studies report the results broken down by different subgroups (e.g., by gender). Only one of these breakouts can be used – choose the one with the most information, or the one most relevant to the review.

**Identifying Information**

D1. Study ID:_____ studid

D2. Module ID:_____ modid

D3. Sample ID:_____ sampid

D4. Coder initials:_____ cosamp

**Sample Description**

D5. Description of treatment group sample:_____ txsamp

_____

_____

D6. Description of comparison group sample:_____ cxlsamp

_____

_____

D7. Total N in treatment group at beginning of study:_____ txn

D8. Total N in comparison group at beginning of study:_____ cxln

*Note: D7 + D8 = total sample size prior to attrition. If multiple samples are being coded, the sum across samples must equal the total sample size prior to attrition.*

D9. Age range of study participants: sampage

1. Adolescent (12-18)
2. Youth (18-21)
3. Adult (21+)
4. Adolescent and youth
5. Youth and adult
6. Adolescent, youth, and adult
8. Other:_____
9. Unspecified/can't tell

D10. Youngest age included in sample *(999 if unknown)*:_____ yage

D11. Oldest age included in sample *(999 if unknown)*:_____ oage

D12. Exact proportion of males in sample (if known):_____ prmale

D13. Approximate gender description of sample: sampgen

1. All male (>90%)
2. More males than females (60-90% male)
3. Roughly equal males and females
4. More females than males (60-90% female)
5. All female (>90%)
9. Can't tell

D14. Race/ethnicity of sample *(999 if unknown)*:

% Asian:_____ rasian % Native American:_____ rnative
% Black:_____ rblack % White:_____ rwhite
% Hispanic:_____ rhisp % Other:_____ rother

D15. General offender type: offtype

1. Violent and/or person crimes
2. Nonviolent and/or nonperson crimes
3. Mixed: violent/nonviolent
4. Specialized caseload: drugs
5. Specialized caseload: sex offenders
6. Specialized caseload: mental health
7. Specialized: domestic violence
8. Other:_____
9. Don't know/can't tell

D16. Composition of supervised offenders: offcomp

1. All probationers
2. All parolees
3. Probationers and parolees

D16a. If combination of probationers and parolees *(999 if unknown)*:

% probation:_____ pcpro % parole:_____ pcpar

D17. Probationer/parolee risk level: offrisk

1. Low risk
2. Medium risk
3. High risk
4. Low and medium risks
5. Medium and high risks
6. All risk levels
7. No risk assessment
8. Other:_____
9. Don't know/can't tell

D18. How was risk determined? riskjmt

1. Statistical model
2. Prior convictions
3. Instant offense
4. Judgment of probation officer/intake
5. Classification instrument
7. N/A
8. Other:_____
9. Don't know/can't tell

D19. Probationer/parolee need level: offneed

1. Low need
2. Medium need
3. High need
4. Low and medium need
5. Medium and high need
6. All need levels
7. No need assessment
8. Other:_____
9. Don't know/can't tell

## E. DEPENDENT VARIABLE CODING SHEET

Instructions: Code each dependent variable reported in the study separately. The same dependent variable measured at multiple times should be coded only once. For non-crime outcomes, code only items E6, E9, and E10.

### Identifying Information

E1. Study ID:_____ studid

E2. Module ID:_____ modid

E3. Sample ID:_____ sampid

E4. Outcome ID:_____ outid

E5. Coder initials:_____ coout

### Outcome Information

E6. Outcome label (label used in the report):_____ outlab

E7. Recidivism construct represented by this measure: rconst

1. Arrest
2. Charge
3. Conviction
4. Technical violation
5. Probation revocation
6. Incarceration
8. Other:_____

E8. Offense types included in recidivism measure:

| | | |
|---|---|---|
| All offenses ('No' for others) | 1. Yes 0. No | oall |
| Drug offenses | 1. Yes 0. No | odrug |
| Person offenses, sexual | 1. Yes 0. No | opsx |
| Person offenses, nonsexual | 1. Yes 0. No | opnsx |
| Person offenses, unspecified | 1. Yes 0. No | opuns |
| Property offenses | 1. Yes 0. No | oprop |
| Weapons offenses | 1. Yes 0. No | oweap |
| Driving offenses | 1. Yes 0. No | odriv |
| Technical or status offenses | 1. Yes 0. No | otech |
| Other:_____ | 1. Yes 0. No | ooth |

E9. Measurement scale: mscale

1. Dichotomous
2. Trichotomous
3. 4-9 discrete ordinal categories
4. >9 discrete ordinal categories/continuous

E10. Source of data: dsrce

1. Self-report
2. Other report (e.g., probation officer)
3. Official records (police, probation, court, etc.)
8. Other:_____
9. Don't know/can't tell

E11. Length of follow-up period: fulng

1. < 6 months
2. 6-12 months
3. > 1, < 2 years
4. > 2 years
8. No follow-up
9. Don't know/can't tell

E12. Is cost/benefit data for the program included in the study? 1. Yes 0. No

## F. EFFECT SIZE LEVEL CODING SHEET

Instructions: Complete a separate coding sheet for each treatment-comparison contrast for each dependent variable.

### Identifying Information

F1. Study ID:_____ studid

F2. Module ID:_____ modid

F3. Sample ID:_____ sampid

F4. Outcome ID:_____ outid

F5. Effect size ID:_____ esid

F6. Coder initials:_____ coes

### Effect Size Information

F7. Effect size type: estype
1. Baseline (pretest; prior to start of intervention)
2. Post-test (first measurement point, post-intervention)
3. Follow-up (all subsequent measurement points, post-intervention)

F8. Which group does the raw effect favor (ignoring statistical significance)? esdir
1. Treatment group
2. Comparison group
3. Neither (ES = 0)
9. Can't tell (ES cannot be used if selected)

F9. Does the investigator report the difference as statistically significant? essig
0. No
1. Yes
8. Not tested
9. Can't tell

F10. If tested, what type of statistical test was used? estest
1. *t* test
2. *F* test
3. $\chi^2$
4. Regression analysis
7. N/A
8. Other:_____
9. Can't tell

F11. Timeframe in months captured by the measure (weeks/4.3)
Minimum:_____ estmin
Maximum:_____ estmax
Mean: _____ estmn
Fixed (same for all subjects) _____ estfix

F12. Timeframe in months from end of program to measurement point (weeks/4.3)
Minimum:_____ esfumin
Maximum:_____ esfumax
Mean:_____ esfumn
Fixed (same for all subjects) _____ esfufix

### Effect size data – all effects

F13. Treatment group sample size for this ES:_____ estxn

F14. Comparison group sample size for this ES:_____ escxl

**Effect size data – continuous outcomes**

F15. Treatment group mean: _____ estxmn

F16. Comparison group mean: _____ escxlmn

F17. Are the above means adjusted? 1. Yes 0. No esmadj

F18. Treatment group standard deviation: _____ estxsd

F19. Comparison group standard deviation: _____ escxlsd

F20. Treatment group standard error: _____ estxse

F21. Comparison group standard error: _____ escxlse

F22. *t*-value from an independent *t*-test or square root of *F* value from a one-way ANOVA with 1 d.f. in the numerator (only 2 groups):_____ estval

F23. Exact probability for a *t*-value from an independent *t*-test or *F*-value from a one-way ANOVA with 1 d.f. in the numerator: _____ estvalp

F24. Correlation coefficient: _____ escorr

**Effect size data – dichotomous outcomes**

F25. Number successful in treatment group: _____ estxs

F26. Number successful in comparison group: _____ escxls

F27. Proportion successful in treatment group: _____ estxspr

F28. Proportion successful in comparison group: _____ escxlspr

F29. Are the above proportions adjusted for pre-test variables?
1. Yes 0. No espradj

F30. Logged odds ratio: _____ eslogor

F31. Standard error of logged odds ratio: _____ eslorse

F32. Logged odds ratio adjusted? (e.g., from logistic regression)
1. Yes 0. No esloradj

F33. $\chi^2$ value with 1 d.f. (2x2 contingency table):_____ eschisq

F34. Correlation coefficient: _____ esdcorr

**Effect size data – hand calculated**

F35. Hand calculated *d*-type effect size: _____ eshand

F36. Hand calculated SE of the *d*-type effect size:_____ eshandse

## Appendix C: Meta-Analytic Procedures

Odds ratio effect size

The odds ratio (OR) is given by the following formula, based on a 2x2 table:

| | Events | Non-Events |
|---|---|---|
| Treatment Group | *a* | *b* |
| Control Group | *c* | *d* |

$$ES_{OR} = \frac{ad}{bc} \tag{1}$$

Analyses are performed on the natural log of the OR and converted back for presentation.

Standard error of the logged OR:

$$SE_{LOR} = \sqrt{\tfrac{1}{a} + \tfrac{1}{b} + \tfrac{1}{c} + \tfrac{1}{d}} \tag{2}$$

Inverse variance weight of the logged OR (fixed effects):

$$w_{LOR} = \frac{1}{SE_{LOR}^2} \tag{3}$$

Heterogeneity (*Q*-statistic)

*Q* is distributed as a chi-square with $k - 1$ d.f.

$$Q = \sum w_i \left( ES_i - \overline{ES} \right)^2 \tag{4}$$

$ES_i$ = individual effect size for $i = 1$ to $k$ (the number of effect sizes).
$\overline{ES}$ = weighted mean effect size over the $k$ effect sizes.
$w_i$ = weight for $ES_i$.

Random effects model

A second variance component $v_\theta$ is computed, reflecting between-study error. We use the DerSimonian and Laird method of moments estimator of $v_\theta$ (5). The inverse variance

weight $w$ is then recalculated (6), with $v_\theta$ added to the within-study variance component ($v_i = SE^2_{LOR}$) from the denominator of (3) above. If $Q$ is smaller than $k - 1$ in (5), $v_\theta$ is set to 0 and the random and fixed effects weights are the same.

$$v_\theta = \frac{Q - (k-1)}{\sum w_i - \left(\sum w_i^2 / \sum w_i\right)} \tag{5}$$

$$w_{RANDOM} = \frac{1}{v_\theta + v_i} \tag{6}$$

Meta-analytic analog to the ANOVA

The total variability $Q$ is partitioned into two groups: variability within categories of the moderator variable ($Q_W$); i.e., the variability of the effect sizes around the category mean, and variability between categories ($Q_B$).

$$Q_W = \sum w_{ij}\left(ES_{ij} - \overline{ES}_j\right)^2 \tag{7}$$

$$Q_B = \sum_{j=1}^{p} (\overline{ES}_j w_j)^2 - \frac{\left(\sum_{j=1}^{p} \overline{ES}_j w_j\right)^2}{\sum_{j=1}^{p} w_j} \tag{8}$$

$j = 1$ to $c$ for $c$ categories of the independent (moderator) variable.
$\overline{ES}_j$ = weighted mean effect size for each group.
$w_j$ = sum of the weights within each group.

In the mixed effects analog to the ANOVA, the moderator variable is treated as fixed and the variability between groups is random. Thus, the estimator for the random effects component is based on $Q_W$, not $Q$ (formula not shown). $Q_W$ estimates from the model are not interpretable. $Q_B$ is the meta-analytic equivalent to the $F$-statistic.

References: Lipsey & Wilson (2001), pp. 47-49, 54, 115-116, 121.; Wilson (2010), pp. 195-198.

## Appendix D: Details of Included and Excluded Studies

Included Studies

| Study | Research Design | Population | Intervention | Comparison | Outcome |
|---|---|---|---|---|---|
| California 1957 (Reimer & Warren, 1957) | Randomized controlled trial | Adult male inmates determined eligible for parole and released. | 15:1 caseload with minimum one contact per week. | 90:1 caseload. No stated requirements for contact. Largely a test of caseload size. | Treatment group less likely to fail on convictions and technical violations. |
| Dorset 1976 (Folkard, Smith, & Smith, 1976) | Randomized controlled trial | Youth and adult male probationers on all types of probation orders. | 20:1 caseload, 1 experimental officer in each of 4 offices, intensive intervention in family, work, and social situations, offenders visited in home environment. | Approx 40-45:1 caseload, traditional probation services, around half as many contacts as treatment group. | Treatment group more likely to fail on convictions. |
| Inner London 1976 (Folkard, Smith, & Smith, 1976) | Randomized controlled trial | Youth and adult male probationers, 1+ prior conviction since age 14. | 20:1 caseload, 1 experimental officer in each of 4 offices, intensive intervention in family, work, and social situations, offenders visited in home environment. | Approx 40-45:1 caseload, traditional probation services, around half as many contacts as treatment group. | Treatment group more likely to fail on convictions. |

| Study | Research Design | Population | Intervention | Comparison | Outcome |
|---|---|---|---|---|---|
| Sheffield (Males) 1976 (Folkard, Smith, & Smith, 1976) | Randomized controlled trial | Youth and adult male probationers on 2-3 year orders with 2+ priors since age 14. | 20:1 caseload, 1 experimental officer in each of 4 offices, intensive intervention in family, work, and social situations, offenders visited in home environment. | Approx 40-45:1 caseload, traditional probation services, around half as many contacts as treatment group. | Treatment group less likely to fail on convictions. |
| Sheffield (Females) 1976 (Folkard, Smith, & Smith, 1976) | Randomized controlled trial | Youth and adult female probationers on 2-3 year orders with 1+ priors since age 14. | 20:1 caseload, 1 experimental officer in each of 4 offices, intensive intervention in family, work, and social situations, offenders visited in home environment. | Approx 40-45:1 caseload, traditional probation services, around half as many contacts as treatment group. | Treatment group more likely to fail on convictions. |
| Staffordshire 1976 (Folkard, Smith, & Smith, 1976) | Randomized controlled trial | Youth and adult male probationers on 2-3 year orders with 2+ priors since age 14. | 20:1 caseload, 1 experimental officer in each of 4 offices, intensive intervention in family, work, and social situations, offenders visited in home environment. | Approx 40-45:1 caseload, traditional probation services, around half as many contacts as treatment group. | Treatment group more likely to fail on convictions. |
| Contra Costa SOP 1986 (Fagan & Reinarman, 1986) | Randomized controlled trial | Juvenile offenders judged 'serious' and a physical threat to others. | 20:1 caseload, 6-month program, weekly contacts. Officers expected to direct clients to treatment and services based on needs. | Routine probation. Larger caseloads, monthly contacts, less opportunity to build understanding of needs. | Treatment group slightly less likely to fail on new arrests. |

| Study | Research Design | Population | Intervention | Comparison | Outcome |
|---|---|---|---|---|---|
| Contra Costa Co CA 1990 (Petersilia & Turner, 1990) | Randomized controlled trial | Adults convicted of felony and misdemeanor drug use, drug dealing, and non-violent drug-related crime. | 40:1 caseload, 1 face-to-face contact/drug test and 2 phone calls per week, gradually reduced. Employment assistance, counseling, and links with law enforcement. | 150-200:1 caseload, infrequent, discretionary contact, random drug tests, employment assistance. | Treatment group more likely to fail on technical violations and arrests. No difference for convictions. |
| Los Angeles Co CA 1990 (Petersilia & Turner, 1990) | Randomized controlled trial | Adults convicted of felonies and classified as high risk. | 33:1 caseload, 3-5 face-to-face contacts and 2 phone calls per week, gradually reducing. Monitoring checks. | 250:1 caseload, 1 face-to-face contact per month. | Treatment group more likely to fail on technical violations, arrests, and convictions. |
| Ventura Co CA 1990 (Petersilia & Turner, 1990) | Randomized controlled trial | Adults convicted of felonies and classified as high risk or convicted of serious crime. | 19:1 caseload, 4 face-to-face contacts, 2 phone calls, 1 drug test per week, plus monitoring checks, gradually reduced. Employment assistance, victim awareness, links with law enforcement. | Existing ISP. 50:1 caseload, 2 face-to-face visits and 1 phone call per month, no victim awareness. | Treatment group less likely to fail on technical violations, arrests, and convictions. |
| Atlanta GA 1992 (Petersilia, Turner, & Deschenes, 1992b) | Randomized controlled trial | Prisoners released to ISP or probationers due to be revoked and returned to prison. Drug involved, high risk/need. | Augmented ISP with increased surveillance (passive monitoring, officer checks). 40:3 caseload. Rehabilitation, employment, drug tests, monitoring of activities. Same contact level as controls. | Existing ISP. 12 face-to-face, 10 phone per month, random drug tests, job verification weekly. | Treatment group more likely to fail on technical violations, arrests, and convictions. No convictions in control group. |

| Study | Research Design | Population | Intervention | Comparison | Outcome |
|---|---|---|---|---|---|
| Dallas TX 1992 (Turner & Petersilia, 1992) | Randomized controlled trial | Adult property offenders (mostly male), initially in prison and high risk of recidivism on parole. | 25:1 caseload with 10 contacts per month – mix of in-person office/home visits & telephone. Employment assistance, discretionary drug testing. | 85:1 caseload, 1 office visit per month, occasional home visits, no drug testing requirement. | Treatment group more likely to fail on technical violations, arrests, and convictions. |
| Des Moines IA 1992 (Petersilia, Turner, & Deschenes, 1992b) | Randomized controlled trial | Probationers and parolees convicted of drug offense or burglary with drug abuse history. | 35:1 caseload. Initially 16 face-to-face, 4 phone contacts and 8 drug tests per month, gradually decreasing. Curfew. Emphasis on urinalysis, unannounced visits and collateral contact Treatment and employment mandated. | 70:1 mixed caseload, routine supervision, risk determines contact levels. Most on highest level: 2 face-to-face and 2 collateral/month. Discretionary testing. | Treatment group more likely to fail on technical violations, less likely on arrests and convictions. |
| Houston TX 1992 (Turner & Petersilia, 1992) | Randomized controlled trial | Adult property offenders (mostly male), initially in prison and high risk of recidivism on parole. | 25:1 caseload with 10 contacts per month – mix of in-person office/home visits & telephone. Employment assistance, discretionary drug testing. | 85:1 caseload, 1 office visit per month, occasional home visits, no drug testing requirement. | Treatment group more likely to fail on technical violations and arrests, but less likely on convictions. |
| Santa Fe NM 1992 (Petersilia, Turner, & Deschenes, 1992b) | Randomized controlled trial | Probationers and parolees with high risk and need. | 35:2 caseload. Initially 12 face-to-face contacts, 8 unannounced home visits, 4 drug tests per month. Therapeutic approach: counseling, job development, group therapy. | 60:1 caseload, routine supervision, 2 face-to-face and 1 office visit per month, discretionary testing and treatment referral. | Treatment group less likely to fail on technical violations, more likely on arrests and convictions. |

| Study | Research Design | Population | Intervention | Comparison | Outcome |
|---|---|---|---|---|---|
| Seattle WA 1992 (Petersilia, Turner, & Deschenes, 1992b) | Randomized controlled trial | Felony drug offenders likely to have high recidivism and drug dependent. | 20:1 caseload. Focus on treatment participation, employment and drug testing. Initially 12 face-to-face contacts and 8 drug tests per month, gradually decreasing. | Routine probation supervision. 85:1 caseload, 4 face-to-face contacts per month, additional contacts/tests discretionary. | Treatment group more likely to fail on arrests, technical violations, slightly less on convictions. |
| Waycross GA 1992 (Petersilia, Turner, & Deschenes, 1992b) | Randomized controlled trial | Prisoners released to ISP or probationers due to be revoked and returned to prison. Drug involved, high risk/need. | Augmented version of existing ISP with increased surveillance (increased drug testing). 40:3 caseload. Rehabilitation, employment, drug testing, monitoring of activities. | Existing ISP. Contact level same in both groups: 12 face-to-face and 10 phone contacts per month, random drug tests, job verification weekly. | Treatment group more likely to fail on technical violations, less likely on arrests. No convictions in either group. |
| Winchester VA 1992 (Petersilia, Turner, & Deschenes, 1992b) | Randomized controlled trial | High risk probationers and parolees with drug offenses and history of abuse. | 24:1 caseload. Initially 12 face-to-face contacts, 4 phone contacts, 4 monitoring checks per month. Discretionary drug testing. Counseling. Halfway house and detox. | Routine supervision: 80:1 caseload. 2 face-to-face and 2 phone contacts per month. Other contact and testing discretionary. | Treatment group more likely to fail on technical violations, arrests, and convictions. |
| Philadelphia PA 1993 (Sontheimer & Goodstein, 1993) | Randomized controlled trial | Serious juvenile offenders | 12:1 caseload, weekly contact with collateral and out of hours contacts mandated. Employment assistance and education. | 70-100:1 caseload, monthly visits, limited access to services. | Treatment group less likely to fail on new arrests. |

| Study | Research Design | Population | Intervention | Comparison | Outcome |
|---|---|---|---|---|---|
| Minnesota ISR 1995 (Deschenes, Turner, & Petersilia, 1995) | Randomized controlled trial | Parolees considered a "public risk monitoring case" (high risk). | 12-15:1 caseload, 4 contacts/month reduced over time, period of house arrest followed by curfew, mandatory 40 hours/week of work, drug treatment or education, random weekly drug tests. | Regular parole supervision. Larger caseloads, fewer contacts (approx. 3 per month). | Treatment group less likely to fail on technical violations and arrests; slightly more likely on convictions. |
| Midwest 1998 (Latessa et al., 1998) | Randomized controlled trial | Rural population, mostly white males (youth and adult), higher risk and need levels. | 20:1 caseload, about 10 contacts a month, drug testing, treatment referrals, probation officers trained in CBT. | 100-200:1 caseload, less frequent contact (about half that in treatment group), no services. | Treatment group more likely to fail on technical violations but less likely to fail on arrests. |
| Northeast 1998 (Latessa et al., 1998) | Randomized controlled trial | Urban population, mostly males, youth and adult, higher risk and need levels. | 25:1 caseload, 1-2 contacts per week, drug testing/treatment, home and office visits, brokerage for needs assessment and services. | 200:1 caseload, less frequent contacts (about half the number in the treatment group, no services). | Treatment group more likely to fail on technical violations but less likely to fail on arrests. |
| CA Drug Test 2002 (Haapanen & Britton, 2002) | Randomized controlled trial | Youthful parolees. Generally serious offenders with substance abuse and other problems. | Parole with enhanced routine unscheduled drug testing requirements: 1 every week in first 90 days of parole then biweekly. | Routine parole with drug testing only after arrest. | Treatment group more likely to fail on new arrests. |

| Study | Research Design | Population | Intervention | Comparison | Outcome |
|---|---|---|---|---|---|
| San Diego CA 2002 (Howard et al., 2002) | Randomized controlled trial | Youths age 15 ½ and under, wards of court, high risk, school behavior issues and family problems. | 15:1 caseload. Intensive supervision and services, whole family interventions, multi-agency working, officers and service providers. Education, home visits, treatment, education, curfew. | 40-50:1 caseload, 1 contact per month. Regular probation or anti-gang unit. | Treatment group more likely to fail on arrests and technical violations after 6 months. |
| Benton/ Franklin WA 2003 (Barnoski, 2003) | Randomized controlled trial | High risk, first time juvenile offenders. | 25:1 caseload, team approach to supervision, individualized case plans for accountability and services. Programs varied across sites. | Traditional probation supervision, 30-100:1 caseload (varied across sites). | Treatment group more likely to fail on felony convictions. |
| Chelan/ Douglas WA 2003 (Barnoski, 2003) | Randomized controlled trial | High risk, first time juvenile offenders. | 25:1 caseload, team approach to supervision, individualized case plans for accountability and services. Programs varied across sites. | Traditional probation supervision, 30-100:1 caseload (varied across sites). | Treatment group more likely to fail on felony convictions. |
| Clallam WA 2003 (Barnoski, 2003) | Randomized controlled trial | High risk, first time juvenile offenders. | 25:1 caseload, team approach to supervision, individualized case plans for accountability and services. Programs varied across sites. | Traditional probation supervision, 30-100:1 caseload (varied across sites). | Treatment group less likely to fail on felony convictions. |

| Study | Research Design | Population | Intervention | Comparison | Outcome |
|---|---|---|---|---|---|
| Clark WA 2003 (Barnoski, 2003) | Randomized controlled trial | High risk, first time juvenile offenders. | 25:1 caseload, team approach to supervision, individualized case plans for accountability and services. Programs varied across sites. | Traditional probation supervision, 30-100:1 caseload (varied across sites). | Treatment group less likely to fail on felony convictions. |
| Cowlitz WA 2003 (Barnoski, 2003) | Randomized controlled trial | High risk, first time juvenile offenders. | 25:1 caseload, team approach to supervision, individualized case plans for accountability and services. Programs varied across sites. | Traditional probation supervision, 30-100:1 caseload (varied across sites). | Treatment group less likely to fail on felony convictions. |
| King WA 2003 (Barnoski, 2003) | Randomized controlled trial | High risk, first time juvenile offenders. | 25:1 caseload, team approach to supervision, individualized case plans for accountability and services. Programs varied across sites. | Traditional probation supervision, 30-100:1 caseload (varied across sites). | Treatment group more likely to fail on felony convictions. |
| Kitsap WA 2003 (Barnoski, 2003) | Randomized controlled trial | High risk, first time juvenile offenders. | 25:1 caseload, team approach to supervision, individualized case plans for accountability and services. Programs varied across sites. | Traditional probation supervision, 30-100:1 caseload (varied across sites). | Treatment group less likely to fail on felony convictions. |

| Study | Research Design | Population | Intervention | Comparison | Outcome |
|---|---|---|---|---|---|
| Pierce WA 2003 (Barnoski, 2003) | Randomized controlled trial | High risk, first time juvenile offenders.. | 25:1 caseload, team approach to supervision, individualized case plans for accountability and services. Programs varied across sites. | Traditional probation supervision, 30-100:1 caseload (varied across sites). | Treatment group less likely to fail on felony convictions. |
| Skagit WA 2003 (Barnoski, 2003) | Randomized controlled trial | High risk, first time juvenile offenders. | 25:1 caseload, team approach to supervision, individualized case plans for accountability and services. Programs varied across sites. | Traditional probation supervision, 30-100:1 caseload (varied across sites). | Treatment group less likely to fail on felony convictions. |
| Snohomish WA 2003 (Barnoski, 2003) | Randomized controlled trial | High risk, first time juvenile offenders. | 25:1 caseload, team approach to supervision, individualized case plans for accountability and services. Programs varied across sites. | Traditional probation supervision, 30-100:1 caseload (varied across sites). | Treatment group less likely to fail on felony convictions. |
| Spokane WA 2003 (Barnoski, 2003) | Randomized controlled trial | High risk, first time juvenile offenders. | 25:1 caseload, team approach to supervision, individualized case plans for accountability and services. Programs varied across sites. | Traditional probation supervision, 30-100:1 caseload (varied across sites). | Treatment group less likely to fail on felony convictions. |

| Study | Research Design | Population | Intervention | Comparison | Outcome |
|---|---|---|---|---|---|
| Whatcom WA 2003 (Barnoski, 2003) | Randomized controlled trial | High risk, first time juvenile offenders. | 25:1 caseload, team approach to supervision, individualized case plans for accountability and services. Programs varied across sites. | Traditional probation supervision, 30-100:1 caseload (varied across sites). | Treatment group more likely to fail on felony convictions. |
| San Francisco 2008 (Guydish et al., 2008) | Randomized controlled trial | Adult female offenders with substance abuse problem. | 50:1 caseload, 2 office, home, or phone contacts per month. Therapeutic/ advocacy orientation, service referral, counseling, treatment, employment assistance. | 100-150:1 caseload, traditional probation supervision. | Treatment group more likely to fail on new arrests. |
| Hawaii HOPE RCT 2009 (Hawken & Kleiman, 2009) | Randomized controlled trial | Drug involved probationers at high risk of failing probation. | Increased drug testing: random tests at least once a week with brief incarceration and non-revocation for failure. | Routine probation with monthly drug tests as part of contact schedule. | Treatment group less likely to fail on new arrests. |
| New Jersey 1988 (Pearson, 1988) | Quasi-experiment (matched sample design) – "Close OTI" only. | Mostly male, convicted of serious nonviolent felonies and previously incarcerated. | No explicit caseload size or contact requirements. Intensive surveillance and services. Median 31 face-to-face contacts, drug tests and curfew checks per month. Community service, employment, offense-specific treatment. | Ordinary parole. No further details, except that contact frequency was much lower than treatment group levels. | Treatment group less likely to fail on convictions after two years. |

| Study | Research Design | Population | Intervention | Comparison | Outcome |
|---|---|---|---|---|---|
| Virginia ISP 1994 (Orchowsky, Merritt, & Browning, 1994) | Quasi-experiment (matched pairs design) | Adults 18 and over at high risk to community and in need of services. | 20:1 caseload. Weekly contact. Home visits, employment verification, collateral contacts, record checks. | Routine probation. Average 68:1 caseload, monthly contact. | Treatment group more likely to fail on arrests. |
| Maryland 2003 (Piquero, 2003) | Quasi-experiment (matched pairs design) | Adult probationers and parolees in program neighborhoods. | No explicit caseload or contact requirements. Intensive community probation: officers in neighborhood offices and conducting home visits and curfew checks along with police officers. | Caseload and contacts not specified. Traditional probation/parole supervision. | Treatment group less likely to fail on arrests, slightly more likely to fail on technical violations. |
| Clark Co NV 2005 (Wiebush et al., 2005) | Randomized controlled trial with high attrition. | High risk male juvenile parolees originally placed in specific correctional facility. | 18:1 caseload, 3.1 contacts per month average later reduced to 1. Drug treatment, education, home visits, house arrest,, reentry services, treatment linkages in community. | 35+:1 caseload, 1.9 contacts per month average, regular parole services. | No difference in arrests. Treatment group less likely to fail on convictions, more likely on technical violations. |
| Denver CO 2005 (Wiebush et al., 2005) | Randomized controlled trial with high attrition. | High risk male juvenile parolees originally placed in specific correctional facility. | 18:1 caseload, 3.1 contacts per month average later reduced to 1. Drug treatment, curfew, reentry services, treatment linkages in community. | 35.1 caseload, 1.7 contacts per month average. Caseloads also reduced to allow access to services. | Treatment group more likely to fail on arrest and convictions, less likely on technical violations. |

| Study | Research Design | Population | Intervention | Comparison | Outcome |
|---|---|---|---|---|---|
| Los Angeles CA 2005 (Zhang & Zhang, 2005) | Randomized controlled trial with high attrition. | Juveniles under court supervision, meeting risk factors for delinquency, drug use and school or family problems. | 25:1 caseload, increased frequency of contact, strong focus on education and ensuring access to social services. | 100-150:1, regular probation supervision. | Treatment group more likely to fail on technical violation but less likely on arrest after 6 months – eventually no difference. |
| New Jersey ISSP 2005 (Paparozzi & Gendreau, 2005) | Quasi-experiment (matched sample design) | Mixed youth and adult parolee sample – high risk/need. | 20-25:1 caseload, treatment-focused program. Drug treatment, employment assistance, public and family provide assistance. | 75-85:1 caseload. Regular parole – also received services but less so. | Treatment group more likely to fail on technical violation but less likely to be revoked for new conviction. |
| Norfolk VA 2005 (Wiebush et al., 2005) | Randomized controlled trial with high attrition. | High risk male juvenile parolees originally placed in specific correctional facility. | 15:1 caseload, 10 contacts per month average gradually decreasing. Drug treatment, education, home visits, reentry services, treatment linkage. | Caseload not specified. Regular parole, including services (less than treatment group). | Treatment group less likely to fail on arrests and convictions, more likely on technical violations. |
| Maryland PCS 2006 (Taxman, Yancey, & Bilanin, 2006) | Quasi-experiment (matched pairs design) | Mostly probationers and some parolees, high risk for recidivism. | 55:1 caseload, emphasis on nature and content of contacts. Officer facilitates behavior management: treatment and pro-social activity. | 100:1 caseload, regular probation/parole supervision. | Treatment group less likely to fail on new arrests and violation warrants. |

Articles Containing Supplementary Details of Included Studies

| Main Study Reference | Additional Documents |
|---|---|
| **Deschenes, Turner, & Petersilia (1995)** | **Deschenes et al. (1995).** A dual experiment in intensive community supervision: Minnesota's prison diversion and enhanced supervised release programs. *Pris. J., 75,* 330-56. |
| **Fagan & Reinarman (1986)** | **Fagan & Reinarman (1991)** The social context of intensive supervision: organizational and ecological influences on community treatment. *In Armstrong (ed.): Intensive interventions with high-risk youths: promising approaches in juvenile probation and parole*, 341-394. |
| **Folkard, Smith, & Smith (1976)** | **Folkard et al. (1974).** *IMPACT Intensive Matched Probation and After-Care Treatment. Volume I: The design of the probation experiment and an interim evaluation.* Home Office Research Study 24. London: HMSO. |
| **Guydish et al. (2008)** | **Chan et al. (2005)**. Evaluation of Probation Case Management (PCM) for drug-involved women offenders. *Crime & Delinquency, 51,* 447-69. |
| **Pearson (1988)** | **Pearson (1985).** New Jersey's Intensive Supervision Program: a progress report. *Crime & Delinquency, 31,* 393-410.<br><br>**Pearson & Bibel (1986)**. New Jersey's Intensive Supervision Program: What is it like? How is it working? *Fed. Prob. 50*, 25-31.<br><br>**Pearson (1987)**. Final report of research on New Jersey's Intensive Supervision Program. Rutgers University. |
| **Petersilia & Turner (1990)** | **Petersilia (1989)**. Implementing randomized experiments: lessons from BJA's intensive supervision project. *Eval. Rev., 13*, 435-458.<br><br>**Petersilia & Turner (1990)**. Comparing intensive and regular supervision for high-risk probationers: early results from an experiment in California. *Crime & Delinq., 36,* 87-111.<br><br>**Petersilia & Turner (1991).** An evaluation of intensive probation in California. *J. Crim. L. Criminol., 82,* 610-58.<br><br>**Petersilia & Turner (1993)**. Intensive probation and parole. *Crime & Just., 17,* 281-335. |
| **Petersilia, Turner, & Deschenes (1992b)** | **Petersilia, Turner, & Deschenes (1992)**. The costs and effects of intensive supervision for drug offenders. |

| | |
|---|---|
| | *Fed. Prob., 56*, 12-17.<br><br>**Turner, Petersilia, & Deschenes (1992)**. Evaluating Intensive Supervision Probation/Parole (ISP) for drug offenders. *Crime & Delinq., 38,* 539-556.<br><br>**Petersilia & Turner (1993).** Evaluating Intensive Supervision Probation/Parole: results of a nationwide experiment. *NIJ Research in Brief.* |
| **Sontheimer & Goodstein (1993)** | **Goodstein & Sontheimer (1997)**. The implementation of an intensive aftercare program for serious juvenile offenders: a case study. *Crim. Just. Behav., 24,* 332-59. |
| **Taxman, Yancey, & Bilanin (2006)** | **Sachwald, Eley, & Taxman (2006).** An ounce of prevention: proactive community supervision reduces violation behavior. *Topics in Community Corrections,* 31-8.<br><br>**Taxman (2006).** A behavioral management approach to supervision: preliminary findings from Maryland's Proactive Community Supervision (PCS) pilot program. *Committee on Law & Just./Nat. Res. Council.*<br><br>**Taxman (2006)**. The role of community supervision in addressing reentry from jails. *Urban Institute/John Jay College/Montgomery Co., MD Reentry Roundtable.*<br>**Taxman (2007).** Reentry and supervision: one is impossible without the other. *Corrections Today (April)*, 98-105.<br><br>**Taxman (2008)**. No illusions: offender and organizational change in Maryland's Proactive Community Supervision efforts. *Crim. Pub. Pol., 7,* 275-302. |
| **Wiebush et al. (2005)** | **Wiebush, McNulty, & Le (2000).** Implementation of the Intensive community-based Aftercare Program. *OJJDP Juvenile Justice Bulletin. U.S. Dept. of Justice, Office of Justice Programs.* |

Excluded Studies

| Study Reference | Reason for Exclusion |
|---|---|
| **Adams (2001)**<br>Specialized sex offender probation in Cook County links supervision, treatment.<br>*IL Criminal Justice Information Auth. "On Good Authority," 4(7),* March. | Unmatched controls. |
| **Adams & Vetter (1971)**<br>Probation caseload size and recidivism rate.<br>*Brit. J. Criminol., 11,* 390-3. | Unmatched controls. |
| **Agopian (1990)**<br>The impact of intensive supervision probation on gang-drug offenders.<br>*Criminal Justice Policy Rev., 4,* 214-22. | Outcomes are only reported for the treatment group. |
| **Altschuler & Armstrong (1994)**<br>Intensive aftercare for high-risk juveniles: policies and procedures. Program summary.<br>*US Department of Justice, Office of Juvenile Justice & Delinquency Prevention.* | No evaluation data reported. |
| **Austin, Quigley, & Cuvelier (1989)**<br>Evaluating the impact of Ohio's community corrections programs: public safety and costs.<br>*National Council on Crime & Delinquency.* | Unmatched controls. |
| **Barnoski (2000)**<br>Intensive parole model for high risk juvenile offenders: interim outcomes for the first cohort of youth.<br>*Washington State Inst. for Public Policy* | Unmatched controls. |
| **Barton & Butts (1990)**<br>Viable options: intensive supervision programs for juvenile delinquents.<br>*Crime & Delinquency, 36,* 238-256.<br><br>**Barton & Butts (1991)**<br>Intensive supervision alternatives for adjudicated juveniles.<br>*In Armstrong (ed.): Intensive interventions with high-risk youths: promising approaches in juvenile probation and parole*, 317-340. | Control group did not receive regular probation supervision (90% incarcerated). |
| **Bayens, Manske & Smykla (1998)**<br>The impact of the 'new penology' on ISP<br>*Criminal Justice Rev., 23,* 51-62. | Unmatched controls.<br>No crime outcomes. |

| Study Reference | Reason for Exclusion |
|---|---|
| **Benekos & Sonnenberg (1998)**<br>An evaluation of Erie County intermediate punishment programs.<br>*PA Commission on Crime & Delinquency.* | Not a test of probation supervision intensity. |
| **Bennett (1987)**<br>A reassessment of intensive service probation.<br>*In McCarthy (ed.): Intermediate punishments: intensive supervision, home confinement and electronic supervision,* 113-132. | Treatment and control conditions not sufficiently different in terms of contact/caseload size changes. More a test of a case planning strategy. |
| **Bonta, Wallace-Capretta, & Rooney (2000)**<br>A quasi-experimental evaluation of an intensive rehabilitation supervision program.<br>*Crim. Justice & Behavior, 27,* 312-329. | Primarily a test of electronic monitoring. |
| **Boudouris & Turnbull (1985)**<br>Shock probation in Iowa.<br>*J. Off. Counseling Services & Rehab., 9(4),* 53-67. | Unmatched controls consisting of mixed corrections population (prisoners, parolees, probationers etc.) |
| **Brownlee & Joanes (1993)**<br>Intensive probation for young adult offenders: evaluating the impact of a non-custodial sentence.<br>*Brit. J. Criminol., 33*, 216-230. | No comparison group. |
| **Burkhart (1969)**<br>The parole work unit programme: an evaluation report.<br>*Brit. J. Criminol., 9,* 125-147. | Unmatched controls. |
| **Clear & Shapiro (1986)**<br>Identifying high risk probationers for supervision in the community: the Oregon model.<br>*Fed. Prob., 50,* 134-141. | Not a test of probation supervision intensity (tests the validity of a classification model). |
| **Cochran, Corbett, & Byrne (1986)**<br>Intensive probation supervision in Massachusetts: a case study in change.<br>*Fed. Prob., 50,* 124-133. | Not a matched design at the subject level. |
| **Cox, Bantley, & Roscoe (2005)**<br>Evaluation of the Court Support Services Division's Probation Transition Program and Technical Violation Unit.<br>*Central CT State Univ.* | Unmatched controls. |
| **Dawson & Cuppleditch (2007)**<br>An impact assessment of the Prolific and other Priority Offender programme.<br>*Home Office Online Report 08/07.* | Insufficient data to calculate effect size. |

| Study Reference | Reason for Exclusion |
|---|---|
| **Deschenes, Turner, & Petersilia (1995)**<br>Minnesota ICS program (ICR is eligible)<br>See reference in Bibliography. | Incarcerated controls. |
| **Diskind & Klonsky (1964)**<br>A second look at the New York State parole drug experiment.<br>*Fed. Prob., 28,* 34-40. | Not a test of probation supervision intensity.<br>No comparison group. |
| **Drake & Barnoski (2006)**<br>The effects of parole on recidivism: juvenile offenders released from Washington State institutions: final report.<br>*Washington State Inst. for Public Policy* | Control group did not receive regular probation/parole supervision (unsupervised release). |
| **England (1955)**<br>A study of postprobation recidivism among five hundred federal offenders.<br>*Fed. Prob., 19,* 10-15. | Not a test of probation supervision intensity.<br>No comparison group. |
| **English, Chadwick, & Pullen (1994)**<br>Colorado's intensive supervision probation: report of findings.<br>*CO Dept of Public Safety.*<br><br>**English, Pullen, & Colling-Chadwick (1996)**<br>Comparison of intensive supervision probation and community corrections clientele.<br>*CO Dept of Public Safety.* | Unmatched controls.<br>Control group did not receive regular probation supervision (halfway house). |
| **Erwin (1987)**<br>Evaluation of intensive probation supervision in Georgia.<br>*Georgia Dept. of Corrections.* | Unmatched controls. |
| **Feinberg (1991)**<br>Juvenile intensive supervision: a longitudinal evaluation of program effectiveness.<br>*In Armstrong (ed.): Intensive interventions with high-risk youths: promising approaches in juvenile probation and parole*, 423-447. | Unmatched controls.<br>Compared programs did not differ substantially in intensity. |
| **GAO (1993)**<br>Intensive Probation Supervision: mixed effectiveness in controlling crime.<br>*U.S. General Accounting Office, Report to the Chairman, Subcommittee on Crime & Criminal Justice, Committee on the Judiciary, House of Representatives.* | Unmatched controls. |

| Study Reference | Reason for Exclusion |
|---|---|
| **Giblin (2002)**<br>Using police officers to enhance the supervision of juvenile probationers: an evaluation of the Anchorage CAN program.<br>*Crime & Delinquency, 48,* 116-137. | Intensive supervision component provided entirely by police officers rather than probation officers. |
| **Gilbert (1977)**<br>Alternate routes: a diversion project in the juvenile justice system.<br>*Evaluation Quarterly, 1(2),* 301-318. | No comparison group.<br>Pretrial program. |
| **Gray et al. (2005)**<br>Intensive Supervision and Surveillance Program: the final report.<br>*Youth Justice Board, U.K.* | Program was not consistently delivered by probation: not possible to distinguish outcomes. Program more of an addition to supervision. |
| **Green & Phillips (1990)**<br>An examination of an intensive probation for alcohol offenders: five-year follow-up.<br>*Int. J. Off. Ther. Comp. Criminol., 34,* 31-42. | No comparison group. |
| **Haas & Latessa (1995)**<br>Intensive supervision in a rural county: diversion and outcome.<br>*In Smykla & Selke (eds.): Intermediate sanctions: sentencing in the 1990s*, 153-169. | Unmatched controls. |
| **Haghighi (1999)**<br>A survey of juvenile intensive supervision probation (ISP) programs in Texas.<br>*TX Juvenile Probation Commission.*` | No evaluation data reported. |
| **Hanley (2002)**<br>Risk differentiation and intensive supervision: a meaningful union?<br>*Ph.D. Diss., Univ. of Cincinnati*<br><br>**Hanley (2006)**<br>Appropriate services: examining the case classification principle.<br>*J. Off. Rehab., 42,* 1-22. | Secondary data analysis of a randomized controlled trial, but random assignment is not maintained. |
| **Hanlon et al. (1998)**<br>The response of drug abuser parolees to a combination of treatment and intensive supervision.<br>*Prison J., 78,* 31-44. | No comparison group. |

| Study Reference | Reason for Exclusion |
|---|---|
| **Hanlon et al. (1999)**<br>The relative effects of three approaches to the parole supervision of narcotic addicts and cocaine abusers.<br>*Prison J., 79,* 163-181. | Primarily a test of a treatment program not delivered by probation/parole officers. |
| **Harrell, Adams, & Gouvis (1994)**<br>Evaluation of the impact of systemwide drug testing in Multnomah County, Oregon<br>*Urban Institute* | Unmatched controls.<br>No offender-level outcomes for post-trial probation experiment. |
| **Harrell et al. (2003)**<br>The impact evaluation of the Maryland Break the Cycle initiative.<br>*Urban Institute.* | Unmatched controls. |
| **Harris, Gingerich, & Whittaker (2004)**<br>The 'effectiveness' of differential supervision.<br>*Crime & Delinquency, 50,* 235-271. | Not a test of probation supervision intensity (more risk classification).<br>Unmatched controls. |
| **Irish (1990)**<br>Crime, criminal justice and probation: preliminary analysis of selected programs in the Criminal Division for 1989.<br>*Nassau Co. Probation Dept.* | Review of programs with unmatched comparisons. |
| **Jernigan & Kronick (1992)**<br>Intensive parole: the more you watch, the more you catch.<br>*J. Off. Rehab., 17(3/4)*, 65-76. | Unmatched controls. |
| **Johnson, Austin, & Davies (2003)**<br>Banking low-risk offenders: is it a good investment?<br>*Inst. on Crime, Justice & Corrections, George Washington Univ.* | Unmatched controls. |
| **Kurtz & Linnemann (2006)**<br>Improving probation through client strengths: evaluating strength based treatments for at risk youth.<br>*Western Criminol. Rev., 7,* 9-19. | Unmatched controls.<br>Some outcomes reported for treatment group only. |
| **Land, McCall & Williams (1990)**<br>Something that works in juvenile justice: an evaluation of the North Carolina court counselors' intensive protective supervision randomized experimental project 1987-1989.<br>*Evaluation Review 14,* 574-606. | Supervision provided by juvenile court to mostly non-criminal juvenile status offenders (runaways, truants etc.) |

| Study Reference | Reason for Exclusion |
|---|---|
| **Lasater et al. (n.d.)**<br>School-based probation intervention results with high-risk youth in Montana.<br>*Character Development Systems, LLC.* | Primarily a test of cognitive-behavioral therapy.<br>Unmatched controls. |
| **Latessa & Travis (1988)**<br>The effects of intensive supervision with alcoholic probationers.<br>*J. Off. Counseling, Services & Rehab., 12(2)*, 175-190. | Limited matching of controls. Higher risk and need offenders selected into treatment group. |
| **Latessa & Vito (1988)**<br>The effects of intensive supervision on shock probationers.<br>*J. Crim. Just., 16,* 319-330 | Not independent of larger study in Latessa (1987). |
| **Lattimore et al. (2005)**<br>Evaluation of the juvenile Breaking the Cycle program<br>*RTI International/NIJ* | Probation component is not a key part of the program. Increased supervision and drug testing are outcomes rather than part of the process. |
| **Martin & Inciardi (1997)**<br>Case management outcomes for drug-involved offenders.<br>*Prison J., 77,* 168-183. | Primarily a test of a treatment program not delivered by probation officers.<br>No crime outcomes. |
| **Maupin (1993)**<br>Risk classification systems and the provision of juvenile aftercare.<br>*Crime & Delinquency, 39,* 90-105. | No comparison group or crime outcomes. Investigates how supervision intensity differs by risk level. |
| **Meisel (2001)**<br>Relationships and juvenile offenders: the effects of Intensive Aftercare Supervision.<br>*Prison J., 81,* 206-245. | No crime outcomes. |
| **MI Dept of Corrections (2002)**<br>530 probationers use automated reporting kiosks (Jan. 17).<br>*MI Dept of Corrections Staff News Bulletin.* | No evaluation data reported. No indication that an evaluation was conducted. |
| **MN Office of the Legislative Auditor (2005)**<br>Community supervision of sex offenders: evaluation report. | No evaluation data or crime outcomes reported. |
| **Nath (1974)**<br>Intensive Supervision Program: final report.<br>*Florida Parole Commission.* | Insufficient data to calculate effect sizes: sample sizes not provided for recidivism outcomes. |
| **National Council on Crime & Delinquency (2001)**<br>Evaluation of the RYSE program: Alameda County Probation Department. | Pre-trial program examining placement into treatment and services. |

| Study Reference | Reason for Exclusion |
|---|---|
| **Otoyo (1983)**<br>A study of the relationship of increased supervisory contacts to recidivism.<br>*Ed.D. Diss., Pepperdine Univ.* | Not a test of probation supervision intensity (looked at type of contact rather than frequency).<br>No comparison group. |
| **Petersilia (1989)**<br>Probation and felony offenders.<br>*Fed. Prob., 49,* 4-9. | Not a test of probation supervision intensity. |
| **Petersilia & Turner (1990)**<br>Los Angeles Electronic Supervision (other programs, including Los Angeles non-EM, are eligible).<br>*See reference in Bibliography.* | Electronic monitoring is the only difference between treatment and control programs. |
| **Petersilia & Turner (1990)**<br>Diverting prisoners to intensive probation: results of an experiment in Oregon.<br>*RAND Corp.*<br><br>**Petersilia & Turner (1993)**<br>Evaluating Intensive Supervision Probation/Parole: results of a national experiment.<br>[Marion Co., OR & Milwaukee, WI only. Other studies eligible.]<br>*NIJ Research in Brief* | Incarcerated controls. |
| **Petersilia, Turner, & Deschenes (1992b)**<br>Macon, GA program (others are eligible)<br>*See reference in Bibliography.* | Electronic monitoring is the only difference between treatment and control programs. |
| **Reichel & Sudbrack (1994)**<br>Differences among eligibles: who gets an ISP sentence?<br>*Fed. Prob., 58,* 51-58. | Not a test of probation supervision intensity. |
| **Rengifo & Scott-Hayward (2008)**<br>Assessing the effectiveness of intermediate sanctions: Multnomah County, Oregon.<br>*Vera Inst. of Justice Report Summary.* | Not a test of probation supervision intensity. |
| **Rhyne & Hamblin (2008)**<br>What works with the DV offender? Services, sanctions and supervision.<br>*Multnomah Co. Dept of Community Justice, Portland, OR.* | No comparison group. |

| Study Reference | Reason for Exclusion |
|---|---|
| **Robertson (2000)**<br>Comparison of community-based models for youth offenders. Part 1: Program effectiveness and cost-effectiveness.<br>*Soc. Sci. Res. Cen., MS State Univ./NIDA* | Unmatched controls. |
| **Robertson & Blackburn (1984)**<br>An assessment of treatment effectiveness by case classification.<br>*Fed. Prob., 48,* 34-38. | Primarily a test of treatment conditions in probation sentences.<br>Unmatched controls. |
| **Romero & Williams (1983)**<br>Group psychotherapy and intensive probation supervision with sex offenders.<br>*Fed. Prob., 47,* 36-41. | Primarily a test of group psychotherapy. |
| **Rossman et al. (1999)**<br>Confronting relapse and recidivism: case management and aftercare services in the OPTS program.<br>*Urban Institute.* | Insufficient data to calculate effect size. |
| **Rubin & Dodge (2009)**<br>Probation in Maine: setting the baseline.<br>*Univ. of Southern ME Muskie School of Public Service/NIC* | Not a test of probation supervision intensity. |
| **Sawyer (1975)**<br>The effects of community probation unit services versus conventional probation services on recidivism by juvenile probationers.<br>*Ph.D. Diss., Brigham Young Univ.* | Not a test of probation supervision intensity. |
| **Seng et al. (2000)**<br>A comparison of evaluation findings on sex offender probation projects in six Illinois counties.<br>*IL Criminal Justice Information Auth.* | No systematic change in intensity.<br>No regular probation comparison group.<br>See also Stalans, Seng, & Yarnold (2002). |
| **Serin, Vuong, & Briggs (2003)**<br>Intensive supervision practices: a preliminary examination.<br>*Correctional Service of Canada.* | Unmatched controls/control group consists of ISP exclusions. |
| **Simon (2008)**<br>Effectiveness of the Probation and Parole Service Delivery Model (PPSDM) in reducing recidivism.<br>*M.A. Thesis, Univ. of Saskatchewan* | Not a test of probation supervision intensity. |

| Study Reference | Reason for Exclusion |
|---|---|
| **Smith (1984)**<br>Alabama prison option: supervised intensive restitution program.<br>*Fed. Prob., 48,* 32-35. | No comparison group.<br>No crime outcomes. |
| **Solomon, Kachnowski, & Bhati (2005)**<br>Does parole work? Analyzing the impact of postprison supervision on rearrest outcomes.<br>*Urban Institute.* | Control group did not receive regular probation supervision (compares parolees to prisoners released without parole). |
| **Stalans, Seng, & Yarnold (2002)**<br>Long-term impact evaluation of specialized sex offender probation programs in Lake, DuPage, & Winnebago Counties.<br>*IL Criminal Justice Information Auth.*<br>**Stalans et al. (n.d.)**<br>Process and initial impact evaluation of the Cook County Adult Probation Department's sex offender program.<br>*No publication details.* | Unmatched controls.<br>See also Seng et al. (2000) |
| **Taxman & Thanner (2006)**<br>Risk, need, and responsivity (RNR): it all depends.<br>*Crime & Delinquency, 52,* 28-51. | Not a test of probation supervision intensity: experiment tests seamless treatment delivery (sometimes via probation, but not always). |
| **Texas Adult Probation Commission (1988)**<br>Recidivism study on intensive supervision, specialized caseloads, and restitution centers for 1985-1987. | Not an evaluation. Outcome comparisons for different types of disposals. |
| **Travis County (n.d.)**<br>The probation experiment.<br>*No publication details.* | Limited information on experiment, but does not appear to be a test of probation supervision intensity.<br>No comparison group. |
| **Trotter (1993 & 1995)**<br>The supervision of offenders – what works?<br>*Australian Criminol. Res. Council.* | Not a test of probation supervision intensity (evaluates a probation officer training program). |
| **Trusty & Arrigona (2001)**<br>Project Spotlight: first year implementation overview and recommendations for improvement.<br>*TX Criminal Justice Policy Council.*<br>**Trusty & Arrigona (2001)**<br>Project Spotlight: program overview, early implementation issues and outcome measures.<br>*TX Criminal Justice Policy Council.* | No comparison group. |

| Study Reference | Reason for Exclusion |
|---|---|
| **Turner & Greene (1999)**<br>The FARE probation experiment: implementation and outcomes of day fines for felony offenders in Maricopa County.<br>*Justice Sys. J., 21,* 1-21. | Not a test of probation supervision intensity (tests fine payment condition with no probation supervision). |
| **Turner & Jannetta (2007)**<br>Implementation and early outcomes for the San Diego High Risk Sex Offender (HRSO) GPS pilot program<br>*Cen. for Evidence-Based Corrections, UC Irvine.* | Unmatched controls (subjects who score highly on an instrument are assigned to treatment). |
| **Turner, Petersilia, Deschenes (1992)**<br>Evaluating Intensive Supervision Probation/Parole (ISP) for drug offenders.<br>*Crime & Delinquency, 38,* 539-556. | Insufficient data to calculate effect sizes.<br>See also Petersilia & Turner (1993). |
| **Turner et al. (2002)**<br>Evaluation of the South Oxnard Challenge Project 1997-2001.<br>*RAND Corporation.* | Supervision for the treatment group not always provided by probation officers. |
| **Ulmer & van Asten (2004)**<br>Restrictive Intermediate Punishments and recidivism in Pennsylvania.<br>*Fed. Sent. Rep., 16,* 182-187. | Unmatched controls. |
| **Van Vleet et al. (2002)**<br>Evaluation of Utah's Early Intervention Mandate: the juvenile sentencing guidelines and intermediate sanctions.<br>*National Institute of Justice.* | Not a test of probation supervision intensity. |
| **Virginia Department of Corrections (1988)**<br>Intensive Supervision Program (ISP) final evaluation report: client characteristics and supervision outcomes: a caseload comparison. | Unmatched incarcerated or probation violator controls. |
| **Vito (1986)**<br>Felony probation and recidivism: replication and response.<br>*Fed. Prob., 50,* 17-25. | Not a test of probation supervision intensity. |
| **Wagner & Baird (1993)**<br>Evaluation of the Florida Community Control Program.<br>*National Institute of Justice.* | Incarcerated controls. |

| Study Reference | Reason for Exclusion |
|---|---|
| **Weatherburn & Trimboli (2008)**<br>Community supervision and rehabilitation: two studies of offenders on supervised bonds.<br>*New South Wales Bureau of Crime Statistics & Research: Crime & Justice Bulletin.* | Not a test of probation supervision intensity. |
| **Wiebush (1993)**<br>Juvenile intensive supervision: the impact on felony offenders diverted from institutional placement.<br>*Crime & Delinquency, 39,* 68-89. | Unmatched controls.<br>Sample may overlap with another eligible study (Latessa & Vito, 1988). |
| **Wilson, Denton, & Williams (1987)**<br>Intensive Supervision Program evaluation: year two.<br>*Kentucky Corrections Cabinet.*<br><br>**Wilson (1987)**<br>Intensive supervision in Kentucky: program procedures and evaluation.<br>*National Institute of Justice.* | No comparison group. |
| **Wilson, Naro, & Austin (2007)**<br>Innovations in probation: assessing New York City's automated reporting system.<br>*JFA Institute.* | No comparison group. |
| **Wisconsin Dept of Health & Social Services (1989)**<br>Reducing criminal risk: an evaluation of the high risk offender intensive supervision project. | Unmatched controls. |
| **Wodahl (2007)**<br>The efficacy of graduated sanctions in reducing technical violations among probationers and parolees: an evaluation of the Wyoming Department of Corrections' intensive supervision program.<br>*Ph.D. Diss., University of NE at Omaha* | Not a test of probation supervision intensity.<br>No comparison group. |
| **Worrall et al. (2003)**<br>Intensive supervision and monitoring projects.<br>*Home Office Online Report 42/03.* | No evaluation data reported. |
| **Worrall et al. (2004)**<br>An analysis of the relationship between probation caseloads and property crime rates in California counties.<br>*J. Crim. Justice 32,* 231-241. | No comparison group (macro-level analysis of the relationship of statewide crime rates to statewide natural increases in caseload size). |

## Appendix E: Philadelphia APPD Low-Intensity Supervision Protocol

- Office reporting: Scheduled office visit once every six months. Contact focused on probation officer review of residence, employment, payments on fines/costs/restitution, and compliance with other conditions.
- Telephone reporting: Scheduled telephone report every six months, occurring approximately midway between office visits. Contact focused on confirmation of details described above. Clients not restricted from initiating telephone contact.
- Drug testing: Only administered if required by court order. Probation officer will order a FIR evaluation after no more than three positive urine tests, and is free to refer offender to drug treatment if the offender requests it.
- Missed contacts: Arrest warrants issued if no case contact for six months. If the offender surrenders positively, the warrant may be removed with no criminal sanction.

(Adapted from Ahlman & Kurtz, 2008).

## Appendix F: Logistic, Zero-Inflated Negative Binomial, and Two-Stage Least Squares Regression Models Without Jail Time Controls

Prevalence of Recidivism (All Offenses, 2-Year Follow Up, Original Table 2.2)

| Logistic Regression | | | Number of observations = 1,559 | |
|---|---|---|---|---|
| | | | Likelihood Ratio $\chi^2$ (9 d.f.) = 47.14 | |
| | | | Pr > $\chi^2$ = .000 | |
| Log likelihood = -787.658 | | | Pseudo $R^2$ = .029 | |
| **Term** | **Odds Ratio** | **S.E.** | **z** | ***p*** |
| Treatment group | 1.03 | .129 | .20 | .844 |
| West probation region | .73 | .109 | -2.12 | .034 |
| Male | 1.35 | .186 | 2.20 | .028 |
| White | 1.06 | .167 | .34 | .733 |
| Age at RA | .98 | .006 | -3.88 | .000 |
| Income $20,000-$29,999 | .72 | .164 | -1.43 | .153 |
| Income $30,000-$39,999 | .57 | .138 | -2.32 | .020 |
| Income $40,000 or more | .29 | .081 | -4.41 | .000 |
| Monthly offending rate 1 year pre-RA (any charged offense) | .79 | .138 | -1.36 | .174 |

Hosmer-Lemeshow $\chi^2$ for goodness-of-fit (8 d.f., 10 groups) = 5.53, $p \leq .699$

Prevalence of Recidivism (Violent Offenses, 2-Year Follow Up, Original Table 2.7)

| Logistic Regression | | | Number of observations = 1,559 | |
|---|---|---|---|---|
| | | | Likelihood Ratio $\chi^2$ (9 d.f.) = 37.48 | |
| | | | Pr > $\chi^2$ = .000 | |
| Log likelihood = -287.943 | | | Pseudo $R^2$ = .061 | |
| **Term** | **Odds Ratio** | **S.E.** | **z** | ***p*** |
| Treatment group | .86 | .205 | -.63 | .527 |
| West probation region | .68 | .193 | -1.36 | .173 |
| Male | 3.82 | 1.323 | 3.87 | .000 |
| White | 1.03 | .305 | .12 | .905 |
| Age at RA | .97 | .012 | -2.89 | .004 |
| Income $20,000-$29,999 | .63 | .253 | -1.14 | .254 |
| Income $30,000-$39,999 | .41 | .179 | -2.04 | .042 |
| Income $40,000 or more | .21 | .113 | -2.92 | .004 |
| Monthly offending rate 1 year pre-RA (charged violent off.) | 2.03 | .921 | 1.55 | .121 |

Hosmer-Lemeshow $\chi^2$ for goodness-of-fit (8 d.f., 10 groups) = 9.63, $p \leq .292$

Prevalence of Recidivism (Drug Offenses, 2-Year Follow Up, Original Table 2.13)

| Logistic Regression | | | Number of observations = 1,559 | |
|---|---|---|---|---|
| | | | Likelihood Ratio $\chi^2$ (9 d.f.) = 34.79 | |
| | | | Pr > $\chi^2$ = .000 | |
| Log likelihood = -471.825 | | | Pseudo $R^2$ = .036 | |
| **Term** | **Odds Ratio** | **S.E.** | **z** | ***p*** |
| Treatment group | .89 | .157 | -.65 | .517 |
| West probation region | .98 | .207 | -.08 | .939 |
| Male | 2.08 | .437 | 3.48 | .001 |
| White | 1.22 | .270 | .92 | .359 |
| Age at RA | .97 | .009 | -3.85 | .000 |
| Income $20,000-$29,999 | .62 | .187 | -1.59 | .113 |
| Income $30,000-$39,999 | .49 | .160 | -2.19 | .029 |
| Income $40,000 or more | .29 | .111 | -3.24 | .001 |
| Monthly offending rate 1 year pre-RA (charged drug off.) | .96 | .503 | -.08 | .937 |

Hosmer-Lemeshow $\chi^2$ for goodness-of-fit (8 d.f., 10 groups) = 14.46, $p \leq .071$

Frequency of Recidivism (All Offenses, 2-Year Follow-Up, Original Table 2.3)

| Zero-Inflated Negative Binomial Regression<br><br>Inflation model = logit<br>Log likelihood= -1692.398 | Number of observations = 1,559<br>Nonzero observations = 335<br>Zero observations = 1,224<br>Likelihood Ratio $\chi^2$ (9 d.f.) = 27.43<br>Pr > $\chi^2$ = .001 | | | |
|---|---|---|---|---|
| **Full Model** | **Incidence Rate Ratio** | **S.E.** | **z** | ***p*** |
| Treatment group | .93 | .142 | -.44 | .658 |
| West probation region | 1.17 | .220 | .83 | .408 |
| Male | 1.98 | .330 | 4.11 | .000 |
| White | 1.13 | .203 | .68 | .495 |
| Age at RA | .98 | .008 | -2.12 | .034 |
| Income $20,000-$29,999 | 1.53 | .425 | 1.55 | .122 |
| Income $30,000-$39,999 | 1.38 | .432 | 1.02 | .310 |
| Income $40,000 or more | 1.10 | .391 | .27 | .789 |
| Monthly offending rate 1 year pre-RA (any charged offense) | 1.24 | .249 | 1.09 | .277 |
| **Inflated Model** | **b** | **S.E.** | **z** | ***p*** |
| Treatment group | -.060 | .154 | -.39 | .696 |
| West probation region | .432 | .189 | 2.28 | .022 |
| Male | -.097 | .176 | -.55 | .582 |
| White | -.018 | .193 | .09 | .925 |
| Age at RA | .023 | .008 | 2.91 | .004 |
| Income $20,000-$29,999 | .612 | .332 | 1.84 | .065 |
| Income $30,000-$39,999 | .858 | .356 | 2.41 | .016 |
| Income $40,000 or more | 1.554 | .397 | 3.91 | .000 |
| Monthly offending rate 1 year pre-RA (any charged offense) | .340 | .195 | 1.59 | .112 |
| Constant | -1.080 | .561 | -1.93 | .054 |
| Ln(Alpha) | .588 | .207 | 2.84 | .005 |
| Alpha | 1.800 | .373 | | |

Vuong test of zero-inflated vs. standard negative binomial: z = 4.03, $p < .001$

Frequency of Recidivism (Violent Offenses, 2-Year Follow-Up, Original Table 2.8)

| Zero-Inflated Negative Binomial Regression | | | Number of observations = 1,559 | |
|---|---|---|---|---|
| | | | Nonzero observations = 77 | |
| | | | Zero observations = 1,482 | |
| Inflation model = logit | | | Likelihood Ratio $\chi^2$ (9 d.f.) = 5.20 | |
| Log likelihood = -509.488 | | | Pr > $\chi^2$ = .817 | |
| **Full Model** | **Incidence Rate Ratio** | **S.E.** | **z** | ***p*** |
| Treatment group | .93 | .218 | -.32 | .748 |
| West probation region | .95 | .288 | -.18 | .858 |
| Male | 1.05 | .373 | .13 | .894 |
| White | 1.03 | .326 | .09 | .928 |
| Age at RA | .99 | .015 | -.75 | .452 |
| Income $20,000-$29,999 | 1.64 | .725 | 1.12 | .263 |
| Income $30,000-$39,999 | 1.18 | .681 | .28 | .779 |
| Income $40,000 or more | 1.02 | .642 | .04 | .969 |
| Monthly offending rate 1 year pre-RA (charged violent off.) | .48 | .263 | -1.34 | .181 |
| **Inflated Model** | **b** | **S.E.** | **z** | ***p*** |
| Treatment group | .133 | .243 | .55 | .584 |
| West probation region | .396 | .292 | 1.36 | .175 |
| Male | -1.340 | .351 | -3.82 | .000 |
| White | -.034 | .300 | -.11 | .909 |
| Age at RA | .034 | .013 | 2.69 | .007 |
| Income $20,000-$29,999 | .517 | .418 | 1.24 | .216 |
| Income $30,000-$39,999 | .918 | .461 | 1.99 | .047 |
| Income $40,000 or more | 1.533 | .553 | 2.77 | .006 |
| Monthly offending rate 1 year pre-RA (charged violent off.) | -1.198 | .871 | -1.38 | .169 |
| Constant | 1.577 | .734 | 2.15 | .032 |
| Ln(Alpha) | -.216 | .283 | -.76 | .445 |
| Alpha | .806 | .228 | | |

Vuong test of zero-inflated vs. standard negative binomial: z = 3.95, $p < .001$.

Frequency of Recidivism (Drug Offenses, 2-Year Follow-Up, Original Table 2.14)

| Zero-Inflated Negative Binomial Regression | | | Number of observations = 1,559 | |
|---|---|---|---|---|
| | | | Nonzero observations = 148 | |
| | | | Zero observations = 1,411 | |
| Inflation model = logit | | | Likelihood Ratio $\chi^2$ (9 d.f.) = 14.38 | |
| Log likelihood = -737.786 | | | Pr > $\chi^2$ = .109 | |
| **Full Model** | **Incidence Rate Ratio** | **S.E.** | **z** | ***p*** |
| Treatment group | .83 | .134 | -1.18 | .237 |
| West probation region | 1.20 | .231 | .97 | .334 |
| Male | 1.47 | .323 | 1.75 | .080 |
| White | .86 | .175 | -.76 | .447 |
| Age at RA | .98 | .008 | -1.97 | .049 |
| Income $20,000-$29,999 | .96 | .261 | -.16 | .876 |
| Income $30,000-$39,999 | .89 | .276 | -.36 | .717 |
| Income $40,000 or more | .59 | .223 | -1.38 | .166 |
| Monthly offending rate 1 year pre-RA (charged drug off.) | 1.81 | .902 | 1.19 | .234 |
| **Inflated Model** | **b** | **S.E.** | **z** | ***p*** |
| Treatment group | .049 | .190 | .26 | .797 |
| West probation region | .084 | .228 | .37 | .712 |
| Male | -.607 | .235 | -2.58 | .010 |
| White | -.268 | .241 | -1.11 | .266 |
| Age at RA | .030 | .010 | .307 | .002 |
| Income $20,000-$29,999 | .482 | .324 | 1.49 | .137 |
| Income $30,000-$39,999 | .699 | .354 | 1.98 | .048 |
| Income $40,000 or more | 1.064 | .425 | 2.51 | .012 |
| Monthly offending rate 1 year pre-RA (charged drug off.) | .168 | .544 | .03 | .757 |
| Constant | .662 | .557 | 1.19 | .235 |
| Ln(Alpha) | -.990 | .403 | -2.46 | .014 |
| Alpha | .372 | .150 | | |

Vuong test of zero-inflated vs. standard negative binomial: z = 3.57, p < .001

Instrumental Variables Model: Prevalence of Recidivism (All Offenses, 2-Year Follow-Up, Original Table 2.5)

| | **First Stage** Treatment Take-Up | **Reduced Form (ITT)** Post-RA Any Off. |
|---|---|---|
| | Observations = 1,559 | Observations = 1,559 |
| | $R^2$ = .655 | $R^2$ = .031 |
| | Adjusted $R^2$ = .651 | Adjusted $R^2$ = .020 |
| **Instruments** | **b (S.E.)** | **b (S.E.)** |
| Assigned LIS | .823 (.089)*** | -.092 (.123) |
| Assigned LIS*West | -.084 (.035)* | .027 (.049) |
| Assigned LIS*Male | .015 (.032) | -.009 (.044) |
| Assigned LIS*White | -.025 (.038) | .036 (.052) |
| Assigned LIS*Age | -.000 (.001) | -.000 (.002) |
| Assigned LIS*Income20 | .034 (.058) | .080 (.080) |
| Assigned LIS*Income30 | .087 (.061) | .091 (.084) |
| Assigned LIS*Income40 | .041 (.067) | .071 (.093) |
| Assigned LIS*Prior offending | -.034 (.027) | .048 (.038) |
| **Exogenous** | | |
| West probation region | .055 (.025)* | -.066 (.035) |
| Male | .004 (.023) | .052 (.032) |
| White | .007 (.027) | -.009 (.037) |
| Age at RA | .001 (.001) | -.004 (.001)** |
| Income $20,000-$29,999 | -.007 (.040) | -.094 (.055) |
| Income $30,000-$39,999 | -.013 (.042) | -.139 (.059)* |
| Income $40,000 or more | -.005 (.047) | -.232 (.065)*** |
| Monthly offending rate 1 year pre-RA (any offense) | -.004 (.022) | -.055 (.031) |
| Constant | -.052 (.064) | .509 (.089)*** |

* $p \leq .05$; ** $p \leq .01$; *** $p \leq .001$.

| Second-Stage Instrumental Variables Regression | | | Number of observations = 1,559 | |
|---|---|---|---|---|
| | | | Wald $\chi^2$ (9 d.f.) = 46.13 | |
| | | | Pr > $\chi^2$ = .000 | |
| Any New Charged Offense | | | $R^2$ = .028 | |
| **Term** | **b** | **S. E.** | **Z** | ***p*** |
| Predicted Treatment Take-up | .006 | .026 | .22 | .827 |
| Constant | .459 | .062 | 7.45 | .000 |

Controlling for region, gender, race, age, SES and offending history.

Instrumental Variables Model: Prevalence of Recidivism (Violent Offenses, 2-Year Follow-Up, Original Table 2.11)

| | **First Stage** Treatment Take-Up | **Reduced Form (ITT)** Post-RA Violent Off. |
|---|---|---|
| | Observations = 1,559 | Observations = 1,559 |
| | $R^2$ = .654 | $R^2$ = .028 |
| | Adjusted $R^2$ = .650 | Adjusted $R^2$ = .017 |
| **Instruments** | **b (S.E.)** | **b (S.E.)** |
| Assigned LIS | .818 (.089)*** | -.089 (.065) |
| Assigned LIS*West | -.083 (.035)* | .018 (.026) |
| Assigned LIS*Male | .014 (.032) | .007 (.023) |
| Assigned LIS*White | -.022 (.038) | .051 (.028) |
| Assigned LIS*Age | -.001 (.001) | -.000 (.001) |
| Assigned LIS*Income20 | .034 (.058) | .065 (.042) |
| Assigned LIS*Income30 | .085 (.061) | .058 (.045) |
| Assigned LIS*Income40 | .036 (.067) | .031 (.049) |
| Assigned LIS*Prior offending | -.011 (.100) | -.093 (.073) |
| **Exogenous** | | |
| West probation region | .055 (.025)* | -.027 (.018) |
| Male | .004 (.023) | .045 (.017)** |
| White | .007 (.027) | -.025 (.020) |
| Age at RA | .001 (.001) | -.002 (.001)* |
| Income $20,000-$29,999 | -.007 (.040) | -.053 (.029) |
| Income $30,000-$39,999 | -.014 (.042) | -.070 (.031)* |
| Income $40,000 or more | -.006 (.047) | -.079 (.034)* |
| Monthly offending rate 1 year pre-RA (violent offense) | -.055 (.083) | .111 (.061) |
| Constant | -.053 (.064) | .167 (.047)*** |

* $p \leq .05$; ** $p \leq .01$; *** $p \leq .001$.

| Second-Stage Instrumental Variables Regression | | | Number of observations = 1,559 | |
|---|---|---|---|---|
| | | | Wald $\chi^2$ (9 d.f.) = 35.44 | |
| | | | Pr > $\chi^2$ = .000 | |
| New Charged Violent Off. | | | $R^2$ = .023 | |
| **Term** | **b** | **S. E** | **z** | ***p*** |
| Predicted treatment take-up | -.008 | .014 | -.61 | .541 |
| Constant | .128 | .033 | 3.91 | .000 |

Controlling for region, gender, race, age, SES and offending history.

Instrumental Variables Model: Prevalence of Recidivism (Drug Offenses, 2-Year Follow-Up, Original Table 2.16)

| | **First Stage** Treatment Take-Up | **Reduced Form (ITT)** Post-RA Drug Off. |
|---|---|---|
| | Observations = 1,559 | Observations = 1,559 |
| | $R^2$ = .665 | $R^2$ = .023 |
| | Adjusted $R^2$ = .651 | Adjusted $R^2$ = .012 |
| **Instruments** | **b (S.E.)** | **b (S.E.)** |
| Assigned LIS | .812 (.089)*** | -.054 (.088) |
| Assigned LIS*West | -.086 (.035)* | .027 (.035) |
| Assigned LIS*Male | .012 (.032) | .034 (.032) |
| Assigned LIS*White | -.023 (.038) | .018 (.038) |
| Assigned LIS*Age | -.000 (.001) | .000 (.001) |
| Assigned LIS*Income20 | .032 (.058) | .004 (.057) |
| Assigned LIS*Income30 | .083 (.061) | -.010 (.060) |
| Assigned LIS*Income40 | .033 (.067) | -.013 (.066) |
| Assigned LIS*Prior offending | .200 (.093)* | .006 (.092) |
| **Exogenous** | | |
| West probation region | .055 (.025)* | -.016 (.025) |
| Male | .004 (.023) | .039 (.023) |
| White | .007 (.027) | .009 (.027) |
| Age at RA | .001 (.001) | -.003 (.001)** |
| Income $20,000-$29,999 | -.007 (.040) | -.046 (.040) |
| Income $30,000-$39,999 | -.013 (.042) | -.060 (.042) |
| Income $40,000 or more | -.005 (.047) | -.098 (.047)* |
| Monthly offending rate 1 year pre-RA (drug offense) | -.010 (.055) | .006 (.055) |
| Constant | -.052 (.064) | .254 (.064)*** |

| Second-Stage Instrumental Variables Regression | | | Number of observations = 1,559 | |
|---|---|---|---|---|
| | | | Wald $\chi^2$ (9 d.f.) = 34.39 | |
| | | | Pr > $\chi^2$ = .000 | |
| New Charged Drug Offense | | | $R^2$ = .102 | |
| **Term** | **b** | **S. E** | **z** | ***p*** |
| Predicted treatment take-up | -.013 | .018 | -.71 | .475 |
| Constant | .230 | .044 | 5.21 | .000 |

Controlling for region, gender, race, age, SES and offending history.

## Appendix G: Conditional Distributions of Selected Model Covariates and Outcome

Age at RA and Probability of a New Charge

| Age | Charged (%) | Not Charged (%) | Total (%) | Age | Charged (%) | Not Charged (%) | Total (%) |
|---|---|---|---|---|---|---|---|
| **19** | 1 (50.0) | 1 (50.0) | 2 (100) | **46** | 14 (24.1) | 44 (75.9) | 58 (100) |
| **20** | 0 (0.0) | 2 (100.0) | 2 (100) | **47** | 13 (17.3) | 62 (82.7) | 75 (100) |
| **21** | 4 (23.5) | 13 (76.5) | 17 (100) | **48** | 9 (17.0) | 44 (83.0) | 53 (100) |
| **22** | 8 (42.1) | 11 (57.9) | 19 (100) | **49** | 7 (14.9) | 40 (85.1) | 47 (100) |
| **23** | 8 (33.3) | 16 (66.7) | 24 (100) | **50** | 16 (25.8) | 46 (74.2) | 62 (100) |
| **24** | 5 (20.8) | 19 (79.2) | 24 (100) | **51** | 6 (17.6) | 28 (82.4) | 34 (100) |
| **25** | 10 (23.3) | 33 (76.7) | 43 (100) | **52** | 11 (26.2) | 31 (73.8) | 42 (100) |
| **26** | 6 (19.4) | 25 (80.6) | 31 (100) | **53** | 4 (12.5) | 28 (87.5) | 32 (100) |
| **27** | 20 (35.7) | 36 (64.3) | 56 (100) | **54** | 6 (15.0) | 34 (85.0) | 40 (100) |
| **28** | 8 (20.0) | 32 (80.0) | 40 (100) | **55** | 6 (22.2) | 21 (77.8) | 27 (100) |
| **29** | 13 (28.3) | 33 (71.7) | 46 (100) | **56** | 3 (20.0) | 12 (80.0) | 15 (100) |
| **30** | 9 (25.7) | 26 (74.3) | 35 (100) | **57** | 2 (12.5) | 14 (87.5) | 16 (100) |
| **31** | 8 (21.1) | 30 (78.9) | 38 (100) | **58** | 3 (21.4) | 11 (78.6) | 14 (100) |
| **32** | 9 (27.3) | 24 (72.7) | 33 (100) | **59** | 0 (0.0) | 15 (100.0) | 15 (100) |
| **33** | 10 (24.4) | 31 (75.6) | 41 (100) | **60** | 0 (0.0) | 7 (100.0) | 7 (100) |
| **34** | 8 (26.7) | 22 (73.3) | 30 (100) | **61** | 0 (0.0) | 7 (100.0) | 7 (100) |
| **35** | 7 (14.3) | 42 (85.7) | 49 (100) | **62** | 0 (0.0) | 5 (100.0) | 5 (100) |
| **36** | 8 (21.6) | 29 (78.4) | 37 (100) | **63** | 0 (0.0) | 7 (100.0) | 7 (100) |
| **37** | 10 (21.7) | 36 (78.3) | 46 (100) | **64** | 0 (0.0) | 3 (100.0) | 3 (100) |
| **38** | 9 (19.1) | 38 (80.9) | 47 (100) | **65** | 1 (25.0) | 3 (75.0) | 4 (100) |
| **39** | 13 (29.5) | 31 (70.5) | 44 (100) | **66** | 2 (50.0) | 2 (50.0) | 4 (100) |
| **40** | 12 (27.3) | 32 (72.7) | 44 (100) | **67** | 0 (0.0) | 5 (100.0) | 5 (100) |
| **41** | 9 (18.4) | 40 (81.6) | 49 (100) | **68** | 1 (50.0) | 1 (50.0) | 2 (100) |
| **42** | 10 (20.4) | 39 (79.6) | 49 (100) | **69** | 0 (0.0) | 2 (100.0) | 2 (100) |
| **43** | 6 (12.8) | 41 (87.2) | 47 (100) | **70** | 1 (50.0) | 1 (50.0) | 2 (100) |
| **44** | 5 (13.5) | 32 (86.5) | 37 (100) | **71** | 1 (100.0) | 0 (0.0) | 1 (100) |
| **45** | 13 (26.0) | 37 (74.0) | 50 (100) | **Total** | 335 (21.5) | 1,224 (78.5) | 1,559 (100) |

Post-RA Months in Jail and Probability of a New Charge

| Months in jail | Charged N (%) | Not Charged N (%) | Total N (%) |
|---|---|---|---|
| **0** | 210 (15.4) | 1,156 (84.6) | 1,366 (100) |
| **1** | 4 (44.4) | 5 (55.6) | 9 (100) |
| **2** | 2 (28.6) | 5 (71.4) | 7 (100) |
| **3** | 1 (33.3) | 2 (66.7) | 3 (100) |
| **4** | 5 (36.5) | 8 (61.5) | 13 (100) |
| **5** | 1 (20.0) | 4 (80.0) | 5 (100) |
| **6** | 4 (44.4) | 5 (55.6) | 9 (100) |
| **7** | 12 (75.0) | 4 (25.0) | 16 (100) |
| **8** | 7 (63.6) | 4 (36.4) | 11 (100) |
| **9** | 7 (70.0) | 3 (30.0) | 10 (100) |
| **10** | 14 (87.5) | 2 (12.5) | 16 (100) |
| **11** | 12 (70.6) | 5 (29.4) | 17 (100) |
| **12** | 56 (72.7) | 21 (27.3) | 77 (100) |
| **Total** | 335 (21.5) | 1,224 (78.5) | 1,559 (100) |

## Appendix H: Diagnostics for Proportional Hazards Models

| Scaled Schoenfeld Residuals<br>All Charged Offenses | | | Time = time<br>d.f. = 1 for all covariates |
|---|---|---|---|
| **Covariate** | **Rho** | $\chi^2$ | ***p*** |
| Treatment group | -.011 | .04 | .836 |
| West probation region | -.008 | .02 | .880 |
| Male | -.044 | .70 | .401 |
| White | -.043 | .71 | .400 |
| Age at RA | -.026 | .23 | .632 |
| Income $20,000-$29,999 | -.009 | .03 | .874 |
| Income $30,000-$39,999 | .025 | .21 | .646 |
| Income $40,000 or more | .015 | .07 | .789 |
| Monthly offending rate 1 year pre-RA (any charged offense) | .045 | .38 | .537 |
| In jail Oct 2007 | -.021 | .15 | .701 |
| In jail Nov 2007 | -.027 | .26 | .613 |
| In jail Dec 2007 | .105 | 3.70 | .054 |
| In jail Jan 2008 | -.119 | 4.75 | .029 |
| In jail Feb 2008 | -.010 | .03 | .861 |
| In jail Mar 2008 | -.019 | .13 | .721 |
| In jail Apr 2008 | .072 | 1.82 | .177 |
| In jail May 2008 | -.019 | .12 | .731 |
| In jail Jun 2008 | -.008 | .02 | .878 |
| In jail Jul 2008 | .081 | 2.26 | .132 |
| In jail Aug 2008 | -.050 | .89 | .345 |
| In jail Sep 2008 | .004 | .01 | .941 |

| Scaled Schoenfeld Residuals<br>Charged Violent Offenses | | Time = time<br>d.f. = 1 for all covariates | |
|---|---|---|---|
| **Covariate** | **Rho** | $\chi^2$ | ***p*** |
| Treatment group | .051 | .21 | .650 |
| West probation region | .201 | 3.01 | .083 |
| Male | -.031 | .08 | .782 |
| White | .196 | 3.46 | .063 |
| Age at RA | -.063 | .33 | .569 |
| Income $20,000-$29,999 | -.054 | .23 | .629 |
| Income $30,000-$39,999 | -.051 | .17 | .680 |
| Income $40,000 or more | -.219 | 3.75 | .053 |
| Monthly offending rate 1 year pre-RA (charged violent offense) | .139 | .81 | .367 |
| In jail 1 year post-RA | -.276 | 6.31 | .012 |

| Scaled Schoenfeld Residuals<br>Charged Drug Offenses | | Time = time<br>d.f. = 1 for all covariates | |
|---|---|---|---|
| **Covariate** | **Rho** | $\chi^2$ | ***p*** |
| Treatment group | -.032 | .17 | .679 |
| West probation region | .043 | .30 | .584 |
| Male | -.107 | 1.82 | .177 |
| White | -.024 | .10 | .756 |
| Age at RA | -.025 | .11 | .739 |
| Income $20,000-$29,999 | .032 | .16 | .687 |
| Income $30,000-$39,999 | .059 | .54 | .464 |
| Income $40,000 or more | .133 | 2.64 | .104 |
| Monthly offending rate 1 year pre-RA (charged drug offense) | .076 | .41 | .522 |
| In jail Oct 2007 | -.100 | 1.48 | .225 |
| In jail Nov 2007 | -.026 | .10 | .747 |
| In jail Dec 2007 | .185 | 5.09 | .024 |
| In jail Jan 2008 | -.169 | 4.26 | .039 |
| In jail Feb 2008 | .019 | .06 | .810 |
| In jail Mar 2008 | -.020 | .06 | .800 |
| In jail Apr 2008 | .050 | .38 | .537 |
| In jail May 2008 | -.002 | .00 | .981 |
| In jail Jun 2008 | .057 | .48 | .488 |
| In jail Jul 2008 | .036 | .19 | .666 |
| In jail Aug 2008 | -.113 | 1.99 | .158 |
| In jail Sep 2008 | .004 | .00 | .957 |

## Appendix J: Sensitivity, Specificity, and Predictive Value

| | Outcome | | |
|---|---|---|---|
| | **No Serious Offense** | **Serious Offense** | |
| **Predicted Low Risk** | True Positive (TP) | False Positive (FP) | *Positive Predictive Value (PPV)* |
| **Predicted Non-Low Risk** | False Negative (FN) | True Negative (TN) | *Negative Predictive Value (NPV)* |
| | *Sensitivity (Sn)* | *Specificity (Sp)* | |

Sensitivity

The proportion of true positives (predicted/actual low-risk) identified by the model.

$$Sn = \frac{\#TP}{\#TP + \#FN}$$

Specificity

The proportion of true negatives (predicted/actual non-low risk) identified by the model.

$$Sp = \frac{\#TN}{\#TN + \#FP}$$

Positive predictive value

The proportion of predicted low-risk cases that are actually low risk (no serious offense).

$$PPV = \frac{\#TP}{\#TP + \#FP}$$

Negative predictive value

The proportion of predicted non-low risk cases that are actually non-low risk (commit serious offense).

$$NPV = \frac{\#TN}{\#TN + \#FN}$$

# BIBLIOGRAPHY

## Chapter 1

** *Denotes the main report of a study that was eligible for inclusion in the meta-analysis.*

American Correctional Association. (2006, August 16). *Public correctional policy on probation*. Retrieved January 26, 2009, from http://www.aca.org/government/policyresolution/view.asp?ID=34

Andrews, D. A., Bonta, J., & Hoge, R. D. (1990). Classification for effective rehabilitation: rediscovering psychology. *Criminal Justice & Behavior, 17,* 19-52. doi:10.1177/0093854890017001004

Aos, S., Miller, M., & Drake, E. (2006). *Evidence-based public policy options to reduce future prison construction, criminal justice costs, and crime rates*. Olympia, WA: Washington State Institute for Public Policy. Retrieved from http://www.wsipp.wa.gov/rptfiles/06-10-1201.pdf

Barnes, G., Ahlman, L., Gill, C., Sherman, L. W., Kurtz, E., & Malvestuto, R. (Forthcoming). Low-intensity community supervision for low-risk offenders: A randomized, controlled trial. *Journal of Experimental Criminology*, in press.

** Barnoski, R. (2003). *Evaluation of Washington State's 1996 juvenile court program for high-risk, first-time offenders: Final report.* Olympia, WA: Washington State Institute for Public Policy. Retrieved from http://www.wsipp.wa.gov/rptfiles/EIPfinal.pdf

Bennett, L. A. (1988). Practice in search of a theory: The case of intensive supervision – an extension of an old practice or a new approach? *American Journal of Criminal Justice, 12,* 293-310. doi:10.1007/BF02888940

Berk, R., Barnes, G., Ahlman, L., & Kurtz, E. (forthcoming). When second best is good enough: A comparison between a true experiment and a regression discontinuity quasi-experiment. *Journal of Experimental Criminology*, in press.

Bonta, J., Rugge, T., Scott, T.-L., Bourgon, G., & Yessine, A. K. (2008). Exploring the black box of community supervision. *Journal of Offender Rehabilitation, 47,* 248-270. doi: 10.1080/10509670802134085

Clear, T. R., & Hardyman, P. L. (1990). The new intensive supervision movement. *Crime & Delinquency, 36,* 42-60. doi:10.1177/0011128790036001004

Cochrane Collaboration. (2008). Review Manager (RevMan). Version 5.0 [Computer software]. Copenhagen: The Nordic Cochrane Centre. http://www.cc-ims.net/revman/about-revman-5

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Boston, MA: Houghton Mifflin.

Crouch, B.M. (1993). Is incarceration really worse? Analysis of offenders' preferences for prison over probation. *Justice Quarterly, 10,* 67-88. Retrieved from http://heinonline.org/HOL/Page?handle=hein.journals/jquart10&div=14&g_sent=1&collection=journals#77

** Deschenes, E. P., Turner, S., & Petersilia, J. (1995). *Intensive community supervision in Minnesota: A dual experiment in prison diversion and enhanced supervised release.* Santa Monica, CA: RAND Corporation. Retrieved from http://www.rand.org/pubs/drafts/2008/DRU777-1.pdf

Erwin, B. S. (1986). Turning up the heat on probationers in Georgia. *Federal Probation, 50(2),* 17-24. Retrieved from http://www.heinonline.org/HOL/Page?handle=hein.journals/fedpro50&id=1&size=2&collection=journals&index=journals/fedpro

** Fagan, J. A., & Reinarman, C. (1986). *Intensive supervision for violent offenders – the transition from adolescence to early adulthood. A longitudinal evaluation.* NCJ 106313. San Francisco, CA: Urban and Rural Systems Associates. See http://www.ncjrs.gov/App/publications/abstract.aspx?ID=106313

Farrington, D. P., Gottfredson, D. C., Sherman, L. W., & Welsh, B. C. (2006). The Maryland Scientific Methods Scale. In L. W. Sherman, D. P. Farrington, B. C. Welsh, & D. L. MacKenzie (Eds.), *Evidence-based crime prevention* (revised edition, pp. 13-21). New York, NY: Routledge.

** Folkard, M. S., Smith, D. E., & Smith, D. D. (1976). *IMPACT Intensive Matched Probation and After-Care Treatment. Volume II: The results of the experiment.* Home Office Research Study No. 36. London: HMSO. Retrieved from http://uk.sitestat.com/homeoffice/rds/s?rds.hors36pdf&ns_type=pdf&ns_url=%5Bhttp://www.homeoffice.gov.uk/rds/pdfs05/hors36.pdf%5D

Gendreau, P., Goggin, C., & Fulton, B. (2001). Intensive supervision in probation and parole settings. In C. R. Hollin (Ed.), *Handbook of offender assessment and treatment* (pp. 195-204). Chichester, UK: Wiley.

Giblin, M. J. (2002). Using police officers to enhance the supervision of juvenile probationers: An evaluation of the Anchorage CAN program. *Crime & Delinquency, 48,* 116-137. doi:10.1177/0011128702048001005

Glaser, D. (1983). Supervising offenders outside of prison. In J. Q. Wilson (Ed.), *Crime and public policy*. (pp. 207-228) San Francisco, CA: Institute for Contemporary Studies.

Glaze, L. E, & Bonczar, T. P. (2009). *Probation and parole in the United States, 2008*. Bureau of Justice Statistics Bulletin, December 2009. NCJ 228230. Washington, DC: U.S. Department of Justice Office of Justice Programs. Retrieved from http://bjs.ojp.usdoj.gov/content/pub/pdf/ppus08.pdf

** Guydish , J., Chan, M., Bostrom, A., Jessup, M. A., Davis, T. B., & Marsh, C. (2008). A randomized trial of Probation Case Management for drug-involved women offenders. *Crime & Delinquency (OnlineFirst)*. doi: 10.1177/0011128708318944

** Haapanen, R., & Britton, L. (2002). Drug testing for youthful offenders on parole: An experimental evaluation. *Criminology & Public Policy, 1,* 217-244. 10.1111/j. 1745-9133.2002.tb00087.x

Hanley, D. (2006). Appropriate services: Examining the case classification principle. *Journal of Offender Rehabilitation, 42(4),* 1-22. doi:10.1300/J076v42n04_01

** Hawken, A., & Kleiman, M. (2009). *Managing drug involved probationers with swift and certain sanctions: Evaluating Hawaii's HOPE.* NCJ 229023. Retrieved from http://www.ncjrs.gov/pdffiles1/nij/grants/229023.pdf

** Howard, L., Misch, G., Burke, C., & Pennell, S. (2002). *San Diego County Probation Department's Repeat Offender Prevention Program: Final evaluation report.* San Diego, CA: SANDAG. Retrieved from http://sandiegohealth.org/crime/

publicationid_753_1432.pdf

Johnson, K. D., Austin, J., & Davies, G. (2003). *Banking low-risk offenders: Is it a good investment?* NCJ 201304. Washington, D.C.: Institute on Crime, Justice, and Corrections, George Washington University. Retrieved from http://www.ncjrs.gov/pdffiles1/nij/grants/201304.pdf

** Latessa, E. J., Travis, L., Fulton, B., & Stichman, A. (1998). *Evaluating the prototypical ISP: Final report.* Cincinnati, Ohio: University of Cincinnati & American Probation and Parole Association. Retrieved from http://www.uc.edu/ccjr/Reports/ProjectReports/ISP.pdf

Lemert, E. M. (1993). Visions of social control: probation considered. *Crime & Delinquency*, *39*, 447-461. doi:10.1177/0011128793039004003

Lipsey, M. W., & Wilson, D. B. (2001). *Practical Meta-Analysis*. Thousand Oaks, CA: Sage.

Lowenkamp, C. T., Latessa, E. J., & Holsinger, A. M. (2006). The risk principle in action: what have we learned from 13,676 offenders and 97 correctional programs? *Crime & Delinquency, 52(1),* 77-93. doi:10.1177/0011128705281747

MacKenzie, D. L. (2006a). *What works in corrections*. New York, NY: Cambridge University Press.

MacKenzie, D. L. (2006b). Reducing the criminal activities of known offenders and delinquents: crime prevention in the courts and corrections. In L. W. Sherman, D. P. Farrington, B. C. Welsh, & D. L. MacKenzie (Eds.), *Evidence-based crime prevention* (revised edition, pp. 330-404). New York, NY: Routledge.

MacKenzie, D. L., & Brame, R. (2001). Community supervision, prosocial activities, and recidivism. *Justice Quarterly, 18,* 429-448. doi:10.1080/07418820100094971

Marlowe, D. B. (2003). Integrating substance abuse treatment and criminal justice supervision. *Science and Practice Perspectives, 2,* 4-14. Retrieved from http://www.ndci.org/sites/default/files/nadcp/NIDAPerspectives-Marlowe%5B1%5D.pdf

** Orchowsky, S., Merritt, N., & Browning, K. (1994). *Evaluation of the Virginia Department of Corrections' Intensive Supervision Program.* NCJ 153677. Richmond, VA: Virginia Department of Criminal Justice Services. See http://www.ncjrs.gov/App/Publications/abstract.aspx?ID=153677

** Paparozzi, M. A., & Gendreau, P. (2005). An intensive supervision program that worked: Service delivery, professional orientation, and organizational supportiveness. *The Prison Journal, 85,* 445-466. doi:10.1177/0032885505281529

** Pearson, F. S. (1988). Evaluation of New Jersey's Intensive Supervision Program. *Crime & Delinquency, 34,* 437-448. doi:10.1177/0011128788034004005

Petersilia, J. (1997). Probation in the United States. *Crime & Justice, 22,* 149-200. Retrieved from http://www.jstor.org/stable/1147573

** Petersilia, J., & Turner, S. (1990). *Intensive supervision for high-risk probationers: Findings from three California experiments*. Santa Monica, CA: The Rand Corporation. Retrieved from http://www.rand.org/pubs/reports/2007/R3936.pdf

Petersilia, J., & Turner, S. (1993). Intensive probation and parole. *Crime & Justice, 17,* 281-335. Retrieved from http://www.jstor.org/stable/1147553

Petersilia, J., & Deschenes, E. P. (1994). Perceptions of punishment: Inmates and staff rank the severity of prison versus intermediate sanctions. *The Prison Journal, 74,* 306-328. doi:10.1177/0032855594074003003

Petersilia, J., Turner, S., & Deschenes, E. P. (1992a). The costs and effects of intensive supervision for drug offenders. *Federal Probation, 56(4)*, 12-17. Retrieved from http://www.heinonline.org/HOL/Page?handle=hein.journals/fedpro56&id=1&size=2&collection=journals&index=journals/fedpro

** Petersilia, J., Turner, S., & Deschenes, E. P. (1992b). Intensive supervision programs for drug offenders. In J. M. Byrne, A. J. Lurigio, & J. Petersilia (Eds.), *Smart sentencing: The emergence of intermediate sanctions* (pp. 18-37). Newbury Park, CA: Sage.

Phipps, P., Korinek, K., Aos, S., & Lieb, R. (1999). *Research findings on adult corrections programs: A review.* Olympia, WA: Washington State Institute for Public Policy. Retrieved from http://wsipp.wa.gov/rptfiles/researchfindings.pdf

** Piquero, N. L. (2003). A recidivism analysis of Maryland's community probation program. *Journal of Criminal Justice, 31,* 295-307. doi:10.1016/S0047-2352(03)00024-2

** Reimer, E., & Warren, M. (1957). Relationship between violation rate and initially small caseload. *Crime & Delinquency, 3,* 222-229. doi:10.1177/001112875700300303

Renzema, M., & Mayo-Wilson, E. (2005). Can electronic monitoring reduce crime for moderate to high-risk offenders? *Journal of Experimental Criminology, 1,* 215-237. doi: 10.1007/s11292-005-1615-1

Rothstein, H. R., & Hopewell, S. (2009). Grey literature. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd edition, pp. 103-125). New York, NY: Russell Sage Foundation.

Ruth, H., & Reitz, K. R. (2003). *The challenge of crime: Rethinking our response*. Cambridge, MA: Harvard University Press.

Sherman, L. W., Gottfredson, D., MacKenzie, D., Eck, J., Reuter, P., & Bushway, S. (1997). *Preventing crime: What works, what doesn't, what's promising*. Washington, D.C.: United States Department of Justice, National Institute of Justice. Retrieved from http://www.ncjrs.gov/works

Skeem, J. L., & Manchak, S. (2008). Back to the future: From Klockars' model of effective supervision to evidence-based practice in probation. *Journal of Offender Rehabilitation, 47*, 220-247. doi:10.1080/10509670802134069

Solomon, A. L., Jannetta, J., Elderbroom, B., Winterfield, L., Osborne, J., Burke, P., Stroker, R. P., Rhine, E. E., & Burrell, W. D. (2008, December 2). *Putting public*

*safety first: 13 strategies for successful supervision and reentry*. Policy brief. Retrieved January 26, 2009, from Urban Institute website: http://www.urban.org/ /url.cfm?ID=411800

** Sontheimer, H., & Goodstein, L. (1993). An evaluation of juvenile intensive aftercare probation: aftercare versus system response effects. *Justice Quarterly, 10,* 197-227. Retrieved from http://www.heinonline.org/HOL/Page?handle=hein.journals/ jquart10&id=1&size=2&collection=journals&index=journals/jquart

Taxman, F. S. (2002). Supervision: Exploring the dimensions of effectiveness. *Federal Probation*, *66(2),* 14-27. Retrieved from http://www.uscourts.gov/fedprob/ 2002sepfp.pdf

Taxman, F. S. (2008a). To be or not to be: Community supervision déjà vu. *Journal of Offender Rehabilitation, 47,* 209-219. doi:10.1080/10509670802134036

Taxman, F. S. (2008b). No illusions: Offender and organizational change in Maryland's proactive community supervision efforts. *Criminology & Public Policy, 7,* 275-302. doi:10.1111/j.1745-9133.2008.00508.x

Taxman, F. S., & Thanner, M. (2006). Risk, need, and responsivity (RNR): It all depends. *Crime & Delinquency, 52(1),* 28-51. doi:10.1177/0011128705281754

** Taxman, F. S., Yancey, C., & Bilanin, J. E. (2006). *Proactive community supervision in Maryland: Changing offender outcomes.* Towson, MD: Maryland Department of Public Safety and Correctional Services. Retrieved from http://www.dpscs.state.md.us/publicinfo/publications/pdfs/PCS_Evaluation_Feb06.pdf

** Turner, S., & Petersilia, J. (1992). Focusing on high-risk parolees: An experiment to reduce commitments to the Texas Department of Corrections. *Journal of Research in Crime & Delinquency, 29,* 34-61. doi:10.1177/0022427892029001003

Weisburd, D., Sherman, L., & Petrosino, A. J. (1990). *Registry of randomized criminal justice experiments in sanctions*. NCJ 129725. Washington, DC: United States Department of Justice, National Institute of Justice. Retrieved from http://www.icpsr.umich.edu/NACJD/nij_pubs.html

** Wiebush, R. G., Wagner, D., McNulty, B., Wang, Y., & Le, T. N. (2005). *Implementation and outcome evaluation of the Intensive Aftercare Program: Final report.* NCJ 206177. Washington, DC: United States Department of Justice, Office of Juvenile Justice and Delinquency Prevention. Retrieved from http://www.ncjrs.org/pdffiles1/ojjdp/206177.pdf

Wilson, D. B. (2002). Meta-analysis macros for SAS, SPSS, and Stata. Retrieved December 7, 2009, from http://mason.gmu.edu/~dwilsonb/ma.html

Wilson, D. B. (2010). Meta-analysis. In A. R. Piquero & D. Weisburd (Eds.), *Handbook of quantitative criminology* (pp. 181-208). New York, NY: Springer.

Wilson, J. A., Naro, W., & Austin, J. F. (2007). *Innovations in probation: Assessing New York City's automated reporting system.* Washington, D.C.: JFA Associates. Retrieved from http://www.nyc.gov/html/prob/downloads/pdf/kiosk_report_2007.pdf

Worrall, J.L., Schram, P., Hays, E., & Newman, M. (2004). An analysis of the relationship between probation caseloads and property crime rates in California counties. *Journal of Criminal Justice*, *32,* 231-241. doi:10.1016/j.jcrimjus.2004.02.003

** Zhang, S. X., & Zhang, L. (2005). An experimental study of the Los Angeles County Repeat Offender Prevention Program: Its implementation and evaluation. *Criminology & Public Policy, 4,* 205-236. doi:10.1111/j.1745-9133.2005.00017.x

## Chapter 2

Agnew, R. (1991). The interactive effects of peer variables on delinquency. *Criminology, 29,* 41-72. doi:10.1111/j.1745-9125.1991.tb01058.x

Akers, R. (1973). *Deviant behavior: a social learning approach.* Belmont, CA: Wadsworth.

Ahlman, L. C., & Kurtz, E. M. (2008). *The APPD randomized controlled trial in low risk supervision.* Internal report. Philadelphia, PA: First Judicial District of Pennsylvania Adult Probation and Parole Department.

Ahlman, L., Kurtz, E., & Manning A. (2008). *Weapons related injury surveillance system (WRISS) report 2002-2007.* Internal report. Philadelphia, PA: First Judicial District of Pennsylvania Adult Probation and Parole Department.

Allison, P. D. (1984). *Event history analysis: Regression for longitudinal event data.* Thousand Oaks, CA: Sage Publications.

Andrews, D. A. (1989). Recidivism is predictable and can be influenced: Using risk assessments to reduce recidivism. *Forum on Corrections Research, 1(2),* 11-17. Retrieved from http://www.csc-scc.gc.ca/text/pblct/forum/special/a1e.pdf

Andrews, D. A., Bonta, J., & Hoge, R. D. (1990). Classification for effective rehabilitation: rediscovering psychology. *Criminal Justice & Behavior, 17,* 19-52. doi:10.1177/0093854890017001004

Angrist, J. D. (2006). Instrumental variables methods in experimental criminological research: What, why, and how. *Journal of Experimental Criminology, 2,* 23-44. doi:10.1007/s11292-005-5126-x.

Angrist, J. D., & Pischke, J-S. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton, NJ: Princeton University Press.

Aos, S., Miller, M., & Drake, E. (2006). *Evidence-based public policy options to reduce future prison construction, criminal justice costs, and crime rates*. Olympia, WA: Washington State Institute for Public Policy. Retrieved from http://www.wsipp.wa.gov/rptfiles/06-10-1201.pdf

Banks, J., Porter, A. L., Rardin, R. L., Siler, T. R., & Unger, V. E. (1976). *Intensive special probation projects – Phase I evaluation report.* NCJ 040512. Washington, D.C.: United States Department of Justice. Retrieved from http://www.ncjrs.gov/App/Publications/abstract.aspx?ID=40512

Barnes, G., Ahlman, L., Gill, C., Sherman, L. W., Kurtz, E., & Malvestuto, R. (Forthcoming). Low-intensity community supervision for low-risk offenders: A randomized, controlled trial. *Journal of Experimental Criminology*, in press.

Bennett, L. A. (1988). Practice in search of a theory: The case of intensive supervision – an extension of an old practice or a new approach? *American Journal of Criminal Justice, 12,* 293-310. doi:10.1007/BF02888940

Berk, R. A. (2005). Randomized experiments as the bronze standard. *Journal of Experimental Criminology, 1,* 417-433. doi:10.1007/s11292-005-3538-2

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 A, 172, Part 1*, 191-211. doi: 10.1111/j.1467-985X.2008.00556.x

Blumstein, A., Cohen, J., & Farrington, D. P. (1988). Criminal career research: Its value for criminology. *Criminology, 26,* 1-35. doi:10.1111/j.1745-9125.1988.tb00829.x

Carter, R., & Wilkins, L. T. (1976). Caseloads: Some conceptual models. In R. Carter, & L. Wilkins (Eds.), *Probation, parole, and community corrections* (2nd edition, pp. 391-401). New York, NY: Wiley.

Clear, T. R,. & Braga, A. A. (1995). Community corrections. In J. Q. Wilson, & J. Petersilia (Eds.), *Crime* (pp. 421-444). San Francisco, CA: Institute for Contemporary Studies.

Clear, T. R., & Hardyman, P. L. (1990). The new intensive supervision movement. *Crime & Delinquency, 36,* 42-60. doi:10.1177/0011128790036001004

Dishion, T. J., & Dodge, K. A. (2006). Deviant peer contagion in interventions and programs: An ecological framework for understanding influence mechanisms. In K. A. Dodge, T. J. Dishion, & J. E. Lansford (Eds.), *Deviant peer influences in programs for youth: Problems and solutions* (pp. 14-43). New York, NY: Guilford Press.

Dishion, T. J., McCord, J., & Poulin, F. (1999). When interventions harm: Peer groups and problem behavior. *American Psychologist, 54,* 755-764. doi:10.1037/0003-066X.54.9.755

Erwin, B. S. (1986). Turning up the heat on probationers in Georgia. *Federal Probation, 50(2),* 17-24. Retrieved from http://www.heinonline.org/HOL/Page?handle=hein.journals/fedpro50&id=1&size=2&collection=journals&index=journals/fedpro

Gambacorta, D. (2009, May 11). Can forecasting tool predict parolees who will commit a crime… and stop them? *Philadelphia Daily News,* p. 3. http://www.philly.com/dailynews/

Glaze, L. E, & Bonczar, T. P. (2009). *Probation and parole in the United States, 2008*. Bureau of Justice Statistics Bulletin, December 2009. NCJ 228230. Washington,

DC: U.S. Department of Justice Office of Justice Programs. Retrieved from http://bjs.ojp.usdoj.gov/content/pub/pdf/ppus08.pdf

Hanley, D. (2006). Appropriate services: Examining the case classification principle. *Journal of Offender Rehabilitation, 42(4),* 1-22. doi:10.1300/J076v42n04_01

Johnson, K. D., Austin, J., & Davies, G. (2003). *Banking low-risk offenders: Is it a good investment?* NCJ 201304. Washington, D.C.: Institute on Crime, Justice, and Corrections, George Washington University. Retrieved from http://www.ncjrs.gov/pdffiles1/nij/grants/201304.pdf

Latessa, E. J., Travis, L., Fulton, B., & Stichman, A. (1998). *Evaluating the prototypical ISP: Final report.* Cincinnati, Ohio: University of Cincinnati & American Probation and Parole Association. Retrieved from http://www.uc.edu/ccjr/Reports/ProjectReports/ISP.pdf

Lemert, E. M. (1993). Visions of social control: Probation considered. *Crime & Delinquency*, *39*, 447-461. doi:10.1177/0011128793039004003

Lowenkamp, C. T., & Latessa, E. J. (2004). Understanding the risk principle: How and why correctional interventions can harm low-risk offenders. *Topics in Community Corrections 2004* (pp. 3-8). Washington, D.C.: U.S. Department of Justice, National Institute of Corrections. Retrieved from http://www.nicic.org/pubs/2004/

period266.pdf

Lowenkamp, C. T., Latessa, E. J., & Holsinger, A. M. (2006). The risk principle in action: what have we learned from 13,676 offenders and 97 correctional programs? *Crime & Delinquency, 52(1),* 77-93. doi:10.1177/0011128705281747

MacKenzie, D. L. (2006). Reducing the criminal activities of known offenders and delinquents: crime prevention in the courts and corrections. In L. W. Sherman, D. P. Farrington, B. C. Welsh, & D. L. MacKenzie (Eds.), *Evidence-based crime prevention* (revised edition, pp. 330-404). New York, NY: Routledge.

Moher, D., Schulz, K. F., & Altman, D. (2001). The CONSORT statement: Revised recommendations for improving the quality of reports of parallel-group randomized trials. *Journal of the American Medical Association, 285,* 1987-1991. doi:10.1001/jama.285.15.1987

Montori, V. M., & Guyatt, G. H. (2001). Intention-to-treat principle. *Canadian Medical Association Journal, 165,* 1339-1341. Retrieved from http://www.cmaj.ca/cgi/reprint/165/10/1339.pdf

Nagin, D. S. (1998). Criminal deterrence research at the outset of the twenty-first century. *Crime & Justice, 23,* 1-42. Retrieved from http://www.jstor.org/stable/1147539

Neithercutt, M. G., & Gottfredson, D. M. (1974). *Case load size variation and difference in probation/parole performance.* NCJ 016576. Pittsburgh, PA: National Center for Juvenile Justice. Retrieved from http://www.ncjrs.gov/App/Publications/abstract.aspx?ID=16576

Paparozzi, M. A., & Gendreau, P. (2005). An intensive supervision program that worked: Service delivery, professional orientation, and organizational supportiveness. *The Prison Journal, 85,* 445-466. doi:10.1177/0032885505281529

Paternoster, R., Saltzman, L. E., Waldo, G. P., & Chiricos, T. G. (1983). Perceived risk and social control: Do sanctions really deter? *Law & Society Review, 17,* 457-480. Retrieved from http://www.jstor.org/stable/3053589

Petersilia, J. (1997). Probation in the United States. *Crime & Justice, 22,* 149-200. Retrieved from http://www.jstor.org/stable/1147573

Petersilia, J., & Turner, S. (1990). *Intensive supervision for high-risk probationers: Findings from three California experiments*. Santa Monica, CA: The Rand Corporation. Retrieved from http://www.rand.org/pubs/reports/2007/R3936.pdf

Petersilia, J., & Turner, S. (1993). Intensive probation and parole. *Crime & Justice, 17,* 281-335. Retrieved from http://www.jstor.org/stable/1147553

Rosch, J. (2006). Deviant peer contagion: Findings from the Duke Executive Sessions on Deviant Peer Contagion. *The Link, 5(2), Child Welfare League of America.* Retrieved from http://www.cwla.org/programs/juvenilejustice/thelink2006fall.pdf

Sarkisian, N. (2009). Topics in multivariate analysis: Count data models. Retrieved February 3, 2010, from http://www.sarkisian.net/sc704/count.pdf.

Schmidt, J. D., & Sherman, L. W. (1993). Does arrest deter domestic violence? *American Behavioral Scientist, 36,* 601-610. doi: 10.1177/0002764293036005005

Shapland, J., Atkinson, A., Atkinson, H., Dignan, J., Edwards, L., Hibbert, J., Howes, M., Johnstone, J., Robinson, G., & Sorsby, A. (2008). *Does restorative justice affect reconviction? The fourth report from the evaluation of three schemes.* Ministry of Justice Research Series 10/08. London, U.K.: Ministry of Justice. Retrieved from http://www.justice.gov.uk/restorative-justice-report_06-08.pdf

Sherman, L. W. (1993). Defiance, deterrence, and irrelevance: A theory of the criminal sanction. *Journal of Research in Crime & Delinquency, 30,* 445-473. doi:10.1177/0022427893030004006.

Sherman, L. W. (2007). The power few: Experimental criminology and the reduction of harm. *Journal of Experimental Criminology, 3,* 299-321. doi:10.1007/s11292-007-9044-y

Sherman, L. W., Gartin, P. R., & Buerger, M. E. (1989). Hot spots of predatory crime: routine activities and the criminology of place. *Criminology, 27,* 27-55. doi: 10.1111/j.1745-9125.1989.tb00862.x

Sherman, L. W., Gottfredson, D., MacKenzie, D., Eck, J., Reuter, P., & Bushway, S. (1997). *Preventing crime: What works, what doesn't, what's promising*. Washington, D.C.: United States Department of Justice, National Institute of Justice. Retrieved from http://www.ncjrs.gov/works

Sutherland, E. H. (1947). *Principles of criminology* (4th edition). Philadelphia, PA: Lippincott.

Taxman, F. S. (2002). Supervision: Exploring the dimensions of effectiveness. *Federal Probation*, *66(2),* 14-27. Retrieved from http://www.uscourts.gov/fedprob/ 2002sepfp.pdf

Taxman, F. S. (2008a). To be or not to be: Community supervision déjà vu. *Journal of Offender Rehabilitation, 47,* 209-219. doi:10.1080/10509670802134036

Taxman, F. S. (2008b). No illusions: Offender and organizational change in Maryland's proactive community supervision efforts. *Criminology & Public Policy, 7,* 275-302. doi:10.1111/j.1745-9133.2008.00508.x

Taxman, F. S., & Thanner, M. (2006). Risk, need, and responsivity (RNR): It all depends. *Crime & Delinquency, 52(1),* 28-51. doi:10.1177/0011128705281754

Warr, M., & Stafford, M. (1991). The influence of delinquent peers: What they think or what they do? *Criminology, 29,* 851-866. doi:10.1111/j.1745-9125.1991.tb01090.x

Weisburd, D., Bushway, S., Lum, C., & Yang, S.-M. (2004). Trajectories of crime at places: A longitudinal study of street segments in the city of Seattle. *Criminology, 42,* 283-321. doi:10.1111/j.1745-9125.2004.tb00521.x

Wilson, J. A., Naro, W., & Austin, J. F. (2007). *Innovations in probation: Assessing New York City's automated reporting system.* Washington, D.C.: JFA Associates. Retrieved from http://www.nyc.gov/html/prob/downloads/pdf/kiosk_report_2007.pdf

Worrall, J.L., Schram, P., Hays, E., & Newman, M. (2004). An analysis of the relationship between probation caseloads and property crime rates in California counties. *Journal of Criminal Justice*, *32,* 231-241. doi:10.1016/j.jcrimjus.2004.02.003

## Chapter 3

Ahlman, L. C., & Kurtz, E. M. (2008). *The APPD randomized controlled trial in low risk supervision.* Internal report. Philadelphia, PA: First Judicial District of Pennsylvania Adult Probation and Parole Department.

Andrews, D. A. (1989). Recidivism is predictable and can be influenced: Using risk assessments to reduce recidivism. *Forum on Corrections Research, 1(2),* 11-17. Retrieved from http://www.csc-scc.gc.ca/text/pblct/forum/special/a1e.pdf

Andrews, D. A., Bonta, J., & Hoge, R. D. (1990). Classification for effective rehabilitation: rediscovering psychology. *Criminal Justice & Behavior, 17,* 19-52. doi:10.1177/0093854890017001004

Barnes, G., Ahlman, L., Gill, C., Sherman, L. W., Kurtz, E., & Malvestuto, R. (Forthcoming). Low-intensity community supervision for low-risk offenders: A randomized, controlled trial. *Journal of Experimental Criminology*, in press.

Berk, R., Barnes, G., Ahlman, L., & Kurtz, E. (forthcoming). When second best is good enough: A comparison between a true experiment and a regression discontinuity quasi-experiment. *Journal of Experimental Criminology*, in press.

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 A, 172, Part 1*, 191-211. doi: 10.1111/j.1467-985X.2008.00556.x

Blumstein, A., Cohen, J., Roth, J. A., & Visher, C. A. (Eds.). (1986). *Criminal careers and "career criminals": Report of the National Academy of Sciences Panel on Research on Criminal Careers.* Washington, D.C.: National Academy Press.

Bonta, J. (1996). Risk-needs assessment and treatment. In A. T. Harland (Ed.), *Choosing correctional options that work: Defining the demand and evaluating the supply* (pp. 18-32). Thousand Oaks, CA: Sage Publications.

Bonta, J. (2002). Offender risk assessment: Guidelines for selection and use. *Criminal Justice & Behavior, 29,* 355-379. doi:10.1177/0093854802029004002

Bridges, G. S., & Steen, S. (1998). Racial disparities in official assessments of juvenile offenders: Attributional stereotypes as mediating mechanisms. *American Sociological Review, 63,* 554-570. Retrieved from http://www.jstor.org/stable/2657267

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* (pp. 221-234). Springfield, IL: Illinois State Board of Parole.

Clear, T. R,. & Braga, A. A. (1995). Community corrections. In J. Q. Wilson, & J. Petersilia (Eds.), *Crime* (pp. 421-444). San Francisco, CA: Institute for Contemporary Studies.

Cohen, M. (1988). Some new evidence on the seriousness of crime. *Criminology, 26,* 343-353. doi:10.1111/j.1745-9125.1988.tb00845.x

Cohen, M. A. (2000). Measuring the costs and benefits of crime and justice. In D. Duffee (Ed.), *Measurement and analysis of crime and justice* (Vol. 4, pp. 263-315). Washington, D.C.: National Institute of Justice. Retrieved from http://www.ncjrs.gov/criminal_justice2000/vol_4/04f.pdf

Cohen, M. A., Rust, R. T., Steen, S., & Tidd, S. T. (2004). Willingness-to-pay for crime control programs. *Criminology, 42,* 89-109. doi:10.1111/j.1745-9125.2004.tb00514.x

Erwin, B. S. (1986). Turning up the heat on probationers in Georgia. *Federal Probation, 50(2),* 17-24. Retrieved from http://www.heinonline.org/HOL/Page?handle=hein.journals/fedpro50&id=1&size=2&collection=journals&index=journals/fedpro

Farrington, D. P. (1998). Predictors, causes, and correlates of male youth violence. *Crime & Justice, 24,* 421-475. Retrieved from http://www.jstor.org/stable/1147589

Federal Bureau of Investigation (2004). *Uniform crime reporting handbook*. Washington, D.C.: U. S. Department of Justice. Retrieved from http://www.fbi.gov/ucr/ucr.htm

Glaze, L. E, & Bonczar, T. P. (2009). *Probation and parole in the United States, 2008*. Bureau of Justice Statistics Bulletin, December 2009. NCJ 228230. Washington, DC: U.S. Department of Justice Office of Justice Programs. Retrieved from http://bjs.ojp.usdoj.gov/content/pub/pdf/ppus08.pdf

Glueck, S., & Glueck, E. (1950). *Unraveling juvenile delinquency.* New York, NY: Commonwealth Fund.

Gottfredson, S. D., & Jarjoura, G. R. (1996). Race, gender, and guidelines-based decision making. *Journal of Research in Crime & Delinquency, 33,* 49-69. doi: 10.1177/0022427896033001004

Gottfredson, S. D., & Moriarty, L. J. (2006). Statistical risk assessment: Old problems and new applications. *Crime & Delinquency, 52,* 178-200. doi:10.1177/ 0011128705281748

Grove, W. M., & Meehl, P. E. (1996). Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical-statistical controversy. *Psychology, Public Policy, & Law, 2,* 293-323. doi:10.1037/1076-8971.2.2.293

Hanley, D. (2006). Appropriate services: Examining the case classification principle. *Journal of Offender Rehabilitation, 42(4),* 1-22. doi:10.1300/J076v42n04_01

Howard, P., Francis, B., Soothill, K., & Humphreys, L. (2009). *OGRS 3: The revised Offender Group Reconviction Scale*. Ministry of Justice Research Summary 7/09. Retrieved from http://www.justice.gov.uk/oasys-research-summary-07-09-ii.pdf

Lipsey, M. W., & Derzon, J. H. (1998). Predictors of violent or serious delinquency in adolescence and early adulthood: A synthesis of longitudinal research. In R. Loeber & D. P. Farrington (Eds.), *Serious and violent juvenile offenders: risk factors and successful interventions* (pp. 86-105). Thousand Oaks, CA: Sage Publications.

Lowenkamp, C. T., & Latessa, E. J. (2004). Understanding the risk principle: How and why correctional interventions can harm low-risk offenders. *Topics in Community Corrections 2004* (pp. 3-8). Washington, D.C.: U.S. Department of Justice, National Institute of Corrections. Retrieved from http://www.nicic.org/pubs/2004/period266.pdf

Lowenkamp, C. T., Holsinger, A. M., & Latessa, E. J. (2001). Risk/need assessment, offender classification, and the role of childhood abuse. *Criminal Justice & Behavior, 28,* 543-563. doi:10.1177/009385480102800501

Marsh, K., & Fox, C., (2008). The benefit and cost of prison in the UK. The results of a model of lifetime offending. *Journal of Experimental Criminology, 4,* 403-423. doi: 10.1007/s11292-008-9063-3.

Marsh, K., Chalfin, A., & Roman, J. K. (2008). What does cost-benefit analysis add to decision making? Evidence from the criminal justice literature. *Journal of Experimental Criminology, 4,* 117-135. doi: 10.1007/s11292-008-9049-1

Miller, T. R., Cohen, M. A., & Wiersema, B. (1996). *Victim costs and consequences: A new look.* NCJ 155282. Washington, D.C.: U. S. Department of Justice, National Institute of Justice. Retrieved from http://www.ncjrs.gov/pdffiles/victcost.pdf

Montori, V. M., & Guyatt, G. H. (2001). Intention-to-treat principle. *Canadian Medical Association Journal, 165,* 1339-1341. Retrieved from http://www.cmaj.ca/cgi/reprint/165/10/1339.pdf

Nagin, D. S., & Paternoster, R. (1991). On the relationship of past to future participation in delinquency. *Criminology, 29,* 163-189. doi:10.1111/j.17459125.1991.tb01063.x

Nagin D. S., & Farrington, D. P. (1992). The stability of criminal potential from childhood to adulthood. *Criminology, 30,* 235-260. doi:10.1111/j.1745-9125.1992.tb01104.x

Pennsylvania Fraternal Order of Police Lodge No. 38 (2008, September 30). *Rendell calls for suspension of paroles.* Retrieved from http://pafop38.com/2008/09/30/ rendell-calls-for-suspenion-of-paroles/ [sic]

Petersilia, J. (1997). Probation in the United States. *Crime & Justice, 22,* 149-200. Retrieved from http://www.jstor.org/stable/1147573

Petersilia, J., & Turner, S. (1990). *Intensive supervision for high-risk probationers: Findings from three California experiments*. Santa Monica, CA: The Rand Corporation. Retrieved from http://www.rand.org/pubs/reports/2007/R3936.pdf

Ramchand, R., MacDonald, J. M., Haviland, A., & Morral, A. R. (2009). A developmental approach for measuring the severity of crimes. *Journal of Quantitative Criminology, 25,* 129-153. doi:10.1007/s10940-008-9061-7

Rubin, S. (1975). Probation or prison: Applying the principle of the least restrictive alternative. *Crime & Delinquency, 21,* 331-336. doi: 10.1177/ 001112877502100404

Ruth, H., & Reitz, K. R. (2003). *The challenge of crime: Rethinking our response*. Cambridge, MA: Harvard University Press.

Sherman, L. W. (1993). Defiance, deterrence, and irrelevance: A theory of the criminal sanction. *Journal of Research in Crime & Delinquency, 30,* 445-473. doi:10.1177/0022427893030004006.

Sherman, L. W. (2007). Use probation to prevent murder. *Criminology & Public Policy, 6,* 843-850. doi:10.1111/j.1745-9133.2007.00461.x

Sellin, T., & Wolfgang, M. E. (1978). *The measurement of delinquency*. Montclair, NJ: Patterson Smith.

Van Voorhis, P., & Brown, K. (1997). *Risk classification in the 1990s.* Washington, D.C.: National Institute of Corrections. Retrieved from http://nicic.org/Downloads/PDF/Library/013243.pdf