Systematic Review Protocol:
Probation Intensity Effects on Probationers’ Criminal Conduct

Lead/contact reviewer:
Charlotte E. Gill, Ph.D.
Center for Evidence-Based Crime Policy
Department of Criminology, Law and Society
George Mason University
4400 University Drive, MS 6D3
Fairfax, VA 22030 USA
cgill9@gmu.edu

Co-authors:
Jordan Hyatt, J.D., M.S.
Jerry Lee Center of Criminology
Department of Criminology
University of Pennsylvania
McNeil Building, Suite 483
3718 Locust Walk
Philadelphia, PA 19104 USA
jhyatt@sas.upenn.edu

Lawrence W. Sherman, Ph.D.
Jerry Lee Centre for Experimental Criminology
Institute of Criminology
Cambridge University
Sidgwick Avenue
Cambridge CB3 9DA UK
ls434@cam.ac.uk

&

Department of Criminology and Criminal Justice
University of Maryland College Park
College Park, MD 20742 USA
sherman1@umd.edu

First submission: December 9, 2009
Latest revision: October 8, 2010
BACKGROUND

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. 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. 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.

Skeem and Manchak (2008) propose that probation supervision may follow one of three broad guiding philosophies: control/surveillance, treatment, or a hybrid of both. 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.

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.

---

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).
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, 1992; 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; 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., 2010; 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.

OBJECTIVES

We will 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 will 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 plan to 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?

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.
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 will be 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 will be eligible as long as the probation officer is the primary supervisor. This limitation will allow us to maintain a degree of equivalence between treatment providers and settings, and between treatment and control group conditions.

We will 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 our sample to randomized controlled trials (RCTs) and highly rigorous quasi-experiments involving subject-level matching and, pre- and post-program measures of offending behavior. 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.

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.
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 will be 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 will 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 will 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 are 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 will be coded.

Types of Outcomes

Eligible studies will measure recidivism in terms of new arrests and/or convictions. Technical violations of probation, such as absconding or failing a drug test, will also be 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. Outcome data will most likely be drawn from official records, but we will also include self-reported data if available.

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.

Settings and Timeframe

Studies will not be excluded on the basis of language or geography. Studies from the late 1950s, when the earliest wave of research on intensive probation began, to the present day will be considered for inclusion.
Search strategy for identification of relevant studies

We will use several strategies to conduct a comprehensive search for literature on probation intensity. The primary literature search will involve keyword searches of online abstract databases and the websites of research organizations and government agencies (see below). Specialist search engines like Google Scholar will also provide a rich source of ‘grey literature.’ We will also consult 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 library book and microfilm collections. Online searches will be supplemented with hand searches of key journals in the field. Every effort will made to locate unpublished material where possible. Eligibility of studies will be assessed by reading titles and abstracts, and obtaining the full text of documents that appeared to be relevant.

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
7. International Bibliography of the Social Sciences
8. ISI Web of Knowledge
9. JSTOR
11. PsycINFO
12. Sage Full Text Collection: Criminology
13. Sage Full Text Collection: Political Science
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
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.)
10. National Probation Service (U.K.)

5 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.
11. Pew Center on the States (U.S.A.)
12. RAND Corporation (chiefly U.S.A.)
14. Urban Institute (U.S.A.)
15. Vera Institute of Justice (U.S.A.)
16. Washington State Institute of Public Policy (U.S.A.)

The following search strings of key words will be used to search the databases and websites, adapted as necessary to meet the requirements of the different search engines. The search terms are deliberately broad (they do not include limiting terms such as ‘evaluation,’ ‘experiment,’ ‘trial’) so that relevant background literature may also be systematically obtained through the searches. ‘*’ indicates where terms are truncated to find all possible variants of the word:

\[ \text{probation}^* \text{ AND supervis}^* \text{ AND case}^* \text{ AND (intens* OR frequen* OR ratio)} \]

\[ \text{AND (recidiv* OR *arrest* OR *convict*)} \]

Description of methods used in primary research

The ‘classic’ model of ISP was tested in the BJA/RAND experiments from the 1980s (Petersilia & Turner, 1993), which 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 analysis we will 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 will initially take the more conservative approach of handling different types of outcome measure separately. However, we will combine arrests and convictions in some analyses. In these cases, arrest outcomes will 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 will 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 will use the data for the full sample or outcome only. Where enough studies provide results broken down by the same types of subgroups, we will analyze those outcomes separately.

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, sample sizes may decrease significantly over time as cases are lost to follow-up. In these cases we will select 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 will count the sample as one study. The report containing the longest follow-up period and/or the most detail is considered the primary study, and other reports will be used to supplement the data from the primary study where necessary. We will ensure that each re-analyses of the same datasets are not inadvertently included with primary evaluation data from the same research project.

Details of study coding categories

The coding protocol developed for this study is reproduced in an appendix to this protocol. 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 will record 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 will serve as both determinants of eligibility and potential moderator variables. We do not expect all these factors to influence outcomes and will not test them all to avoid the problem of finding statistically significant results merely by chance. However, we also aim to be as inclusive as possible so that potentially relevant information is not missed. The main items of interest relate to the four research questions set out above. The first and second authors will double-code a subset of studies and separately code those remaining to ensure that both coders share the same understanding of the coding protocol.
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.

Assessment of reporting bias

Our search strategy covers a range of databases that will enable us to identify unpublished literature (e.g., dissertations and technical reports) as well as published works. We will compare results from published and unpublished studies to estimate reporting bias, and if data are sufficient we will statistically test for publication bias using the funnel plot and trim-and-fill methods (Duval & Tweedie, 2000).

Analytic strategy

Meta-analytic procedures will be used to quantitatively combine effect size data from the eligible studies where appropriate (i.e., where two or more studies are available that measure a common outcome, such as arrests, and contained sufficient information to calculate an effect size). Effect sizes for each outcome measure in the studies will be encoded according to the 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 will 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 of an event (e.g., arrest) occurring. 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.6

Meta-analytic methods will also be 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 will use the meta-analytic analog to the analysis of variance (ANOVA) to test whether each these factors separately might account for any variability in the observed effect sizes from each study (Lipsey & Wilson, 2001, pp. 120-122).

Due to the greater risk of bias in non-randomized studies, experimental and quasi-experimental results will be 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 will be analyzed with the quasi-experiments. The concern with such experiments is that the attrition

6 Note that although results will be presented 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).
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.

Computations of effect sizes and related statistics will be performed using meta-analysis macros for STATA software (Wilson, 2002). We will use RevMan software (Cochrane Collaboration, 2008) to construct forest plots for the graphical representation of meta-analysis results.

**SOURCES OF SUPPORT**

We are grateful to the Jerry Lee Foundation and the School of Arts and Sciences at the University of Pennsylvania for funding the first author’s doctoral studies and dissertation, on which this review is based.

**ACKNOWLEDGMENTS**

We thank David Wilson, the Campbell Collaboration Crime and Justice Group Steering Committee, and the anonymous peer reviewers for their helpful feedback on this protocol. Charlotte Gill would like to thank Larry Sherman, John MacDonald, and Paul Allison for their support during the original development of this project as a doctoral dissertation.

**DECLARATIONS OF INTEREST**

All three authors have been involved in conducting randomized trials of probation supervision. The second and third authors are in the process of developing further experiments in this field. However, the authors are committed to the objective use of rigorous research to inform crime policy, and their conclusions here are not affected by the outcomes of their primary research.

**REFERENCES**


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
   8. Other: report, conference paper

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? prodesc
   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: Yes0. Noprogir
   Program decreases ratio of clients to probation officers: Yes0. Noprogdr
   Program increases frequency of contact with probation officer: Yes0. Noprogif
   Program decreases frequency of contact with probation officer: Yes0. Noprogdf
   Program increases drug testing requirements: Yes0. Noprogdt
   Program decreases drug testing requirements: Yes0. Noprogddt
   Other: Yes0. No progoth

C6c. If Yes for any of the above, state exact numbers if available (999 if not):
   Control ratio: / Treatment ratio: racx1/ratx
   Control freq: / Treatment freq: frcx1/frtx
   Control drug tests: / Treatment drug tests: drcx1/drtx
   Other: txcxloth

C6d. Additional program components (indicate whether present or not):
   Curfew: Yes0. Noaddcomp_curf
   Drug treatment: Yes0. Noaddcomp_drug
Electronic monitoring. Yes 0. No
Employment program/assistance. Yes 0. No
Halfway house. Yes 0. No
Home visits. Yes 0. No
House arrest. Yes 0. No
Offense-specific treatment. Yes 0. No (e.g., sex offender treatment)
Other treatment. Yes 0. No
Other: 1. Yes 0. No

C7. What happened to the comparison group?
‘Supervision as usual’
Other:

C8. Was supervision for treatment group provided by anyone other than probation officer?
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: txlmm
Fixed (same for all subjects): txlfix

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

C11. What supervision philosophy was stated?
Control/surveillance 8. Other:
Treatment 3. Hybrid 9. Don’t know/not stated

C12. Did the intervention remain consistent over time?
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 1. Yes

C14. Subject-level matching?
0. No 1. Yes

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

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

C17. Rating of initial similarity between treatment and control group:
<p>| | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
</tbody>
</table>

(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: $D7 + D8 = \text{total sample size prior to attrition}$. 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
   7. Adult (21+)
   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%)
   6. Female
   7. Male
   8. Other:
   9. Unspecified/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
6. N/A
7. No need assessment
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
   7. Other:
E8. Offense types included in recidivism measure:
   All offenses (‘No’ for others): 1. Yes 0. No
   Drug offenses: 1. Yes 0. No
   Person offenses, sexual: 1. Yes 0. No
   Person offenses, nonsexual: 1. Yes 0. No
   Person offenses, unspecified: 1. Yes 0. No
   Property offenses: 1. Yes 0. No
   Weapons offenses: 1. Yes 0. No
   Driving offenses: 1. Yes 0. No
   Technical or status offenses: 1. Yes 0. No
   Other: 1. Yes 0. No
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.)
E11. Length of follow-up period: fulng
   1. < 6 months
   2. 6-12 months
   3. > 1, < 2 years
   4. > 2 years
   5. No follow-up
   6. 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)
   4. Follow-up (all subsequent measurement points, post-intervention)
F9. Does the investigator report the difference as statistically significant?essig
   0. No
   1. Yes
   2. Not tested
   3. Can’t tell
F10. If tested, what type of statistical test was used?estest
    1. t test
    2. F test
    3. χ²
    4. Regression analysis
F11. Timeframe in months captured by the measure (weeks/4.3)
    Minimum:estmin
    Maximum:estmax
    Mean:estm
F12. Timeframe in months from end of program to measurement point (weeks/4.3)
    Minimum:esfumin
    Maximum:esfumax
    Mean:esfum

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: esexls
F27. Proportion successful in treatment group: estxspr
F28. Proportion successful in comparison group: esexlspr
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