Low-intensity community supervision for low-risk offenders: A randomized, controlled trial

6 authors, including:

Geoffrey C. Barnes
University of Cambridge
22 PUBLICATIONS **1,502** CITATIONS

SEE PROFILE

Lindsay Ahlman
University of Maryland, College Park
7 PUBLICATIONS **326** CITATIONS

SEE PROFILE

Charlotte Gill
George Mason University
30 PUBLICATIONS **1,326** CITATIONS

SEE PROFILE

Lawrence William Sherman
University of Cambridge
176 PUBLICATIONS **19,237** CITATIONS

SEE PROFILE

Low-intensity community supervision for low-risk offenders: a randomized, controlled trial

**Geoffrey C. Barnes · Lindsay Ahlman ·
Charlotte Gill ·Lawrence W. Sherman ·
Ellen Kurtz · Robert Malvestuto**

Abstract The Philadelphia Low-Intensity Community Supervision Experiment provides evidence on the effects of lowering the intensity of community supervision with low-risk offenders in an urban, US county community corrections agency. Using a random forests forecasting model for serious crime based on Berk et al. *Journal of the Royal Statistical Society, Series A*, *172*(Part 1), 191–211, 2009, 1,559 low-risk offenders were identified and randomly assigned to either standard or reduced frequency of mandatory office visits. Treatment as assigned was substantially delivered at 4.5 probation visits per year versus 2.4, for as long as offenders remained on active probation or parole. In a one-year follow-up for all cases, outcomes examined were the prevalence, frequency, seriousness and time-to-failure of arrests for new crimes committed after random assignment was implemented. No significant differences ($p=.05$) in outcomes were found between standard and low-intensity groups. Non-significant differences for offense seriousness favored the low-intensity group. We conclude that lower-intensity supervision at the tested level of dosage can allow fewer officers to supervise low-risk offenders in the community without evidence of increased volume or seriousness of crime.

G. C. Barnes (✉) · C. Gill
Jerry Lee Center of Criminology, University of Pennsylvania, Philadelphia, PA, USA
e-mail: gbarnes@sas.upenn.edu

L. Ahlman · E. Kurtz · R. Malvestuto
Adult Probation and Parole Department of the First Judicial District of Pennsylvania, Philadelphia, PA, USA

L. W. Sherman
University of Cambridge, Cambridge, UK

1 Risk-based policy is not inherently evidence-based

There is a growing consensus among reformers that corrections should employ both risk classification and evidence-based practices. This recommendation may imply that policies requiring a case-by-case analysis of offender risk are "risk-based" and thus inherently *evidence*-based. They are not. The evidence requirements for treatment are completely separate from the evidence needed for diagnosis. Just because a *diagnosis* or prognosis is based on evidence, that does not mean that the prescribed *treatment* has any evidence for its success. It is only when a risk-based policy of classification has been successfully tested for outcomes of specific treatments that it can become an evidence-based treatment policy. Yet this point may not be readily apparent to even the most sophisticated observers of correctional policies.

The recent Pew Center on the States (2009) report on community corrections, for example, has two leading recommendations that can be summarized as (1) using risk-based sorting of offenders, for (2) choosing evidence-based interventions. Yet in the absence of evidence about the *effects* on recidivism of corrections agencies using risk information, risk-based sorting does not (yet) constitute, in itself, an evidence-based practice. The two key recommendations of this influential and well-researched Pew report, then, are in stark contradiction. Until each separate risk-based treatment has been tested, using any of them would violate the report's injunction to employ evidence-based practice. In order for risk-based corrections to *also* constitute evidence-based practice, the use of risk information combined with treatment comparisons must itself be subjected to empirical testing in order to assess its consequences for key outcomes.

One major form of risk-based corrections is the growing use of low-intensity supervision for low-risk probationers and parolees. As Jacobson (2005: 169) recommends, risk assessment instruments can be used to identify "relatively low-risk parolees who require only minimal, if any, supervision." Jacobson himself introduced such risk-based supervision in the New York City Probation Department in the mid-1990s, with up to 70% of probationers now assigned to low-*intensity* supervision because they were assessed as low-*risk* (Wilson et al. 2007). Probation officer caseload size for low-risk probationers increased from 283:1 to 492:1.

The New York model of low-intensity supervision consisted of an automated kiosk reporting system. Low-risk probationers check in at the computerized kiosk (similar in design to a banking ATM) once a month to answer basic questions about their contact details, employment, and new arrests. The process takes around four minutes. If any issues arise, such as new arrests or a change of details, or if the client is selected for random drug testing, the client must see an officer in person before leaving. Compliance with this requirement is enforced by kiosk attendants.

Trend analysis of the New York program shows that arrests of low-risk probationers have declined since the program was fully implemented in 2003–2006 (Wilson et al. 2007). That trend has assured its administrators of its safety and effectiveness. Yet a before–after, no comparison-group analysis is not sufficiently rigorous for low-intensity supervision to qualify as evidence-based policy, by almost any definition. Such definitions are hard to find in authoritative sources, although all of them clearly imply the importance of unbiased comparisons. A recent National

Research Council (NRC) report, for example, used the phrase "evidence-based" repeatedly in the report and its glossary, but shied away from ever explicitly defining the term itself (O'Connell et al. 2009: 371). The Coalition for Evidence-Based Policy (2010) nonetheless found enough material in the NRC report to cite it as consistent with the Coalition's central theme that "evidence of effectiveness generally cannot be considered definitive without ultimate confirmation in well-conducted randomized controlled trials."

The Maryland Report to the US Congress (Sherman et al. 1997) used a more relaxed definition of "evidence-based," requiring two or more quasi-experimental studies using similar sample characteristics in test and comparison units to show that a program "worked," in the absence of substantial evidence to the contrary from other tests. A before–after trend with no comparison group, however, did not even meet this more generous definition of "evidence-based" policy.

Because there is no evidence available on the New York model that includes any kind of comparison group, elected officials could rightly ask whether its low-intensity probation program is truly safe or effective. Therein lies the potential confusion between risk assessment and evidence-based practice. If municipal leaders want to know whether low-intensity supervision is as safe for the public as "normal" supervision is, the risk-based answer would be "yes, because of the evidence that they are low-risk." The *evidence*-based answer, however, would be this: that in a fair comparison of low-intensity and normal supervision, low-risk probationers had no different rates of prevalence, frequency, seriousness, or time-to-failure in repeat offending.

This answer captures the efforts of the Obama administration to re-brand two decades of "evidence-based" practice in medicine with what may be a much clearer name: "comparative effectiveness" research. It is only by comparing two different practices that any controlled testing can generate answers to the questions of "what works" or "how cost-effective is this treatment." Thus, in testing the consequences of risk-based practices, the appropriate comparison is with practices that are not based on risk classification. This comparison simultaneously incorporates the use of *both* a classification system *and* a difference in treatment that is applied to *each specific* risk level when it is isolated from the broader population of offenders at *all* risk levels. The comparison is between a "one-size-fits-all" (OSFA) treatment (when used for that risk level) versus a unique-to-that-risk-level (UTTRL) treatment. A transition from risk-blind to risk-based treatments, then, would logically require separate comparisons between the OSFA and the UTTRL for each level of risk. It may even require comparing different treatments that are UTTRL for each level of risk, either with each other or with the standard OSFA treatment.

This article reports the first in a series of planned comparisons between OSFA and UTTRL treatments at different levels of risk. It reports on the first such comparison, conducted at the lowest risk level. The article also reports on what appears to constitute the first randomized trial of low-intensity probation or parole supervision at any risk level. It may also be the first evidence-based answer to the key questions about any model of low-intensity supervision, or supervision of low-risk offenders.

Commendably, the community corrections agency conducting the experiment did so in order to decide whether or not to adopt risk-based resource allocation, rather than as an after-the-fact examination of whether making such a change had been

beneficial. It was only after the results of the experiment were obtained that the agency began to convert its city-wide operations to risk-based classifications for different levels of intensity of supervision. The goal of that conversion, and the reason for conducting the experiment, was to reallocate scarce resources in a way that might yield more crime prevention per dollar invested in probation and parole, and specifically more prevention of homicide (Sherman 2007).

This 2007–2008 field experiment tested low-intensity supervisory practices based in part on the New York model. The test was conducted by the Adult Probation and Parole Department (APPD) of the First Judicial District of Pennsylvania in Philadelphia. The APPD agreed to use random assignment in reducing probation or parole supervision visits from a policy of about once every month to a policy of once every six months. This reduction was assigned to half of a large population of offenders who had previously been identified as "low-risk" by a statistical model developed from over 30,000 previous cases and tested on an independent sample (see below). These offenders included both persons sentenced to a term of community supervision, and those who were paroled to community supervision from a term of imprisonment of less than 24 months in the county (city) jail system.[1]

Classification of all such offenders as either low-risk, high-risk, or neither high nor low was performed independently by criminologists at the University of Pennsylvania, in collaboration with the research staff of the APPD. The random assignment sequences and outcome analyses reported here were also generated at the University. The low-intensity model of supervision, like the OSFA standard model of supervision, was designed and implemented by the APPD.

This article reports the conduct and results of what we call the "Philadelphia Low-Intensity Community Supervision Experiment." It begins with a review of the relevant criminological theory and related previous research on which the experiment builds. It then describes the procedures used for risk assessment, the eligibility criteria, the sample selection procedures (and their errors), and random assignment procedures. Evidence on the implementation and variable experiences of offenders in the two groups completes the description of the experimental conditions. Evidence on outcomes is then offered for measures of prevalence, frequency, and time to failure. A power analysis concludes the data presented, showing how much chance the design gave for any true underlying differences to emerge in the comparison.

2 Community supervision and criminological theory

There are at least three perspectives in criminological theory that can be applied to predict the result of reducing the intensity of community supervision of sentenced offenders. One of them is specific deterrence, which predicts that less supervision will cause more crime. The other two theories are defiance and deviant peer contagion, both of which predict the opposite effect from deterrence: less crime with less supervision. While the experiment was not designed explicitly to test any of

---

[1] The operating practices of the APPD do not distinguish between probationers and parolees, largely because many offenders have multiple cases simultaneously at different stages of the system. It is possible, for example, to be on probation for one offense at the same time as being on parole for another offense.

these three perspectives, it is useful to note the possibilities and limitations of what the experiment may have to say about each of them.

2.1 Specific deterrence: more crime predicted

Specific (or "special") deterrence is the doctrine that applying sanctions to individuals who have already committed crimes will reduce the likelihood of their committing more crimes in the future (Zimring and Hawkins 1973; Gibbs 1975). In its classical utilitarian formulation, it had three components: certainty, severity and celerity (speed). When applied to the question of lowering the frequency of required contacts with probation or parole officers, specific deterrence predicts that less severity of the sanction will be less effective in reducing future offending. Because it inflicts less pain (or even inconvenience), and because offenders under community supervision are free to offend, the effect of less "severe" supervision is predicted to occur immediately.

Severity is, of course, only one of six potential constructs affecting decisions to offend. Deterrence doctrine also says that general deterrence—including *general* severity, certainty, and celerity—affects decisions to commit crime. Those three constructs combine with *specific* severity, certainty, and celerity to sum up to six elements of deterrence that could, in theory, be manipulated. This experiment manipulated only one of those six elements. Moreover, deterrence is only a doctrine, and not a theory (cf. Gibbs 1975), precisely because it offers no logic for assigning equal or different weighting to each of the six constructs. Nor does it postulate any conditional relationships, such as the necessity of a certain threshold of severity to be met before certainty or celerity can cause any effect on behavior.

In the absence of any theoretical algorithm, we can only observe that changing one element in six leaves the majority of deterrence doctrine unchanged. Theoretically, then, a marginal reduction in severity may not be enough to matter. As long as offenders are swiftly and consistently declared in technical violation as soon as they miss an appointment or fail a drug test, the other conditions of specific deterrence would remain unchanged. And as long as overall sentencing patterns, arrest risks, and the number of police remain unchanged, the experimental intervention may likely not constitute much change in deterrence. Even more important may be a lack of change in police patterns of serving bench warrants for fugitives (see Goffman 2009)—or not serving them, as is far more typical. While what actually matters is an empirical question, the theoretical argument for predicting more crime from less supervision appears manifestly weak.

That argument is good news from a policy perspective, of course. But it is bad news for the advancement of deterrence theory. Whatever the findings show, they will not make much of a dent in the fog of non-specification that surrounds the deterrence doctrine. This is especially important in light of other experiments finding both crime reductions and crime increases from the same intervention, depending upon the offenders' social bonds (e.g., Sherman 1993).

2.2 Defiance theory: less crime predicted

Defiance theory (Sherman 1993: 450) predicts that criminal "sanctions provoke future *defiance* of the law (persistence, more frequent or more serious violations) to

the extent that offenders experience sanctioning conduct as illegitimate, that offenders have weak bonds to the sanctioning agent and community, and that offenders deny their shame and become proud of their isolation from the community." Under this theory, the less often sanctions are applied to offenders who meet the conditions that the theory specifies, the less crime such offenders will commit in response to the provocation of criminal sanctions.

While not all offenders sentenced to probation or parole may meet the conditions of defiance theory, we can assume that at least some of them will meet some or all of the three conditions the theory postulates: (1) illegitimate sanctioning, (2) weak bonds to the community, and (3) pride rather than shame at a criminal conviction.

First, many APPD clients will find reason to challenge the *legitimacy* of the system. There are great frustrations in traveling to the APPD office from the far reaches of a big city, and of enduring often-long waits in crowded conditions. Offenders could quite easily become angry at even the prospect of "going downtown," let alone when returning to their homes after what they may see as a humiliating day of forced submission to authority. To the extent that such reactions may occur after each and every probation supervision visit, this may erode their moral intuitions that this is a fair and reasonable punishment. To the extent that they witness a high concentration of minorities and few middle-class whites attending the probation office, this may also create a sense of unfairness (however unjustified) about apparent racial and class inequality of criminal justice in Philadelphia.

Second, many APPD clients are also poorly bonded to society, with little "stake in conformity." Their employment prospects are already spoiled by their criminal records. They have very high rates of school dropout and low levels of educational attainment. Many have unstable family or housing situations, leaving them one argument away from homelessness. Those who have jobs may risk losing them by simply complying with the requirement to visit APPD during the conventional work-week. They may have little to lose by any further criminal sanctions, but much to "gain" from either technical violations or committing new crimes.

Third, many APPD clients may live in areas with high prevalence of criminal convictions, especially among young males. In such a social context, there may be few sources of honor or prestige. Being "bad" enough to come to the attention of the law may become a source of pride (Sherman 1993), especially given the prestige already accorded to other young men who have "stood up" to the system. Not letting anyone push you around—not a cop, not a judge, not a probation officer—can be part of the general "code of the street" (Anderson 1999).

The experiment is unable to explore the differences *within* the groups on the elements or indicators of defiance theory. The low prevalence of recidivism requires that statistically powerful tests be limited, by design, to the main effects *between* groups. The experiment is thus no more a test of defiance theory than it is of deterrence doctrine. Its results can be *consistent* with one or the other of those perspectives. But the results of the main effect design—in what was intended to be a policy experiment—cannot constitute a fully elaborated test of any theory of the criminal sanction.

2.3 Deviant peer contagion and deviancy training: less crime predicted

The same can be said for the third theoretical perspective that predicts the main effects of this experiment: deviant peer contagion. The fact that all probation visits during this experiment occurred in the same location is central to this prediction (Dodge et al. 2006). The same is true for the theory's corollary, *deviancy training* theory (Dodge and Dishion 2006). This theoretical perspective has been induced, in part, from substantial empirical evidence (Lipsey 2006) that even successful treatment programs are often less successful when delivered in a group setting rather than in individual training. The theory therefore predicts the following:

> "When prevention and treatment programs are administered in ways that place deviant youth with deviant peers, programs are more likely to have adverse effects—the conduct of adolescents in these programs often worsens. This may be due to the processes of labeling, communication and acquisition of cultural norms, reinforcement, and deviancy training" (Rosch 2006:4).

At first glance, the APPD is forced to see its clients under conditions highly typical of the kinds of deviant peer concentrations described by the Duke University Executive Session on deviant peer contagion (Dodge et al. 2006). Efficiency has long required that over 40,000 Philadelphians at any given time be scheduled to journey to one central location for all their probation or county parole contacts. This concentration often results in long lines of sentenced offenders waiting outside (in all weather) while they await scanning by metal detectors and other procedures before they are allowed into the building. After rising early in the morning (often before dawn) to arrive on time, the offenders may then wait inside for hours until they are able to see their probation officer for a brief meeting. Until that meeting occurs, they have almost no one else to talk to except other offenders. This may expose them to what Dishion describes as "deviancy training".

> "Deviancy training occurs when a peer displays antisocial behavior or talks about it and other peers positively reinforce that behavior by smiling or giving verbal approval and high status to the first peer. A youth observes this norm and then engages in similar talk or behavior, which is also reinforced. Soon, the youth is drawn into the peer culture and becomes more deviant." (Dodge et al. 2006)

> "Evidence also suggests that young adolescents are most susceptible to deviant peer influence, and the effects are *most severe for youth with modest levels of delinquency.*" [emphasis added] (Rosch 2006: 4).

The fact that the sample in this experiment consists of low-risk offenders may suggest that they are particularly susceptible to deviancy training reactions. Nonetheless, that effect may be mitigated by the older age of the sample than in many of the studies reviewed by Dodge et al (2006). Because the sample in the present experiment had an average age of 40, the peer contagion perspective may not necessarily predict higher rates of offending with more frequent deviant concentrations, nor predict lower rates of offending with less frequent mixing of offenders

with high concentrations of deviant peers. As Rosch (2006:4) observes about delinquents, but not about adult offenders:

> An array of moderating factors, such as age of the children in the program, the program's duration, the experience level of program staff, and the structure of the program, may either exacerbate or minimize the adverse effects of interaction with deviant peers. Because the possibility of deviant peer effects has rarely been the explicit focus of rigorous academic research, there is much we do not know."

In summary, the state of theory requires an open mind about the effects of low-intensity supervision on offending. There is no theoretical consensus to cite as a reason *not* to undertake such a test. Regardless of the result, there is no way to say it was clearly predictable in advance.

We are also conscious that the experiment is very limited in what it can say about any of these theoretical perspectives. While the main effects could in principle be consistent with one or more of the predictions, they cannot constitute a "crucial" test of any of the theories. Embedding such a test in the research design would have required either (1) many more cases (to examine moderator effects of pre-random assignment characteristics), (2) many more treatment conditions (to map possible dose-response curves), (3) direct but costly measures of offender experiences in probation (such as "deviancy training"), as well as their pre-post differences in attitudes and perceptions under different treatments as assigned, or (4) all of the above. All that was beyond the scope of the present experiment.

What can be said is this: specific deterrence doctrine predicts more crime with less supervision. Defiance and deviant contagion predict less crime with less supervision. Neither result is predictable based on previous research on high-intensity supervision. As apparently the first experiment in reducing supervision (see next section), the present experiment has no priors as a basis for evidence-based prediction. As an attempt to falsify any and all predictions based on theory alone, it will provide some insight to theory, and possibly far more to policy.

3 Risk classification and supervision intensity: previous research

Most of the research on supervision intensity has tested *intensification* of supervision programs (ISPs) for serious (and less serious) offenders on probation and parole. Much less is known about the effects of *reduction* of the intensity of supervision of clients, regardless of their risk of any new crimes or very serious crimes. We have been unable to find any previous randomized trial (or well-controlled quasi-experimental study) that specifically tested the impact of reducing probation contacts and/or increasing officer caseload size on the recidivism of low-risk offenders. The only available evidence on our specific research question is a non-causal, cross-sectional study of county probation agencies showing that counties with higher probation caseloads in California have higher property crime rates (Worrall et al. 2004). That result, of course, could be entirely spurious in relation to funding as the underlying cause, with higher crime counties having greater financial strain and fewer probation officers per case.

A systematic review of the effects of probation intensity on recidivism (Gill 2010) has identified a considerable number of potentially-relevant evaluations from nearly 4,000 database hits. Almost all these studies are concerned with the effects of increasing, rather than reducing, the frequency of contacts and reducing officer caseloads—the hallmarks of ISP (Petersilia and Turner 1993). Overall, the evidence on the effectiveness of these programs at reducing recidivism of high-risk offenders is doubtful, given very few significant differences and a split in the direction of the effect (MacKenzie 2006: 311).

More relevant is the evidence that low-risk offenders fare worse under higher-intensity supervision models than under "standard" intensity of supervision (Erwin 1986; Hanley 2006). Erwin (1986) conducted a small evaluation of an ISP in Georgia, in which program participants were matched with regular probation clients on age, sex, race, crime type, risk score, and need score (risk and need were assessed by the Risk/Need Assessment instrument, the standard tool used in the jurisdiction). ISP participants were placed in caseloads of 25 per officer and were subject to 5 face-to-face contacts per week initially (eventually decreasing to 2) as well as mandatory curfew, employment, and community service, among other requirements. Analysis of recidivism outcomes was stratified by assessed risk level. Erwin found that low-risk cases had more recidivism under the more severe intervention than under the less severe treatment. Of those placed in intensive supervision, yet deemed to be low-risk, 42% were re-arrested, compared to 27% of the low-risk offenders in the regular probation group. Only those low-risk clients who were sentenced to incarceration were re-arrested at a higher rate (46%). However, these sample sizes were small; the re-arrest rate of 42% represents only 5 out of 12 low-risk clients in the intensive supervision group. In contrast, Erwin found that only the highest-risk group (which constituted a larger proportion of the sample) reduced their recidivism compared to regular probation after going through ISP.

Hanley (2006) found a similarly unfavorable, but non-significant, effect of intensive supervision for low-risk clients. The finding comes from her secondary analysis of data collected from areas across the US that were selected by the Bureau of Justice Assistance in 1986 to receive funding to explore the effects of ISP. The main effects of these programs are reported in Petersilia and Turner (1993). Hanley selected 1,100 cases from the dataset that were deemed to be low or high risk based on risk factors measured in the original study. She measured treatment "appropriateness" according to whether or not high-risk clients received more contacts and low-risk clients received fewer contacts. Controlling for program site and legal status (probationer or parolee), Hanley found that low-risk offenders who received risk-inappropriate services (i.e., intensive supervision) were more likely to fail than those left on standard supervision, although the relationship was non-significant. Conversely, high-risk offenders had lower re-arrest rates when they received more contacts.

There is little in the present evidence base to alter the predicted effects of the low-intensity supervision. While the direction of effect favors the predictions of defiance theory or deviant peer contagion, the unstable estimates make the findings far from definitive. No data clearly contradict either of the directions of effect posited by the competing theories. Instead, we find ample basis for a claim to equipoise, or the morally balanced position of not knowing which of the two choices is better for the

community and the convicted offenders (Federal Judicial Center 1981). It is in this moral position that we were confident in the ethical basis for conducting a randomized controlled trial.

4 Forecasting and sample eligibility

4.1 Forecasting low risk of serious crime

In order to identify offenders who would be at very low risk of committing a "serious offense"—defined as murder, attempted murder, aggravated assault, robbery, or sex crimes—the APPD asked the University of Pennsylvania's Richard Berk to modify his existing random forests model, which was (at that time) designed only to forecast homicide or attempted homicide (Berk et al. 2009). 'Low risk' in the present study is defined as a *forecast* of no charges for serious crimes within two years of the probation or parole case start date (regardless of date of random assignment). Berk et al. (2009) provide a detailed description of the basic statistical techniques used in this revised model to forecast risk. The basic method is to use prior criminal record and other baseline data for each offender to forecast risk at the beginning of each APPD supervision, based on the recent two-year outcomes of APPD clients with similar characteristics. Each individual forecast is a unique calculation based on a model that was developed with one large sample and tested with an independent sample.

Briefly, random forests methods were applied to a dataset containing all new probation or parole cases which began between 2002 and 2004, with information about their background characteristics and post-probation arrest outcomes of the offenders. The predictors used in the model reflected only information routinely available at probation intake. Using the existence of any charges for a serious offense as the dependent variable, the model tabulated false positives and false negatives in order to yield the possibility of a higher-than-low-risk offender being identified mistakenly as low risk—the 'worst case scenario'—of approximately 5%. The model was tested on an independent sample, and then used to identify, by design, the lowest-risk 61% of the entire APPD caseload. This stratification, and the false positive and false negative rates associated with it, were acceptable to APPD from an operational perspective. The APPD then implemented an experimental design in which those offenders forecasted to be low risk were randomly assigned to either 'supervision as usual' (about 150 cases per officer) or to the experimental group on a low-intensity supervision caseload with a goal of a 400:1 offender-to-officer ratio. (The actually delivered ratio during the experiment was an average of 323:1.)

4.2 Screening for eligibility

Once the revised crime risk forecasting model was completed in midsummer 2007, the APPD decided to locate the randomized trial in the regional supervision units for the West and Northeast areas of Philadelphia. "Regional supervision", at that time, was the default supervision for any offender who was not assigned by court mandate

to a specialized unit, such as sex offenders, drunk driving, gun court, and others. While these special units supervise over half of all offenders, there is little administrative discretion located within the APPD in how to supervise them. The purpose of the experiment was to test an alternative to the regional supervision model, for which caseloads per officer had grown steadily higher for decades.

The following discussion describes how the "pipeline" of eligible cases was screened in order to yield the final "batch" sample of cases eligible (or almost) for being "dropped down" the model and then subjected to random assignment to regional versus low-intensity supervision. The discussion tracks the numbers presented in the CONSORT diagram in Fig. 1.

The CONSORT diagram is recommended by the Consolidated Statement on Reporting of Trials (www.consort-statement.org). The CONSORT Group was founded by health care researchers to undertake initiatives "to alleviate the problems arising from inadequate reporting of randomized controlled trials (RCTs)." These reporting issues affect RCTs in other fields as well, and could be remedied in experimental criminology by wider use of the kind of diagram presented in Fig. 1. This diagram increases transparency about both internal and external validity of an experiment. It shows the reader the exact populations from which sample cases are drawn, and how they were removed from, or remained under, the experiment as conducted.

On 27 July 2007, a total of 12,233 probation cases, involving 7,830 different offenders, were under active supervision in the West and Northeast regions. Data on these cases were extracted from the APPD's case management database and then linked to court history data from the Court of Common Pleas Case Management System. In order for their cases to be eligible for a low-risk assessment, an offender had to have a valid local police ID number that appeared in both databases. A total of 431 offenders did not have a valid ID that could be linked to court processing data. Therefore, the low-risk model made a risk assessment for 95% of the active cases, or a total of 7,399 offenders. These offenders were then "dropped down" the forecasting model one at a time, in order to yield an individualized risk forecast.

The random forest model gives each risk assessment what is essentially a vote count, indicating what proportion of the model's numerous decision trees classified the case as low risk (Berk et al. 2009). Only cases where more than 50% of the votes forecasted an absence of serious offending were ultimately classified as low risk and eligible for random assignment. A total of 576 offenders (or less than 8%) could not be classified due to missing predictor values. A further 3,964 offenders' assessments (or 54%) received less than the required number of votes to be forecast as "low risk".

On 24 August 2007, the remaining 2,859 offenders were screened for exclusionary criteria prior to random assignment. Not all offenders classified as low risk were eligible for inclusion in the low-intensity experiment. Ineligible offenders included those whose cases were scheduled to end within 60 days of the start of the experiment, offenders who were court-ordered to a specialized unit, offenders supervised in an existing low-risk caseload[2], or who were in potential

---

[2] APPD already had a low-risk caseload outside the regional units before the experiment was implemented. However, assignment to the caseload was based on a different risk tool that predicted arrest for any new offense. Offenders assigned to this caseload had reduced reporting requirements, so had already experienced supervision levels similar to those being tested in the experiment.
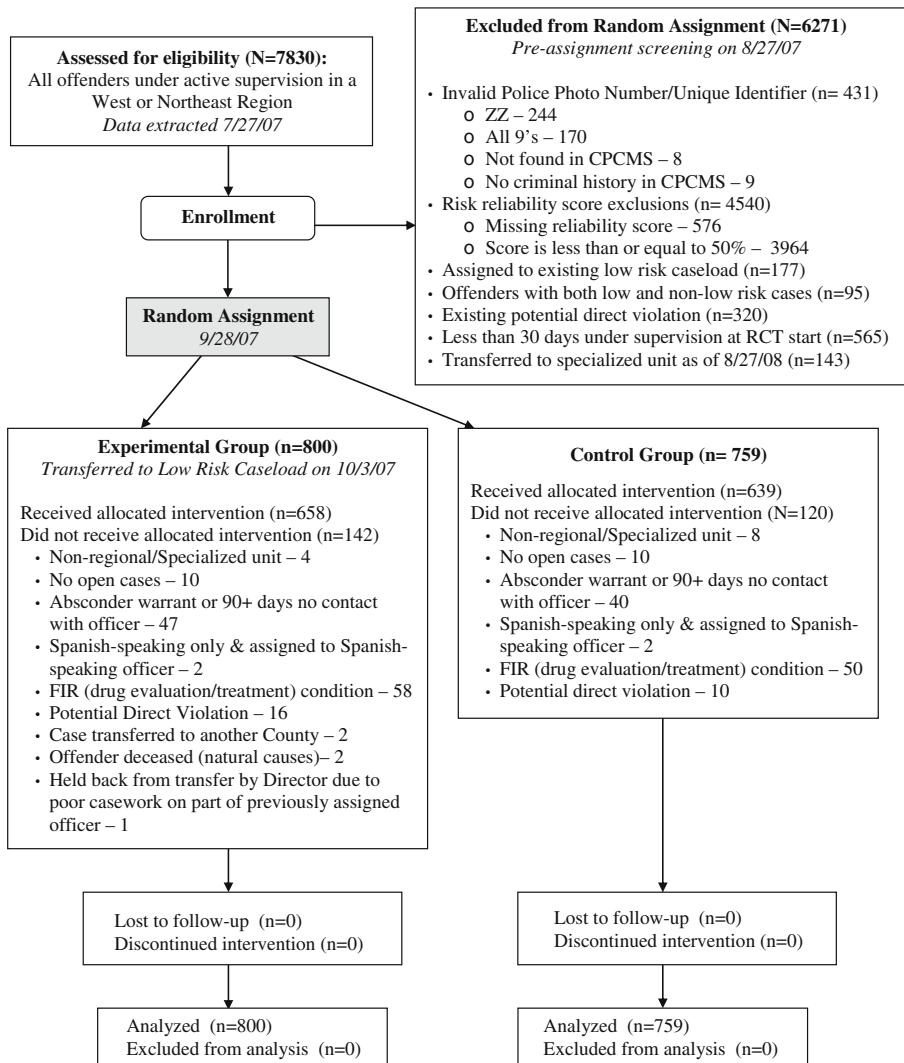
Assessed for eligibility (N=7830):
All offenders under active supervision in a West or Northeast Region
Data extracted 7/27/07

Excluded from Random Assignment (N=6271)
Pre-assignment screening on 8/27/07

- Invalid Police Photo Number/Unique Identifier (n= 431)
  - ZZ – 244
  - All 9's – 170
  - Not found in CPCMS – 8
  - No criminal history in CPCMS – 9
- Risk reliability score exclusions (n= 4540)
  - Missing reliability score – 576
  - Score is less than or equal to 50% – 3964
- Assigned to existing low risk caseload (n=177)
- Offenders with both low and non-low risk cases (n=95)
- Existing potential direct violation (n=320)
- Less than 30 days under supervision at RCT start (n=565)
- Transferred to specialized unit as of 8/27/08 (n=143)

Enrollment

Random Assignment
9/28/07

Experimental Group (n=800)
Transferred to Low Risk Caseload on 10/3/07

Received allocated intervention (n=658)
Did not receive allocated intervention (n=142)
- Non-regional/Specialized unit – 4
- No open cases – 10
- Absconder warrant or 90+ days no contact with officer – 47
- Spanish-speaking only & assigned to Spanish-speaking officer – 2
- FIR (drug evaluation/treatment) condition – 58
- Potential Direct Violation – 16
- Case transferred to another County – 2
- Offender deceased (natural causes)– 2
- Held back from transfer by Director due to poor casework on part of previously assigned officer – 1

Control Group (n= 759)

Received allocated intervention (n=639)
Did not receive allocated intervention (N=120)
- Non-regional/Specialized unit – 8
- No open cases – 10
- Absconder warrant or 90+ days no contact with officer – 40
- Spanish-speaking only & assigned to Spanish-speaking officer – 2
- FIR (drug evaluation/treatment) condition – 50
- Potential direct violation – 10

Lost to follow-up (n=0)
Discontinued intervention (n=0)

Lost to follow-up (n=0)
Discontinued intervention (n=0)

Analyzed (n=800)
Excluded from analysis (n=0)

Analyzed (n=759)
Excluded from analysis (n=0)

Fig. 1 CONSORT diagram of the Philadelphia low-intensity community supervision experiment

direct violation of their sentence due to an arrest occurring after the start of their supervision. Offenders with any known disqualifying condition were excluded from the sample prior to random assignment.

A total of 1,084 low-risk offenders were thus excluded from random assignment due to disqualifying condition(s) found in the eligibility screening prior to random assignment. The majority of these losses (669 offenders) were due to cases that were scheduled to end within 60 days of the start of the experiment. Further exclusions prior to random assignment featured 177 offenders supervised in existing low-risk units, and 143 offenders who had been transferred to a specialized unit after data were extracted but before random assignment occurred. Finally, 95 offenders were excluded from random assignment due to having multiple simultaneous sentences

(of different starting dates) for which the model forecast both low-risk and non-low-risk outcomes.

Shortly before the experimental start date, APPD staff realized that clients would also have to be excluded if they had 'potential direct violations.' These were clients who had potentially violated their sentence by being arrested, and the case was in the process of being returned to court to deal with the new charge. Given the increased size of their caseload, dealing with the court procedures would be too much work for the low-risk officers, and these cases therefore could not be included in the experimental caseload. An additional pre-random assignment screening removed 320 of these cases. This reduced our total sample size (experimental plus control) to 1,455, which was too small to satisfy the constraint of maintaining the two low-risk officers' caseloads at 400 cases each. To increase sample size, we relaxed the 60 days remaining on probation requirement to 30 days. The reduction in time remaining allowed us to retain an additional 104 cases from the original 1,775, giving us an acceptable sample size of 1,559.

5 Random assignment and attrition

A random assignment sequence was applied to the surviving 1,559 cases on 28 September 2007, and the experiment officially began on 1 October 2007. In the West APPD region, 400 clients were assigned to the low-risk caseload and 401 to control. In the Northeast APPD region, 400 clients were assigned to the low-risk caseload and 358 to control. Experimental group offenders were formally transferred from their existing probation officers to the low-risk officers during the first two weeks of October 2007. The 759 offenders selected for the control group remained in general "regional" supervision. The latter's probation/parole officer was not informed of the offender's inclusion in the experiment, nor of the offender's low-risk assessment. The experiment was thus a "single-blind" study, in which the experimental cases were identified to the APPD staff but the control cases were not.

As is often the case in randomized controlled trials, a non-trivial number (16.8%) of offenders included in random assignment were later determined to be ineligible for low-risk supervision based on a variety of disqualifying conditions. A total of 262 offenders (142 in the experimental group, 120 in the control group) were determined, post-random assignment, to have been ineligible for the experiment by the time of implementation of the massive "batch" random assignment. Sixty of these offenders were declared ineligible based upon the same exclusion criteria that were applied during pre-assignment screening. The remaining 202 offenders became ineligible due to new and previously unanticipated conditions that made them difficult or impossible to assign to a low-intensity caseload.[3]

---

[3] One example of these unanticipated conditions was the court-ordered FIR (Forensic Intensive Recovery) program, a drug evaluation and treatment regime. Within weeks of the experiment's start date, APPD administrators decided that the intensive monitoring required for offenders in drug treatment was impossible to provide within the experimental officers' large caseloads. As a result, 108 offenders (58 experimental, 50 control) with FIR conditions became ineligible for low-intensity supervision.

5.1 Intention-to-treat analysis

Regardless of the reason for their disqualification, all the cases subject to treatment attrition after random assignment are retained in the present analysis, based on the treatment groups to which they were originally assigned. For the offenders who were assigned to the control group, any ineligibility for the RCT had little or no bearing on the way they were treated by their assigned probation officer, since the control officers were not aware who was and was not part of the experiment. The ineligible cases that were originally assigned to the experimental group, however, were either never transferred into the low-risk caseload, or were removed from it when their disqualifying condition was identified. The analysis that we present here, therefore, follows a standard intention-to-treat (ITT) protocol, despite the fact that a noteworthy proportion (17.8%) of the experimental offenders were not permitted to participate in their assigned treatment.

6 Baseline equivalence and sample demographics

Because random assignment gave each offender an equal probability of being assigned to the experimental low-intensity supervision, there is little concern that the two treatment groups are systematically different from one another. Nevertheless, an examination of the groups' baseline equivalence is important, both for demonstrating this lack of randomization bias, and for revealing the types of offenders who participated in the research.

Thirty-nine offender characteristics, measured at the start of RCT, have been tested to determine whether the two treatment groups were statistically indistinguishable from one another when the experiment began. The key tests are presented in Tables 1 and 2, with further tests in the Appendix. No systematic differences were found between the experimental and control groups. Only 2 of these 39 baseline equivalence tests found a statistically significant difference, which is almost exactly the 5% of significant results that one would expect by chance.

Table 1 shows the test results for the key personal demographic variables, in which there were no significant baseline differences between the treatment and

Table 1 Demographics at RCT start

	Experimental		Control		
	n	Mean	*n*	Mean	*p*
Male	799	.665	759	.676	.635
White	800	.428	759	.389	.119
African-American	800	.480	759	.484	.889
Age	799	40.797	759	40.576	.680
Active probation cases	800	1.604	759	1.639	.621
Minimum forecast score	800	.652	759	.650	.706

Table 2 Prevalence of offending and incarceration at any time prior to RCT start

	Experimental		Control		
	n	Mean	n	Mean	p
Prevalence of any prior charges	800	.996	759	1.000	.083
Prevalence of serious offenses	800	.294	759	.278	.492
Prevalence of violent offenses	800	.378	759	.361	.500
Prevalence of sexual offenses	800	.029	759	.018	.179
Prevalence of property offenses	800	.590	759	.622	.198
Prevalence of firearm offenses	800	.109	759	.125	.314
Prevalence of drug offenses	800	.653	759	.659	.795
Prevalence of local jail incarceration	800	.661	759	.651	.666

control groups. The sample as a whole was two-thirds male, and nearly half African-American. Their average age was 41 years. In comparison to all of the offenders under APPD supervision, this low-risk sample was more likely to be female (32.9%, compared to 23.0% in the full APPD standing caseload), more likely to be white (40.9%, compared to 33.7%), less likely to be African-American (48.2%, compared to 59.0%), and approximately 3 years older (40.7 years, compared to 37.4 years).

The average number of active probation or parole cases per offender at the start of the RCT was also the same in the two groups, at 1.621 cases each. One-third of the offenders in the sample had multiple cases open at APPD when the study began. Since a separate forecast was made for each case (as opposed to each offender), these offenders would have had multiple reliability scores generated by the random forest model. In these cases, the eligibility rules demanded that each offender's minimum voting score be in excess of 0.50. The forecasting model results are nearly identical for the two groups, which means that neither group was any more "low risk" than the other.

Table 2 shows what proportion of the offenders had been charged with various sorts of offenses, in the Philadelphia court system, prior to the start of the RCT. Both felonies and misdemeanor offenses are included in the data. Again, no significant differences existed between the two treatment groups, but the numbers do reveal a bit about the type of offenders involved in the research.

Nearly every offender in the sample[4] had a prior criminal record of some kind. Serious offending—which encompasses the same offenses (murder, attempted murder, aggravated assault, robbery, and sexual offenses) that the forecasting model was built around—was much more rare. Nevertheless, nearly 29% of the RCT offenders had some history of serious offending. Violent offending, which includes

---

[4] Two experimental offenders appear to have had their prior criminal records expunged from the court database, and now have no previous criminal history, despite the fact that both of them were on probation and were enrolled into the RCT.

some more minor charges that the "serious offense" category does not, was present in histories of 37% of the offenders. A prior record of sexual or gun crime was fairly rare, but a majority of the sample had at least one charge for property offenses, drug crimes, or both. Nearly two-thirds of the offenders had spent at least some amount of time incarcerated in the local jail system. All of this may illustrate the difference between a subjective and a statistical approach to risk classification.

A similar set of comparisons between the two treatment groups, based upon the raw number of prior charges for these same crimes, reached similar conclusions. On average, the offenders in the sample had 20.9 charges for offenses committed prior to the beginning of the experiment, and had spent a mean of 144.0 days in the custody of the local jails. Only one of these tests found a significant difference between the treatment and control groups; the experimental group had an average of 0.17 charges for prior sexual offenses, compared to just 0.07 charges in the control group (p=0.044).

The two groups also exhibited no significant differences in their mean ages at the time of first contact with the Philadelphia criminal courts. Taken as a whole, these 1,559 offenders were 29.1 years of age when they committed the offense that led to them being charged (as an adult) for the *first* time. This is roughly four years later than average age across the entire APPD caseload (25.2 years), and confirms a large body of evidence in criminology about the lower risks associated with late-onset offenders (e.g., Farrington et al. 2006).

7 Treatment protocols

The experiment was largely limited to the manipulation of caseloads, with probation officers retaining discretion in exactly how to apply the protocol guidelines in each group. This section explains how the supervision was structured. The next section then reports on our measures of what supervision elements were delivered in each group and to what degree.

7.1 Control group

Just under half of the offenders under APPD supervision are placed into "general supervision", meaning that they have no special court-ordered requirements for how their sentences are to be managed. All of the offenders in the RCT were selected from this population of probation clients, and those in the control group simply remained there after random assignment. Nearly 200 officers work in general supervision, with a targeted caseload of around 150–200 offenders per officer. Offenders in these caseloads are normally required to make monthly office visits to the centralized probation/parole office in downtown Philadelphia. Sentencing conditions, however, can shape the basic framework of supervision, often mandating drug testing, educational services, or financial sanction.

APPD officers in general supervision exercise considerable discretion in determining the substance and frequency of supervision. While it is the officer's responsibility to carry out the judicial orders, reporting frequency can be set, and

other sanctions applied, at the discretion of the supervising officer. Typically, officers increase the offenders' mandated reporting frequency in response to a new arrest or positive drug test, although such supervisory sanctions are determined on a case-by-case basis by the officer and unit supervisor. New arrests for certain dangerous offenses can require the officer to lodge a probation warrant against the offender, holding him or her in custody in spite of other bail or pre-trial release options.

All of these variable circumstances can result in a range of different experiences for different sorts of offenders. The exact requirements that an offender must meet during their period of supervision may vary from case to case, and may also change over time.

7.2 Experimental treatment group

Once offenders were assigned to the experimental treatment group, a number of things changed about the way their cases were managed by the APPD. The low-risk, low-intensity supervision for these offenders consisted of placement into a (fairly large) caseload of other low-risk offenders. The experimental supervision was delivered by two probation officers, one for the West region and one for the Northeast. The treatment protocol was designed so that each probation officer received and maintained a caseload of approximately 400 offenders. This number was chosen not only for statistical power, but also for operational reasons. With such a large number of clients, APPD administrators intended to constrain a low-risk officer's ability to invest a large amount of time in each case.

One challenge with the experimental treatment group was that the offenders who were randomly assigned to it necessarily had to depart as time went on. Some simply reached the end of their sentences, while others were transferred out of low-intensity supervision when they committed a new offense or failed to maintain contact with their probation officer. In order to maintain the experimental treatment caseload of approximately 400 offenders per officer, new low-risk offenders were identified at various times during the next year, and were used to keep the caseloads at the desired level. These "backfill" cases were not counted as part of either the experimental or control groups, but were necessary to maintain "real-world" conditions for purposes of external validity. The decline over time in the proportion of cases in the officers' caseloads that was actually included in the experimental sample was balanced by an increasing proportion made up of "backfill" offenders who were not included in the study sample.

At the offender's first contact with the low-risk officer, officers formally told offenders that they were in a low-risk caseload subject to the reduced reporting requirements described below. They were also told that they would be transferred back to standard supervision if they were rearrested for a new crime, and that an arrest warrant would be issued if they had no contact for six months. Both the offender and officer signed a statement of understanding of these conditions.

Low-risk offenders in the experimental caseloads were meant to receive a considerably reduced level of supervision compared to the standard model received

by those in general supervision. The low-risk supervision protocols, described by Ahlman and Kurtz (2009), included the following:

> ▪ Office reporting: Offenders were to have one scheduled office visit every six months. These contacts focused on a review of the offender's residence, employment, payments on fines/costs/restitution, and compliance with other conditions.
> ▪ Telephone reporting: Offenders were scheduled for one telephone report every six months, occurring approximately midway between office visits. These contacts focused on confirmation of details described above. The clients were not restricted from initiating additional telephone contact.
> ▪ Drug testing: Only administered if required by court order. Probation officers were instructed to order a drug evaluation after no more than three positive urine tests, and were free to refer the offender to drug treatment if the offender requested it.
> ▪ Missed contacts: Arrest warrants were issued if there was no contact with the offender for six months. If the offender surrendered voluntarily, the warrant could be removed with no criminal sanction.

Low-risk officers were not responsible for the case once a new arrest occurred, but they were still expected to handle technical violations (e.g., missed contacts, noncompliance with judicial conditions) that did not result in arrest or a warrant. Thus, low-risk clients who committed a new offense were essentially dropped from the experimental treatment program, and—for the remainder of their time under APPD supervision—received the same treatment as the control group. For the ITT analysis, however, they remained in the experimental group for measurement purposes.

8 Treatment as delivered

Figure 2 shows how all of the 1,559 offenders in the RCT were supervised by APPD for one full year before and after the start of the experiment. Some of these offenders had been on probation, under general supervision, for a year or more before the RCT began, while others had only recently begun serving their sentences. The manner in which the cases were selected and screened, as described above, required that offenders were scheduled to be under APPD supervision between late July through early November 2007. That is therefore the one period of time when nearly all of the offenders in both treatment groups were in the active caseload of a probation officer.

Because the sample was pulled from APPD's standing caseload, all members of the experimental group experienced at least some time in general supervision prior to the RCT start date. The experiment is therefore most generalizable to changes to the amount and intensity of supervision that are applied to offenders who have *already had at least two months experience* with the higher supervision levels of the control treatment beforehand.[5]

---

[5] That fact makes this experiment most useful in the short run, when the APPD's caseload is in transition for a gradual shift of existing cases from OSFA to risk-based treatment unique to that risk level. The present experiment is perhaps less valid as an assessment of differences from the initiation of probation or parole sentence, as will become the case in the long run.
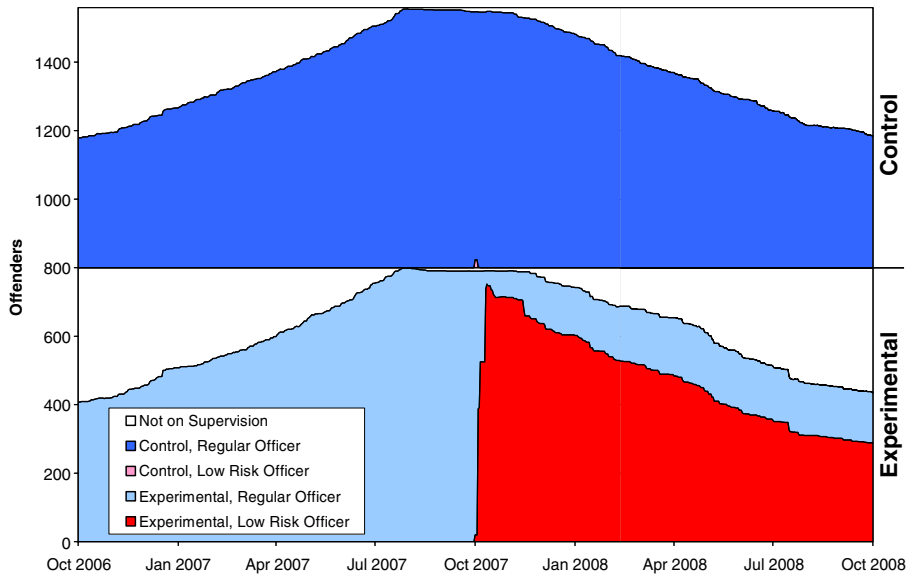
Fig. 2 Treatment as delivered to experimental and control offenders for one year before and one year after RCT start

Beginning in early October 2007, the vast majority of offenders in the experimental group were moved into the caseload of one of the two low-intensity probation officers. At the same time, the small number of control group offenders who had been previously assigned to one of these officers were transferred to another officer in the general supervision pool. Within two weeks, all the offenders who could be transferred were successfully placed into their respective treatment groups. Shortly thereafter, a number of experimental offenders were found to be ineligible for low-intensity treatment (see above), and were quickly transferred back into general supervision. By mid-November 2007, these transfers out of the experimental group were complete, and that group reached a certain degree of stability.

From November 2007 on, the sample in both groups began to decrease naturally, as offenders reached the end of their sentences and departed from APPD supervision entirely. In the experimental group, meanwhile, some offenders began to get arrested for new offenses or have warrants sworn out against them, and were therefore returned to general supervision, where they essentially began to receive the control treatment. What Fig. 2 shows is the proportion of offenders in the sample who were actually receiving the different possible treatment conditions at any given time, regardless of the reasons for the delivered treatment status. The lower half depicts the cases randomly assigned to low-intensity supervision; the upper half depicts the cases assigned to the control.

Table 3 breaks down the two treatment groups by the amount of time the offenders were assigned to one of the two low-risk probation officers after the RCT began. The 24 (out of 759) control group offenders who were assigned to these officers were all transferred elsewhere within a few days, never met with these officers after the experiment began, and were never informed that they had been assessed as low risk. In essence, therefore, none of the control group offenders

Table 3 Days of low-risk treatment delivered

	Experimental		Control			
	n	Mean	*n*	Mean	*P*	
Days assigned to low-risk P/PO	800	215.595	759	.094	.000	*
1 or more days of low risk	800	.945	759	.032	.000	*
30 or more days of low risk	800	.891	759	.000	.000	*
60 or more days of low risk	800	.790	759	.000	.000	*
90 or more days of low risk	800	.750	759	.000	.000	*
180 or more days of low risk	800	.605	759	.000	.000	*
270 or more days of low risk	800	.444	759	.000	.000	*
Full year assigned to low risk	800	.361	759	.000	.000	*

*Statistically significant

actually experienced any of experimental treatment, despite the fact that a small number of them were assigned to these officers for a few days.

In the experimental treatment group, the average offender experienced more than 7 months of low-risk treatment by the end of the first year. Nearly 95% of them were assigned to a low-risk officer for at least one day, even if some of them were soon found to ineligible and transferred back to general supervision. Nearly 90% of the experimental group received at least 30 days of low-intensity supervision, and 75% stayed with their assigned treatment for at least 90 days. By the end of the first year, however, sentence completions, new arrests, and absconder warrants had combined to leave just 36% of the experimental group still receiving their assigned treatment.

The treatment experiences depicted in both Fig. 2 and Table 3 require some further explication. First, we should stress that all offenders in the sample had the identical period of outcome measurement (1 October 2007 through 30 September 2008), regardless of how long they remained in their randomly assigned treatment. Second, in both groups, the offenders could only experience their assigned treatment for as long as their sentences lasted, so that both sides of the equation examined the persisting effects of the initial differences in treatment. Third, of the experimental offenders, only 44 (5.4%) were never transferred into a low-risk caseload. A more substantial 177 (22.2%) of the ITT experimental offenders were later moved back into general supervision after spending some time in low-intensity treatment for reasons of policy—which makes the experiment generalizable to the way the APPD operates. More than a quarter of the experimental-group offenders spent some amount of time experiencing the control treatment after the experiment began, many because they violated their conditions of (low-risk) probation. Only slightly more than a third of those initially assigned experienced the experimental treatment for the full year of the RCT, again because many of them reached the end of their sentence—but were still tracked for re-arrest outcomes. The delivery of the low-risk treatment, therefore, was uneven by design, with the dosages received by the experimental offenders necessarily variable.

Table 4 shows the average caseload sizes experienced by the offenders in the two treatment groups, on the 15th day of each month, for 12 months before and after the

Table 4 Average supervising officer caseload sizes per offender on active supervision, October 2006 through September 2008 (snapshot values on the 15th of each month)

	Experimental		Control			
	n	Mean	n	Mean	P	
October 2006	382	144.3	362	148.5	.363	
November 2006	405	145.8	390	147.8	.669	
December 2006	452	145.6	422	149.0	.403	
January 2007	486	150.3	464	152.9	.502	
February 2007	527	152.6	506	154.1	.695	
March 2007	562	151.6	539	152.5	.784	
April 2007	606	151.7	577	148.8	.382	
May 2007	657	149.6	621	145.4	.182	
June 2007	722	151.7	681	147.2	.172	
July 2007	773	150.0	720	147.3	.357	
August 2007	787	152.9	748	149.4	.238	
September 2007	774	156.9	739	152.7	.184	
October 2007 (RCT Begins)	772	357.1	726	138.0	.000	*
November 2007	764	355.5	701	141.7	.000	*
December 2007	722	334.3	655	144.7	.000	*
January 2008	680	316.9	618	144.6	.000	*
February 2008	651	336.8	571	144.3	.000	*
March 2008	629	322.2	543	145.5	.000	*
April 2008	603	329.1	513	146.2	.000	*
May 2008	539	293.5	469	147.2	.000	*
June 2008	503	331.0	453	144.9	.000	*
July 2008	479	316.8	415	147.3	.000	*
August 2008	430	300.8	382	149.8	.000	*
September 2008	416	281.6	370	149.9	.000	*
Average caseload during RCT		322.97		134.51		

*Statistically significant

start of the experiment. These values represent the average across all of the offenders who were under active supervision on the appropriate date, including (after the RCT began) those experimental offenders who had returned to general supervision. In order to provide a more understandable metric, offenders who had absconded from supervision, and whose cases had been handed off to one of the "Wanted Card" caseloads at APPD, are excluded from these figures. The "Wanted Card" caseloads can easily number into the thousands of offenders, and including them would skew the results upwards, overestimating the number of cases handled by the APPD's officers. APPD itself did not count such cases in the caseloads. By APPD definition, the average caseload offenders experienced during the RCT was 323 active clients for experimental officers and 135 for control officers, or 140% higher caseloads for the experimentals.

Table 4 shows that, prior to the start of the experiment, there was virtually no difference in the caseload size of the average offender's probation officer. For the entire year that the experiment was in operation, however, the average caseloads more than doubled for offenders in the experimental group. The experimental offenders had far more competition for their supervising officers' attention than did the offenders in the control group.

Table 5 shows the average number of contacts that the offenders' assigned officers had (or attempted to have) with them, both before and after random assignment. The values are annualized based on the number of days that the offenders were under active[6] supervision, and thus represent how many contacts an average offender would have experienced during a full year of supervised time. During the year prior to the start of the experiment, there was no significant difference between the two treatment groups.

Once the RCT began, the number of contacts experienced by the experimental group was reduced by approximately 45%, while the control group had almost exactly the same amount of contact as they had encountered during the previous year. While the experimental protocol had aimed to compare *monthly* office visits in the control group to *twice-yearly* visits under low-intensity supervision (i.e., a six-fold decrease in visits), the delivered difference was much smaller (two-fold). The control offenders—along with the offenders in both groups during the year prior to random assignment—simply did not meet with their assigned probation officers on anything close to the policy of a once-per-month schedule. Yet after random assignment, the two groups were clearly subjected to different numbers of contacts with their probation officers. The experimental treatment appears to have been delivered as designed, even if the dosage differential was not as strong as anticipated.

9 Post-assignment offending

Recidivism was measured using court records, tracking any charges for offenses which took place after the start of the experiment. The available data were limited to offenses which were dealt with by the Philadelphia courts. Most offenses which took place outside the city limits are therefore excluded from our analysis.

There is no evidence that either group was more criminally active than the other during the year after the start of the experiment. Table 6 shows the *prevalence* of one or more new criminal charges for any offenses committed within this one-year period. There were no significant differences, in any offense category, between the experimental and control groups. In addition, neither group proved more likely to end up incarcerated in the local jail system than the other.

Table 7 examines the *frequency* of offending during the one-year observation period. No significant differences in the frequency are observed between those assigned to low-intensity supervision and those assigned to the control group. If anything, the control group seems to have been slightly more criminally active than those on low-

---

[6] As before, "active supervision" excludes any time when the offender had absconded from supervision and had been placed into one of the "Wanted Card" caseloads. The same pattern of results, however, is found when this "Wanted Card" time is included in the calculations.

Table 5 Annualized contacts between offenders and probation officers, October 2006 through September 2008

	Experimental		Control			
	n	Mean	*N*	Mean	*p*	
One year prior to RCT start						
Attempted contacts	799	16.95	758	16.37	.268	
Successful contacts	799	11.04	758	10.50	.091	
Successful in-person office contacts	799	4.52	758	4.63	.633	
One year after RCT start						
Attempted contacts	787	9.25	743	16.82	.000	*
Successful contacts	787	5.50	743	10.00	.000	*
Successful in-person office contacts	787	2.43	743	4.52	.000	*

*Statistically significant

intensity supervision, although the effect is small and the difference is not significant. The multiple comparisons are not independent, so it would be misleading to estimate the probability of most of them favoring the experimental group by chance.

Both Tables 6 and 7 also portray *seriousness* of offending. There are no significant differences in the prevalence or frequency of serious offending, although the difference for prevalence is close to significant. The effect is very small, however, with a difference of 1.6% versus 3% of offenders charged with a murder, rape, robbery, or aggravated assault. The fact that the control group had the higher rate of serious offenses maintains the consistent finding of no evidence that the experimental group had any increased risk.

With no significant differences found in either the prevalence, frequency, or seriousness of post-assignment offending, there remains one other way in which the two treatment groups could be different from one another. Figure 3 shows a time to failure graph, showing when members of each group committed their first post-assignment offense. There is no significant difference here either (*p*=.649). The first

Table 6 Prevalence of offending and incarceration for one year after RCT start

	Experimental		Control		
	n	Mean	*n*	Mean	*p*
Prevalence of any new charges	800	.160	759	.150	.593
Prevalence of new serious offenses	800	.016	759	.030	.067
Prevalence of new violent offenses	800	.025	759	.040	.106
Prevalence of new sexual offenses	800	.001	759	.000	.318
Prevalence of new property offenses	800	.055	759	.067	.316
Prevalence of new firearm offenses	800	.005	759	.011	.215
Prevalence of new drug offenses	800	.064	759	.063	.967
Prevalence of new jail incarceration	800	.150	759	.165	.426

**Table 7** Frequency of offending and incarceration for one year after RCT start

	Experimental		Control		
	n	Mean	*n*	Mean	*p*
Charges for any kind of new offenses	800	.874	759	1.055	.366
Charges for new serious offenses	800	.073	759	.100	.548
Charges for new violent offenses	800	.145	759	.244	.198
Charges for new sexual offenses	800	.031	759	.000	.318
Charges for new property offenses	800	.245	759	.267	.760
Charges for new firearm offenses	800	.033	759	.061	.389
Charges for new drug offenses	800	.158	759	.198	.354
New entries into local jail system	800	.188	759	.204	.509
Days incarcerated in local jail system	800	15.824	759	16.195	.891

six months of the experiment show a very slightly (and non-significant) increased prevalence of offending in the experimental group as compared to the control offenders, while the second six months show the two groups essentially tied throughout the remainder of the year.

Even when later offending, which occurred after the first anniversary of the experiment's start date, is considered, the results are the same. There is simply no evidence that reducing the intensity of supervision had any effect on the subsequent criminal behavior of these low-risk offenders.

10 Statistical power

These null findings for post-assignment offending largely conform to what the research team had expected to find when the experiment was first conceived. But
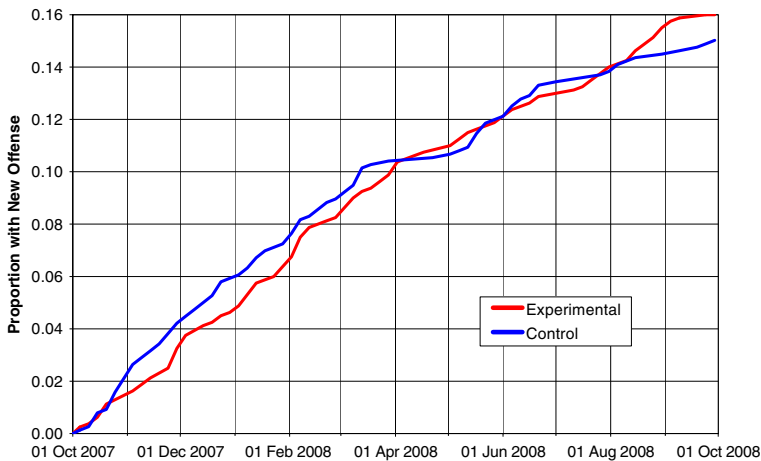
Fig. 3 Time to the first new offense after RCT start, October 2007–September 2008

null findings are also fairly easy to produce, simply by using an insufficient sample size, selecting an excessively heterogeneous sample, or otherwise creating an underpowered research design. We took particular pains, therefore, to ensure that the measures presented here would have enough statistical power to detect as significant any reasonably-sized differences between the two treatment groups.

Cohen (1988) defines a system of standardized effect sizes that can be used when determining the power of a statistical test, in which an effect size of 0.80 is described as "large", 0.50 is referred to as "moderate", and 0.20 is depicted as "small". Based on these definitions, our sample of 1,559 offenders is exceptionally powerful, with a probability of 0.976 to find a significant difference even when the effects of the experimental treatment are "small". For any test in which the full sample could be employed, a power of 0.80—which is the usual standard applied when developing a research design—is achieved with effect sizes of just 0.142. All the differences we report between groups were therefore below that less-than-"small" level.

Another way of presenting the power of these analyses is to estimate how large a sample would have been needed to detect the differences found above as statistically significant. In order to reach a statistical power of 0.80, the frequency comparison for any post-assignment offending would have required us to identify and randomly assign a sample of 14,843 offenders. The test for violent offenses would have required a smaller, but arguably still over-powered, sample of 7,362 different offenders. For the most modest differences, such as the comparison of days spent incarcerated in the local jail system, a sample of more than 600,000 offenders would have been required to produce the desired level of statistical power. Clearly, any of these samples would have been so vastly over-powered as to find virtually any difference, no matter how meaningless, as statistically significant.

Instead, the sample size used here seems to have been almost ideal to detect any important differences, while ignoring any trivial ones. Thus, the fact that no significant differences were found in post-assignment offending can be regarded, insofar as any null finding can, as a valid conclusion, and not merely the artifact of an underpowered research design.

11 Conclusion

This experiment constitutes a weak test of theory, but a stronger test of policy. That disjunction is made possible by its empirical capacity to *predict* without necessarily being able to *explain*. While the external validity of any policy test should be enhanced by clearer theoretical implications, there are many effective policies that lack clear specification of such mechanisms. Big effects of policies, particularly those established in RCT designs, have been widely replicated without a clear theoretical basis—simply on the strength of the evidence. Hot spots police patrols (Braga and Weisburd 2010), for example, have substantial evidence consistent with deterrence doctrine, but the precise theoretical mechanisms by which intensified police patrols cause less crime cannot yet be well specified. This

limitation can be readily acknowledged. But it makes the policy implications no less compelling.

Similarly, this experiment provides a reasonably unbiased estimate of the differences in crime between a low and even lower level of dosage in probation supervision. As a policy experiment, it constitutes strong evidence that agencies with low frequency of visits (i.e., 4.5 visits annually) can safely cut that frequency roughly in half (to 2.4 visits annually), at least for low-risk offenders identified by the kind of tools used in Philadelphia. Whether this finding is generalizable can only become knowable by further testing. But this evidence provides strong justification for doing just that—more testing. Community corrections agencies around the world now have a firm basis for undertaking their own replications of our design, using these results to defend against criticism that even an experiment would be unsafe.

The interpretation of this experiment differs, however, in important respects from one with a "big effects" result. Our evidence does not show big effects. Rather, it shows the potential for big cost reductions, which can be an equally compelling factor in public policy. The difference between big effects and cost reductions, from a hypothesis testing standpoint, is that we have failed to reject the null hypothesis, rather than rejecting it by a large margin. Important as the cost-reduction implications may be, we do not claim to have "proven a negative" or accepted the null hypothesis (Weisburd et al. 2003). Rather, we have used conventional power levels (or better) to attempt to detect a discernible difference. Whether greater power would have revealed such a difference, we cannot say.

This is a problem that often confronts public health research. Do cell phones cause cancer? Do high-voltage power lines? Or dental X-rays? The low base rates of these very serious events make it very difficult to discern causation. They make it equally difficult to reject causation. Similarly, we cannot say in a scientific sense that reduced frequency of probation is truly "safe." All we can say is that, in this research design, there is no evidence that it is unsafe.

The experiment has other limitations. By focusing on "dosage" of supervision visits to the probation office, it may have overlooked important differences in the nature of the experience that offenders received under the two randomly assigned conditions. The fact that only two officers delivered the low-intensity treatment leaves the experiment somewhat vulnerable to a rival interpretation: that the result depended on their personalities, which may have counteracted the effect of a dosage cut. Under this interpretation, the findings would not generalize to similar dosage cuts with officers of different personalities. The plausibility of that interpretation is mitigated, however, by several factors. One is that the supervision meetings tend to be fairly brief (15 minutes or so) and highly structured (following a checklist of questions about current residence, employment, relationships, etc). The offenders spend far more time in travel and waiting at the APPD office than they do in actually talking with their probation/parole officers. It is hard to imagine just how a personality difference could causally override a contrary effect of the dosage difference. But we must at least note the possibility.

Moreover, we do not have exact measures of how much time the officers spent with the offenders in the two conditions. If, for example, the low-intensity offenders

received longer visits, or had more referrals, or other unobserved dimensions of service, that might have counteracted the reduction in number of visits. Yet given the enormous caseloads the experimental officers took on, that possibility seems unlikely.

The clearest limitation is the treatment attrition of 17.8% of the offenders assigned to the experimental low-intensity dosage. Because these offenders were declared ineligible after random assignment, they received the standard (control) dosage rather than their assigned dosage. It is therefore possible that their inclusion in treatment as delivered could have altered the result. It is not possible, however, to tell from the evidence whether the result would have shown more offending in the lower or higher supervision intensity conditions.

As limited as this policy conclusion may be, it must still be recognized as a substantial step forward from previous evidence on reduced intensity of community supervision. At the least, it contradicts previous evidence from trend analysis that suggested reduced supervision would reduce recidivism (Wilson et al. 2007). While we have some non-significant indications of that result in our frequency data, we cannot say that reduced supervision for low-risk probationers "works" to reduce recidivism. What it does "work" to do is to cut costs for low-risk probationers, freeing up resources for higher-risk cases.

As a weak test of theory, this experiment suggests that predictions from all three relevant theories were wrong. This is only a weak test by design, since we did not have the resources to construct the more elaborate design needed to specify different theoretical conditions. This means that one or more of the theories may still be correct, under different test conditions. Perhaps the differences in dosage, for example, were not large enough to meet the thresholds required by these theories. Or perhaps the theories only applied to certain subgroups within the low-risk population, even though that population is itself a subgroup of the larger APPD clientele.

The test of deviant peer contagion is especially problematic, given the average age of the sample at 41 years. Deviant peer contagion theory may be most appropriate for younger offenders. But since youth is the strongest predictor of high risk status for APPD cases (Berk et al. 2009), it is likely that low-risk cases will be older offenders. This may mean that a more appropriate test of deviant peer contagion would be an *increase* in supervision among high-risk cases, which are likely to be younger offenders.

Whatever other designs might have shown about these theories, this design failed to find any statistically significant confirmation of any theoretical prediction of a difference between low-intensity and higher-intensity supervision, regardless of direction. Put another way, having a mean of 2.4 office contacts a year versus 4.5 yields virtually identical results in criminal behavior. Whether the same conclusion would be reached by comparing such bigger differences as 12 versus 2 contacts annually, however, is a very different question. Thus, it is important not to draw too broad a conclusion about the effects of *any* differences in supervision intensity. We cannot generalize from this particular result to reduced supervisions at all levels of dosage. The external validity of this finding is clearly limited to the levls of dosage we studied.

Theoretically, there could also be much more harmful effects of higher-intensity supervision for low-risk offenders. Based on evidence from the ISP experiments (Hanley 2006), for example, as well as the non-significant differences in this experiment, much higher level of supervision for low-risk offenders might well raise their offending rate even more when compared to the counterfactual of only 2.4 contacts per year for the low-intensity supervision group. It could thus arguably be unethical to raise contact levels sufficiently to perform such an experiment, as it surely would be in Philadelphia, given the present results.

As a test of policy, this experiment shows that cutting the specified delivered supervision levels by 58% is not a discernibly unsafe policy for APPD to adopt. How much lower the APPD could go could depend in part on considerations of retribution in sentencing as well as crime prevention. This experiment has no evidence to offer on public perceptions of just deserts, but it is to be expected that public officials must consider other factors besides crime prevention.

In sum, we believe the experiment supports the following conclusions:

1. This experiment provides no evidence to support a claim that among probationers and parolees identified as low-risk by our statistical model, a reduction in supervisory office visits from 4.5 to 2.4 annually caused any increase in the prevalence, frequency, seriousness, or speed of re-offending.
2. The absence of evidence that the reduction caused any harm can provide some assurance to policymakers that the 140% increase in caseload size for low-risk offenders is justifiable, and has some potential to reduce the costs of supervising offenders of this type. Moreover, re-allocations of personnel to higher-risk offenders may be in the public interest—if those personnel can yield greater effects in other tasks.
3. The experiment cannot reject the null hypothesis that there is no deterrent, defiant, or peer contagion effect of 4.5 visits annually over 2.4.

From a crime prevention standpoint, the findings supported a re-allocation of resources of probation contacts away from low-risk and toward high-risk offenders. Such a reallocation, however, still needs to be supported by evidence that more contact with probation can reduce offending, at least for higher-risk offenders. This experiment is therefore only a first step in testing a risk-based strategy in comparison to the predominant OSFA strategy. It answers the first question: can supervision be cut for low-risk offenders without increasing crime. That answer is yes.

This experiment does not, however, answer the remaining question: can a reallocation of probation officers away from low-risk offenders and into more supervision of high-risk offenders help to reduce crime? For that, a separate experiment is required, one that the Philadelphia APPD began in February 2010: one which tests a very different strategy that is unique to that risk level. Only with the results of both kinds of experiments can it be said that a risk-based strategy for crime prevention is also evidence-based.

Appendix

Table 8 Additional measures of baseline equivalence

	Experimental		Control		
	n	Mean	*n*	Mean	*p*
Offender residence at RCT start					
Within city limits	781	.864	746	.853	.512
Distance from city limits	781	8.441	746	5.542	.390
Zipcode population	781	43,927	746	43,819	.903
Zipcode households	781	18,323	746	18,510	.612
Zipcode persons per household	780	2.408	746	2.379	.029 *
Zipcode proportion African-American	780	.404	746	.437	.072
Zipcode household income	779	32,738	746	31,922	.171
Zipcode house value	779	66,614	746	65,492	.570
Frequency of prior offending, probation cases, and jail stays					
Charges for any kind of offenses	800	20.930	759	20.822	.935
Charges for serious offenses	800	1.014	759	.881	.237
Charges for violent offenses	800	2.713	759	2.455	.308
Charges for sexual offenses	800	.170	759	.074	.044 *
Charges for property offenses	800	8.696	759	8.802	.910
Charges for firearm offenses	800	.461	759	.503	.607
Charges for drug offenses	800	3.421	759	3.957	.054
Prior probation cases	800	2.661	759	2.736	.544
Entries into local jail system	800	1.764	759	1.730	.755
Days incarcrated in local jail system	800	140.116	759	148.038	.555
Frequency of prior offending, among only those with one or more priors					
Any kind of offense	798	29.499	759	28.675	.090
Serious offense	247	33.237	231	33.031	.832
Violent offense	315	33.233	294	32.448	.339
Sexual offense	24	31.206	14	27.808	.282
Property offense	480	29.950	489	29.272	.261
Firearm offenses	91	29.031	103	30.975	.166
Drug offenses	527	32.415	512	32.300	.856

*Statistically significant

References

Anderson, E. (1999). *Code of the street*. Boston: Norton.

Ahlman, L. C. & Kurtz, E. M. (2009). *The APPD Randomized Controlled Trial in Low Risk Supervision: The Effects of Low Risk Supervision on Rearrest*. Philadelphia: Adult Probation and Parole Department.

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

Braga, A., & Weisburd, D. (2010). *Policing problem places*. NY: Oxford University Press.

Coalition for Evidence-Based Policy (2010). Home Page. Downloaded February 19 from http://coalition4evidence.org/wordpress/.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale: Lawrence Earlbaum Associates.

Dodge, K. A., & Dishion, T. J. (2006). Deviant peer contagion in interventions and programs: an ecological framework for understanding influence mechanisms. In K. A. Dodge, T. J. Dishion, & J. E. Landsford (Eds.), *Deviant peer influences in programs for youth* (pp. 14–43). New York: Guilford Press.

Dodge, K. A., Dishion, T. J., & Lansford, J. E. (2006). *Deviant peer influences in programs for youth*. New York: Guilford Press.

Gill, C. E. (In press). "Intensity of probation supervision: A systematic review." Jerry Lee Center for Criminology, University of Pennsylvania.

Goffman, A. (2009). On the run: Wanted men in a Philadelphia Ghetto. *American Sociological Review, 74*, 339–357.

Erwin, B. S. (1986). Turning up the heat on probationers in Georgia. *Federal Probation, 50*, 17–24.

Farrington, D., Coid, J. W., Harnett, L., Jolliffe, D., Soteriou, N., Turner, R., et al. (2006). *Criminal careers and life success: new findings from the Cambridge Study in Delinquent Development*. London: Home Office Findings #281.

Federal Judicial Center. (1981). *Experimentation in the Law*. Washington: Federal Judicial Center, Administrative Office of the U.S. Courts.

Gibbs, J. D. (1975). *Crime, punishment and deterrence*. NY: Elsevier.

Hanley, D. (2006). Appropriate services: examining the case classification principle. *Journal of Offender Rehabilitation, 42*, 1–22.

Jacobson, M. (2005). *Downsizing prisons*. NY: NYU Press.

Lipsey, M. (2006). The effects of community based group treatment for delinquency: A meta-analytic search for cross-study generalizations. In K. A. Dodge, T. J. Dishion, & J. E. Landsford (Eds.), *Deviant peer influences in programs for youth* (pp. 162–184). New York: Guilford Press.

MacKenzie, D. (2006). *What works in corrections: Reducing the criminal activities of offenders and delinquents*. New York: Cambridge University Press.

O'Connell, M. E., Boat, T., Warner, K. E. (Eds.) (2009). *Preventing mental, emotional, and behavioral disorders among young people: Progress and possibilities*. Washington, D.C.: Committee on the Prevention of Mental Disorders and Substance Abuse Among Children, Youth and Young Adults: Research Advances and Promising Interventions. Institute of Medicine; National Research Council, National Academies Press.

Petersilia, J., & Turner, S. (1993). Intensive probation and parole. In M. Tonry (Ed.), *Crime and Justice: A Review of Research*, *17*, 281–335.

Pew Center on the States. (2009). *One in 31: The long reach of American corrections*. Washington: The Pew Charitable Trusts.

Rosch, J. (2006). *Deviant peer contagion: Findings from the Duke executive sessions on deviant peer contagion*. The Link 5, 1–17. Child Welfare League. Downloaded on May 4, 2009 from http://www.cwla.org/programs/juvenilejustice/thelink2006fall.pdf.

Sherman, L. W. (1993). Defiance, deterrence and irrelevance: A theory of the criminal sanction. *Journal of Research in Crime and Delinquency, 30*, 445–473.

Sherman, L. W. (2007). Use probation to prevent murder. *Criminology and Public Policy, 6*, 843–849.

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: U.S. Department of Justice.

Weisburd, D., Lum, C., & Yang, S. M. (2003). "When can we conclude that treatments or programs "Don't Work"? *Annals of the American Academy of Political and Social Science, 587*, 31–48.

Wilson, J. A., Naro, W., & Austin, J. F. (2007). *Innovations in probation: Assessing New York City's automated reporting system*. Washington: JFA Associates.

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.

Zimring, F. E., & Hawkins, G. (1973). *Deterrence: The legal threat in crime control*. Chicago: University of Chicago Press.

Geoffrey Barnes is Research Assistant Professor of Criminology at the Jerry Lee Center of Criminology, University of Pennsylvania.

Lindsay Ahlman is Research Associate in the Adult Probation and Parole Department of the First Judicial District of Pennsylvania.

Charlotte Gill is a doctoral candidate in criminology at the University of Pennsylvania and Managing Editor of the Campbell Collaboration Crime and Justice Steering Group.

Lawrence W. Sherman is Director of the Jerry Lee Center of Criminology at the University of Pennsylvania and Wolfson Professor of Criminology at the University of Cambridge.

Ellen Kurtz is Director of Research for the Adult Probation and Parole Department of the First Judicial District of Pennsylvania.

Robert Malvestuto is the Chief Probation and Parole Officer of the Adult Probation and Parole Department of the First Judicial District of Pennsylvania.