The effects of community-based supervision on juvenile recidivism: A meta-analysis of intensive supervision probation and aftercare/re-entry programs

by

Jessica April Bouchard

B.A., St. Thomas University, 2012

Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of
Master of Arts

in the
School of Criminology
Faculty of Arts and Social Sciences

SIMON FRASER UNIVERSITY

Fall 2015

Approval

Name:	**Jessica April Bouchard**
Degree:	**Master of Arts (Criminology)**
Title:	***The effects of community-based supervision on juvenile recidivism: A meta-analysis of intensive supervision probation and aftercare/re-entry programs***
Examining Committee:	**Chair:** Martin Andresen Professor

Jennifer Wong
Senior Supervisor
Assistant Professor

William Glackman
Supervisor
Associate Professor

Jodi Viljoen
External Examiner
Associate Professor
Department of Psychology
Simon Fraser University

Date Defended/Approved: December 03, 2015

Abstract

Community correctional sentences are administered to more juvenile offenders in North America than any other judicial sentence (Hockenberry & Puzzanchera, 2014; Torbet, 1996). Particularly prominent in juvenile corrections is intensive supervision probation and aftercare/re-entry, yet the effect of these supervision-oriented interventions on recidivism is mixed. The purpose of this meta-analytic study was to determine the effect of intensive supervision probation and aftercare/re-entry on juvenile recidivism. An extensive search of the literature and the application of strict inclusion criteria resulted in the selection of 27 studies that contributed 55 individual effect sizes. Studies were pooled based on intervention type (intensive supervision probation or aftercare/re-entry) and outcome measure (alleged offenses or convicted offenses). The pooled analyses yielded contradictory results with respect to outcome measure; in both cases, supervision had a positive effect on alleged offenses and negatively impacted convicted offenses. Implications of this pattern across intervention type and outcome measure, as well as recommendations for future research are discussed.

Keywords: aftercare; re-entry; intensive supervision probation; juvenile; meta-analysis; recidivism

Dedicated to all who have encouraged me to chase my dreams

Acknowledgements

This thesis would not be complete without acknowledging a few individuals who have played a significant role in my journey through grad school and to whom no number of “thank yous” will ever be enough.

To my family, thank you for supporting my 5,000+ km move across the country to pursue this degree and for continuously supporting me despite my inability to properly answer the question “so, what exactly are you doing?” for most of it.

Thank you to my New Brunswick friends who have graciously stayed in touch over the past four years, and for always welcoming me home with excitement and a backlog of stories. To my Vancouver friends, thank you for standing by my side from beginning to end, and for the endless support and encouragement. But of above all, thanks for the memories, it’s been a pleasure!

Heather, Nicole, and Rahul, thank you for making early morning classes and late-night assignments (and everything in between) a little more bearable. Melissa, Alicia, and Sebastian, you three have contributed to my grad school experience in immeasurable ways. Because of your friendships, I am where I am today.

Thank you to Dr. Bill Glackman for introducing me to the world of program evaluation and for instilling much of the confidence and curiosity I have today. I would also like to thank Dr. Jodi Viljoen for contributing her valuable time and insight to the ever-important final touches of this thesis.

Last but not least, endless thanks to Dr. Jennifer Wong for taking me under her wing and teaching me everything I know about meta-analysis. Jennifer, over the last 3 years you have allowed me to learn and grow, and have provided me with numerous opportunities and invaluable experiences that will have major impacts on my future endeavours, and for that, I will always be grateful.

Table of Contents

List of Tables

List of Figures

Chapter 1. Background

1.1. Review of the literature: Community-based interventions and intermediate sanctions

1.1.1. The emergence of community-based interventions and intermediate sanctions

Deciding what to do with those who disobey the law is a complex but central issue in correctional research (Cullen & Gendreau, 2000). Since Martinson's infamous report concluded that when it comes to rehabilitating offenders "[w]ith few and isolated exceptions, the rehabilitative efforts that have been reported so far have had no appreciable effect on recidivism", the impact of correctional programs has been highly debated (Martinson, 1974, p. 25). The conclusions drawn from Martinson's report altered existing correctional practices and fuelled an unprecedented shift in attitude and ideology (Cullen & Gendreau, 2001).

In essence, Martinson's conclusions refuted that rehabilitative practices were successful at correcting criminal behaviour (which was later reduced to 'nothing works' in rehabilitating offenders) (Cullen & Gendreau, 2000). Social and political platforms exacerbated the 'nothing works' allegations by calling attention to increasing crime rates. Meanwhile, concern for public safety intensified as questions arose concerning the capacity of the criminal justice system to correct criminal behaviour; creating a compelling argument to adopt an alternate (i.e., more punitive) approach to dealing with offenders (Lemert, 1993; Pratt, Franklin, & Gau, 2011). Furthermore, the declaration of 'war on crime' triggered a close investigation into the practices of the criminal justice system, which subsequently unveiled strong political opinions that the system was inherently lenient and flawed. The heightened scepticism and concern, coupled with questions relating to the legitimacy of the criminal justice system, served as a catalyst for 'get tough' attitudes and policies. These conclusions led to new sentencing guidelines

that focused on crime control and a “desire for retribution and just deserts”, whereby offenders would serve longer and determinate sentences (Pratt et al., 2011, p. 81). In short, it did not take much time, or effort, for this ideology to be accepted as factual (Pratt et al., 2011).

However, this drastic change in ideology did not go unchallenged (Pratt et al., 2011). Martinson’s conclusions were met with widespread criticism from academics who challenged the notion that ‘nothings works’ (Cullen & Gendreau, 2000; Gendreau, 1981; Gendreau & Ross, 1979; Palmer, 1975). In response to this drastic shift in ideology and in an attempt to salvage the reputation of treatment interventions and revive rehabilitative efforts, a proliferation of evidence emerged with an aim to dispel the claim that treatment could not rehabilitate offenders (Cullen & Gendreau, 2000; Gendreau & Ross, 1979; Gendreau & Ross, 1987). The emergence of new data demonstrated favourable effects of treatment interventions (Cullen & Gendreau, 2000) and that treatment programs could indeed have a positive impact on recidivism (Gendreau, 1981). Although the new evidence was encouraging, there was no denying that there was some truth to Martinson’s conclusions; not all interventions were effective, and existing effects were small (Cullen & Gendreau, 2000). Although Martinson later retracted his claims that 'nothing works' (Cullen & Gendreau, 2000), the statement was influential and had a lasting impact on future correctional practices and ideologies.

The socio-political context and the sequential declaration of the ‘war on drugs’ and the ‘war on crime’ resulted in an unprecedented growth and overcrowding in the American prison system (Shelden, 2004). Within a period of 10 years (1975-1985) incarceration rates in the United States nearly doubled (Wakefield & Uggen, 2010). Due to this vast influx of individuals entering the criminal justice system, the existing resources could not keep up with the demand and were quickly exhausted (Cullen & Gendreau, 2000; Langan, 1994; Petersilia, 1999). As a response to the huge increase in incarcerated offenders, community-based correctional programs emerged as a potential alternative to the punitive approach. This new option held the promise of alleviating logistical and economic pressures while holding offenders accountable for their actions and maintaining public safety (Petersilia, 1999; Redondo, Sanchez-Meca & Garrido, 1999).

Newly emerging evidence demonstrated promise that diverting offenders could decrease recidivism and community-based diversion programs soon gained credence as a viable alternative to formal processing (Cullen & Gendreau, 2001). While evaluative efforts transitioned to a focus on interventions that took place outside of secure detention facilities, 'tough' sanctions remained at the forefront of the public agenda. As such, research on the effectiveness of community-based supervision programs manifested as an area of widespread attention. Although the implementation of community corrections would require a fundamental re-thinking of correctional sentencing, community correctional sentences promised harsh penalties while avoiding further exhaustion of correctional funds (Byrne, 1990). As it stood, the only available options were either probation or incarceration; probation was considered too lenient for offenders, while, due to limited capacity, incarceration was reserved for severe offenders. In showing that rehabilitation did, in fact, decrease recidivism and that tough on crime policies would not be sustainable, it was undeniable that 'tough' community alternatives (what is known today as 'intermediate sanctions') were needed (Clear & Hardyman, 1990)

Three sequential conditions are said to have contributed to the emergence of intermediate sanctions: 1) prison overcrowding, 2) overuse of probation (and subsequent perception of correctional failure), and 3) a lack of sentencing choices (Baird, 1990, Byrne, 1990; Caputo, 2004; Clear & Hardyman, 1990; Petersilia, 1999; Taxman, 2002).

1) Prison overcrowding: In the 1980s, the prison population increased dramatically over a short period (Clear & Hardyman, 1990; Wakefield & Uggen, 2010). While the prison population continued to grow, systemic resources were not expanded, leading to system overcrowding that eventually exhausted available resources (Byrne, 1990; Caputo, 2004; Kellum, 2006). While building additional facilities to adequately house the growing population of offenders was an apparent response, the endeavour would be costly (Petersilia & Turner, 1993). This predicament led policymakers to turn to the community and consider the possibility of community-based initiatives as viable options to relieve the pressures in areas that the criminal justice system was lacking (Byrne, 1990).

2) Problems with probation: As the prison system became increasingly overcrowded and ran out of space and resources, the community was deemed an alternative setting worthy of experimentation. As such, to alleviate prison overcrowding, more serious offenders were sentenced to probation. The probation population continued to grow and diversify, and probation officers could not keep up with the heightened demand. Furthermore, the increase in the number of high-risk offenders in the community (Corbett, 1999; Lemert, 1993, Torbet, 1996) required more intensive supervision and programming than could be allocated (Caputo, 2004; Petersilia, 1999). As concern for public safety was widespread, perhaps unsurprisingly, probation began to lose credibility as a viable alternative to incarceration as recidivism rates for probationers increased.

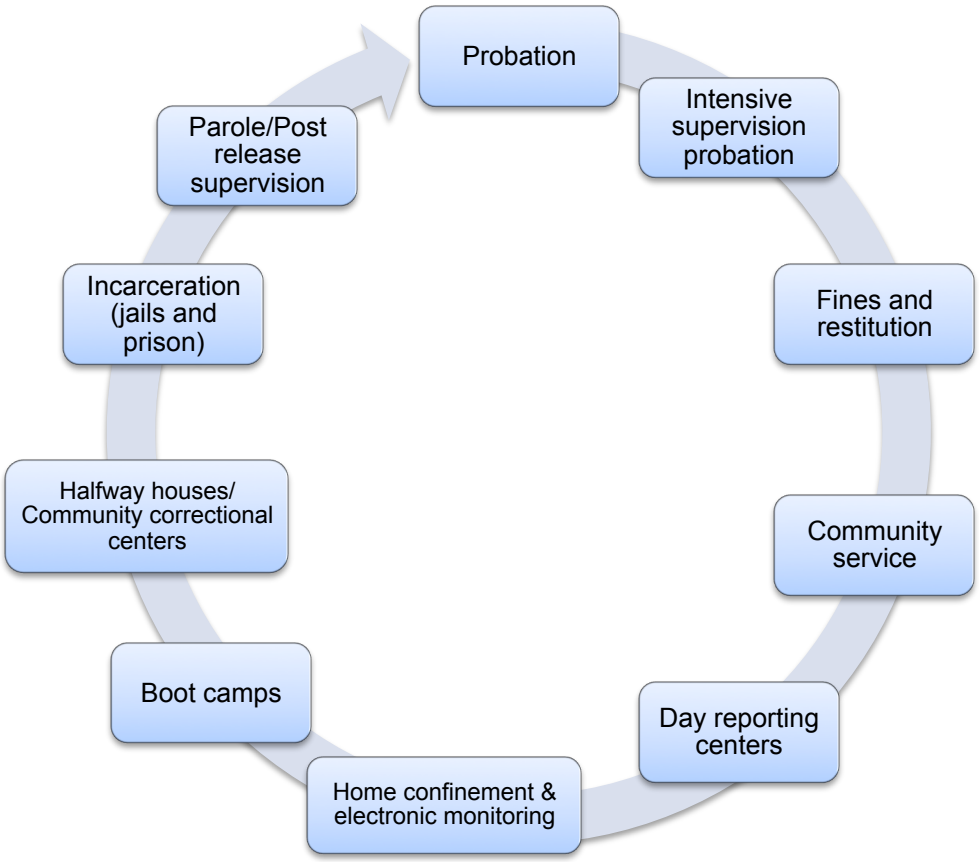
3) Lack of sentencing choices: Prior to the 1980s, there were very few viable sentencing options other than probation and incarceration. This lack of selection resulted in a polarizing choice between ‘soft’ and ‘tough’ alternatives (Caputo, 2004; Petersilia, 1999). Traditionally, probation was reserved for less serious offenders (whose crimes were not serious enough to warrant incapacitation) while incarceration was reserved for more serious offenders (Caputo, 2004; Clouser, 1996). However, the newly adopted ‘get tough’ policies and attitudes quickly caused an influx of offenders to be processed through the criminal justice system. Consequently, the criminal justice system could not keep up with the demand, and increasingly dangerous offenders were being sentenced to the community as opposed to institutions. Inevitably, probation grew to be interpreted by the public as too lenient. In short, a sentencing option that struck a balance between 'too harsh' and 'too lenient' was required; an option that fell in between probation and incarceration for youth whose criminal activity was severe enough to warrant attention and intervention but was not severe enough for incarceration (Clouser, 1996; Petersilia, 1999). As such, intermediate sanctions offered an appealing solution to all of these issues; it would alleviate pressures on jails and prisons and would provide a continuum of sentencing choices (Armstrong, 1990).

1.1.2. Intermediate sanctions as an alternative community-based correctional intervention

Correctional interventions are commonly classified among three categories: incarceration, community corrections, and intermediate sanctions (Caputo, 2004). While incarceration refers to incapacitation in prisons, jails, or in a residential facility, community corrections refer to interventions, including supervision such as probation, which take place in the community. Intermediate sanctions refer to a continuum of correctional options that fall in between traditional probation and incarceration, and range in level of supervision, control, severity, and punitiveness (Caputo, 2004; Petersilia, 1999).

As intermediate sanctions were designed to offer a range of correctional options, a spectrum of interventions was developed to reflect the range of severity in crimes and risk levels of offenders. Caputo (2004) highlights eight principal forms of intermediate sanctions that fall along the continuum: intensive supervision programs, monetary penalties/restitution (including fines), community service, day reporting centers, home confinement (including electronic monitoring), boot camps, day halfway houses, and aftercare/post-release supervision. See Figure 1 for the continuum of correctional options that are classified as intermediate sanctions, as well as a brief definition of each type of sanction.

Figure 1: Continuum of intermediate sanction interventions



Probation: Traditional community-based supervision (by a probation officer) where offenders comply with certain conditions.

Intensive supervision probation: Enhanced form of probation that typically involves smaller caseloads, increased contacts, and greater monitoring and/or increased services.

Fines and restitution: Primarily sanction-based; can include monetary fines as compensation for harm caused to the victim (Caputo, 2004).

Community service: A court ordered sanction that requires an offender to perform unpaid work in place of an alternate sanction (i.e., avoid financial restitution) (Caputo, 2004).

Day reporting centers: A structured supervision-oriented sanction that resembles traditional probation, however, it "[c]ombines high levels of control with intensive delivery of services. They require offenders to report to a specific location on a routine, pre-arranged basis, usually daily, where they participate in structured activities such as counselling and job training" (Caputo, 2004, p. 11).

Home confinement & electronic monitoring: Specialized supervision that is either home-based or the offender is tracked by GPS. Offenders must comply to certain conditions such as a curfew, or must remain in a pre-established proximity to their home. Offenders may also be required to remain "in their homes for a specified number of hours per day or week [and] are permitted to leave for approved activities such as employment and religious services" (Caputo, 2004, p. 12).

Boot camps: Structured/militaristic residential program that generally focus on discipline.

Halfway houses/Community correctional centers: Commonly used as a transitional phase for offenders post-incarceration, these sanctions are "[c]ommunity-based, minimum security residential facilities that provide offenders and released inmates with housing, some treatment services, and access to community resources for employment and education" (Caputo, 2004, p. 12).

Incarceration (jails and prison): Confinement in an institution.

Parole/post-release supervision (i.e., aftercare/re-entry): Community-based intervention in which offenders receive supervision after being released from a closed institutional setting.

Over the years, extensive research has been conducted on the effectiveness of intermediate sanctions in preventing and deterring criminal activity. Some research on intermediate sanctions is conclusive: for example, military-style boot camps and Scared Straight programs are not effective in reducing recidivism (Petrosino, Turpin-Petrosino, & Buehler, 2004; Wilson, MacKenzie, & Mitchell, 2005). As such, the presence of these correctional practices has diminished substantially. Conversely, intensive supervision probation (ISP) programs and aftercare/re-entry programs are still prevalent in today's criminal justice system. Still, there is relatively little determinate evidence on the effectiveness of these approaches. This thesis seeks to expand upon the existing evidence to determine the effect of intensive supervision probation and aftercare/re-entry on juvenile recidivism.

1.1.2.1. Aftercare/Re-entry

Aftercare/re-entry programs were developed through the theoretical lens that increased supervision and monitoring will lead to a decrease in criminal activity. The goal of aftercare/re-entry is to prevent recidivism among offenders who are released

from custodial institutions (Kellum, 2006). Put simply, aftercare/re-entry programs aim to help offenders transition from the prison setting to the community setting (Fairfull, 2013). Through transitional guidance and community reintegration, these interventions seek to provide "supervision, services, and supports [that are] designed to safeguard rehabilitative gains made in placement while guiding the [offender's] safe transition back to an ordinary life in the community" (Griffin, 2004, p. 2). Generally speaking, aftercare/re-entry programs include supervision as well as any service that is deemed to assist the successful transition and reintegration of prisoners from custody to the community (Petersilia, 2004).

The term 'aftercare' suggests that the priorities of the intervention lie in the 'after' portion of a custodial sentence. However, the importance of the process beginning while the offender is in custody cannot be understated (Griffin, 2004). Aftercare/re-entry is not just a post-institutional supervision program (Griffin, 2004). Petersilia (2004) highlights the importance of each successive step throughout the re-entry process. That is, pre-planning for re-entry upon admission, the supports the youth receives throughout the transition period, as well as the supervision and support that extend beyond initial reintegration. In short, experts suggest that from the point of admission, every step taken should be intentional to prepare the offender for successful reintegration upon release (Petersilia, 2004).

Various authors have noted that for aftercare/re-entry interventions to be effective in reducing recidivism they must contain the following characteristics:

First, re-entry planning should take place immediately upon commitment to the institution (Byrnes, Macallair, & Shorter, 2002). If preliminary intervention and release planning can commence prior to release, the offender is more likely to be successful upon re-entry (Travis & Petersilia, 2001). Second, the initial phase of transition should be highly structured and intensive (Griffin, 2004; Kellum, 2006). Research has suggested that the period immediately following release is when offenders are most likely to relapse into criminal/deviant behaviour (known as the "danger time") (Griffin, 2004, p.5). Moreover this period is when supervision should be the most intense and when resources are most necessary (Griffin, 2004; Soloman, 2008). The intensive

structure of the program (highly concentrated resources and reporting schedule) should gradually become less intensive over time (Griffin, 2004; Soloman, 2008). Third, it is suggested that the re-entry process should be a collaborative effort between various agencies and community partners, and should be a continuum of care wherein services expand across settings (custody and community) (Griffin, 2004; Kellum, 2006; Petersilia, 2004; Soloman, 2008; Travis & Petersilia, 2001). Furthermore, aftercare/re-entry programs should link professionals from a variety of areas of expertise. As the barriers that offenders face upon re-entry are numerous, attending to risk factors (such as education, housing, employability, substance abuse, and mental illness) through the provision of services will optimize the likelihood of a successful transition into the community (Byrnes et al., 2002; Soloman, 2008).

1.1.2.2. Intensive Supervision Probation (ISP)

Intensive Supervision Probation (ISP) is a generic term for an intervention that is anything but generic (Petersilia & Turner, 1993); ISP programs can be highly diverse in design and overall goals. ISP programs are the most commonly used intermediate sanction (Caputo, 2004). As a 'tough' community alternative, these programs highlight increased intensity, control and supervision in comparison to traditional probation (Taxman, 2002). Although ISP programs are diverse, they are characterized by three primary features: 1) small caseloads, 2) intensive supervision, and 3) strict conditions (Petersilia & Turner, 1993; Tonry, 1996).

1) Small caseloads: The development of intensive supervision probation programs was based on the assumption that the number of clients in each caseload would be indicative of the intensity of supervision each probation officer could provide to each youth (Clear & Hardyman, 1990). Intensive probation programs typically have caseloads of 20 to 30 probationers (in comparison to traditional probation that can have caseloads of over 100 probationers) (Caputo, 2004). Reducing caseloads allows probation officers to better focus on their clients, interact more frequently and directly, build rapport, and have higher quality relationships (Taxman, 2002).

2) Intensive supervision: A second feature of ISP programs is the intensive supervision component. First, it is not uncommon for ISP programs to be characterized

by a specialized supervision team (probation officers, police officers, case managers, social workers, community service providers) who work collaboratively to provide intensive surveillance (Caputo, 2004). Second, this team provides access/referral to community-based services, wherein the offender can learn valuable skills (vocational, academic, interpersonal, and intrapersonal) that may help them steer clear of future criminal activity (Petersilia, 1999; Petersilia & Turner, 1993; Taxman, 2002).

3) Strict conditions: In addition to the two features discussed above, offenders participating in intensive supervision probation must adhere to strict conditions such as curfews, frequent and unannounced visits, drug testing, restitution and/or community service (Caputo, 2004). Together, these features of ISP programs result in a harsher and more punitive intervention than traditional probation (Tonry, p. 84).

In sum, intensive supervision probation programs are an enhanced/intensive form of standard probation (Caputo, 2004; Clouser, 1996). In theory, intensive supervision probation operates on similar principles as traditional/regular supervision with similar core principal intervention components (supervision/monitoring and services/treatment). The difference lies in the intensity of the practice, i.e., more frequent supervision contacts and service delivery, service referrals, and service brokering through a specialized assessment (Clear & Hardyman, 1990). The increased monitoring, coupled with the small caseloads (more frequent contacts through smaller caseloads) along with strict conditions of compliance, will provide offenders with less opportunity to commit crime, and serve as a specific deterrent to criminal activity.

Although the terminology 'intensive supervision probation' does not imply the incorporation of additional rehabilitative practices, services and treatment are also considered important components of ISP programs (Armstrong, 1990). Moreover, it has been suggested that surveillance-oriented interventions will only produce beneficial effects if they are paired with rehabilitative intervention (Altschuler, Armstrong, & MacKenzie, 1999; Byrne, 1990, Petersilia, 1999; Stone & Fulton, 1995; Taxman, 2002). Thus, while ISP is surveillance-oriented, surveillance is not always the sole focus of the intervention (Byrne, 1990).

1.1.3. Criticisms of aftercare/re-entry and intensive supervision probation

As discussed above, intermediate sanctions were implemented with the expectation that the 'tough' interventions would decrease the prison population and reduce recidivism. However, although many offenders were diverted to community-based interventions, the incarcerated population continued to grow (Merrington, 2006). This can ultimately be attributed to three reasons: 1) faulty theory, 2) high rates of failure, and 3) net widening.

1) Faulty Theory

The main criticism of intermediate sanctions in general and intensive supervision probation programs in particular is that they lead to increased recidivism. That is, heightened supervision (offenders being watched more closely) may lead to increased detection of criminal activity (a higher likelihood of being caught) and subsequently result in increased (official) rates of recidivism (Caputo, 2004; Petersilia, 1999; Tonry, 1990). ISP programs were developed on the theoretical reasoning that smaller caseloads and more frequent contacts would lead to more intensive supervision. In turn, the intensive supervision would serve as a specific deterrent for youth, and result in a reduction in recidivism (Baird, 1990; Bennett, 1988). As such, the theory that 'more supervision will be better' may not be sound (Clear & Hardyman, 1990).

2) High rates of failure

Intensive supervision is typically accompanied by strict conditions and strict enforcement (Caputo, 2004; Clear & Hardyman, 1990). Hence, in addition to the watchfulness associated with intense monitoring, the vigilance related to strict conditions may lead to more technical violations (Caputo, 2004; Clear & Hardymnan, 1990; Petersilia & Turner, 1993). It is important to note that low levels of compliance may be indicative of unrealistic demands, implying that failure is more attributable to strict enforcement than criminal activity (Caputo, 2004; Merrington, 2006). If the point regarding increased technical violations is valid, then ISP programs altogether may be working against their goal of reducing recidivism.

3) Net-widening

Net-widening refers to programs expanding beyond the population the intervention intended to target and placing offenders "in more restrictive sanctions than their offense warrant[s]" (Caputo, 2004, p. 52; Merrington, 2006). Expanding on the issue of high rates of failure, Caputo (2004) notes that

> "[w]hen offenders who are not in need of imprisonment and who would be adequately supervised on regular probation or parole are placed into intermediate sanctions, technical program failures become much more concerning. When they fail as a result of rule violations, their punishments are typically more severe and when these punishments involve incarceration, intermediate sanctions works against attempts to reduce correctional populations and costs" (Caputo, 2004, p. 189).

1.1.4. The evidence for community-based supervision interventions on juvenile recidivism

The following section highlights findings from the most recent meta-analyses of community-based supervision programs for juveniles. Notably, there has been a vast increase in the number of studies evaluating aftercare/re-entry and ISP evaluations. Furthermore, there is little consistency with respect to positive or negative treatment effects, and, generally, researchers have not found statistically significant effects.

1.1.4.1. Aftercare/re-entry

Weaver and Campbell (2015) assessed the effects of 30 independent aftercare programs on the recidivism of juvenile offenders. The results demonstrated a small positive effect, however, the effect was non-significant (RR=0.931, p=0.117). Risk ratios ranged from 0.391 to 2.095, showing some variability in the effects of the selected interventions. Strong program implementation, randomized designs, as well as peer-reviewed sources, were significant moderators. Additionally, James, Stams, Asscher, De Roo, & van der Laan (2013) examined the effects of 22 independent aftercare/re-entry programs on recidivism for juveniles and young adult offenders. The analysis demonstrated a small, positive, and significant effect, meaning that in comparison to the control group (care as usual), youth who participated in an aftercare program post-

release from a secured facility were significantly less likely to recidivate (d=0.12, p<0.001).

Recent meta-analyses on the effects of aftercare programs for youth have demonstrated conflicting results. The studies above show that there is variability in effect sizes across studies; some programs are effective at reducing recidivism, while some are not. Notably, although Weaver and Campbell (2015) and James et al. (2013) both assessed the effect of aftercare programs on juvenile recidivism, the studies resulted in conflicting overall findings. The differences between the studies primarily lie in the operationalization of 'aftercare' interventions and the inclusion criteria. One difference between the two studies that is worth noting is that Weaver and Campbell (2015) included programs in which the intervention served as aftercare to shock incarceration programs as well as boot camp programs while James et al. (2013) did not include such interventions. Furthermore, James et al. (2013) only included aftercare programs that "incorporated a treatment modality such as skills training, counseling, and cognitive behavioral therapy" and excluded studies where the intervention focused on discipline or surveillance, such as certain intensive supervision probation programs (p. 265). Although the premise of Weaver and Campbell (2015) and James et al. (2013) are the same, the differing operationalization of 'aftercare' and differing inclusion criteria results in slightly different sets of studies, which subsequently resulted in contrasting findings and conclusions towards the effectiveness of aftercare programs.

1.1.4.2. Intensive supervision probation

Studies with negative findings

Sarver, Molloy, and Butters (2012) conducted a meta-analysis of the effects of intensive supervision probation/parole for juvenile recidivism. The literature was searched from 1987-2011 and resulted in 19 studies. In 13 of these studies the comparison group was standard probation, in five studies outcomes for intensive supervision programs were compared to outcomes for youth in custody, and in one study two different intensive supervision programs were compared. The analysis of the complete set of studies produced an odds ratio of 0.85 (p<.05), suggesting that youth who participate in intensive supervision probation/parole programs are significantly more

likely to recidivate than the youth in the comparison group. Moreover, the meta-analysis that specifically analyzed the 13 studies in which the comparison group was standard probation resulted in an odds ratio of 0.88 ($p<.01$), suggesting, again, that youth participating in intensive supervision probation are significantly more likely to commit a new offense in comparison to youth on standard probation.

Studies with positive findings

A meta-analysis conducted by Aos et al. (2001) revealed inconclusive evidence for the effects of intensive probation and parole programs for youth; intensive probation programs that were an alternative to incarceration resulted in a 0% change when compared to youth receiving traditional probation. Farrington and Welsh (2005) conducted a meta-analysis on community experiments, specifically, intensive supervision and aftercare programs. The results demonstrated a positive effect for intensive supervision programs in reducing recidivism (0.02) however this effect was not statistically significant. Finally, Lipsey (2009) meta-analyzed 548 independent study samples of juvenile offenders. The study samples were categorized into one of the following intervention approaches: no supervision, diversion, probation/parole, and institutionalization. Although the number of studies falling within each of the above-mentioned categories was not specified, the probation/parole category yielded a slightly positive, but non-significant standardized regression coefficient of 0.01.

In sum, recent meta-analyses on the effects of aftercare programs and intensive supervision programs for youth have demonstrated inconsistent results with respect to effectiveness of reducing juvenile recidivism. The studies above show that there is variability in effect sizes across studies; some programs are effective at reducing recidivism, while some are not. Notably, two recent meta-analyses on aftercare programs differed in operationalization of the intervention as well as selection criteria, which resulted in differing samples and findings. Finally, the most recent literature on intensive supervision probation programs demonstrates mixed findings; positive findings are small and not statistically significant, while negative findings are supported by statistically significant results. Given the inconsistent findings between studies, it is important to be aware of operationalization and selection criteria, and draw conclusions accordingly.

1.2. Review of the literature: The role of meta-analysis in evaluation

1.2.1. Why use meta-analysis to synthesize research?

There exists a variety of techniques to synthesize and summarize research, each with their own advantages and disadvantages. First, researchers may use a qualitative approach to summarize findings from multiple studies. One common qualitative technique to synthesize research is the narrative review. As the name suggests, researchers narratively summarize a body of literature to draw conclusions. As there is no systematic technique to summarizing the literature and drawing conclusions, of primary concern with this method of synthesizing research is the subjectivity of the review. In contrast, the use of quantitative techniques to synthesize research increases objectivity. One quantitative method used to synthesize research is called 'vote counting'. In vote counting, researchers draw conclusions based on significance (p-value). What Card (2011) refers to as "majority rules criteria" (p. 7), the researcher counts how many studies in their sample produced a significant positive finding, how many studies produced a statistically significant negative finding, and how many studies produced non-significant findings. Once all of the 'votes' are counted (i.e., the significance of each individual study is accounted for), the researcher draws conclusions in favour of whichever side received the most counts. Although this technique is systematic and objective, it does not take into account the magnitude (size) of the effect (Wilson, 2001).

A meta-analysis is a quantitative technique used to summarize research (Wilson, 2001). By applying statistical techniques that allow for the synthesis and analysis of multiple study outcomes (effect sizes), researchers can calculate an overall pooled estimate (Card, 2011, Lipsey & Wilson, 2001a). By using the effect size, the research is taking into account both the magnitude and direction of an effect, not just whether the result was statistically significant. This pooled estimate represents the average effect across studies, from which researchers can draw conclusions about a specific body of literature (Card, 2011). The processes associated with a systematic review of literature (systematic searching, a priori criteria, transparent decision-making) combined with the

steps and statistical techniques involved with a meta-analysis (described further in chapter 3) contribute to maximize objectivity and minimize bias (Lipsey & Wilson, 2001a).

1.2.2. The advantage of using meta-analysis in evaluation research

Wells (2009) advocates for the use of meta-analytic procedures to summarize research when attempting to navigate through a large body of criminal justice literature that is fraught with inconclusive and inconsistent evidence. As discussed above, the results produced by individual studies might be negligible, mixed, or biased from methodologically weak designs (discussed further below and in Chapter 3), causing confusion that renders it difficult to compare studies and their outcomes. However, when analyzed together through meta-analysis, information can be standardized, strengthened through numbers, and conclusions can be drawn more confidently (Garfield, 1991).

Lipsey (2002) outlines six lessons that can be learned from drawing conclusions about meta-analytic syntheses of evaluation studies, four of which pertain to the importance of methodological variables. Points 1, 2, and 3 (below) focus on methodological limitations of individual studies that are more robust when synthesized through meta-analysis.

> 1) many social programs are more effective than generally realized, 2) individual evaluations can easily produce erroneous results, 3) the methods used in an evaluation play a large role in the program effects found in the evaluation (Lipsey, 2002, p. 194).

One of the benefits of meta-analysis that can be taken from these cautions about methodology is that individual studies generally focus on statistical significance and derive the importance of their findings from this single feature (the previously favoured method of 'vote counting'). On the other hand, meta-analysis focuses on the magnitude of the effect of each individual study rather than on the significance of each individual study. Lipsey closes with an encompassing lesson that there is "safety in numbers" (Lipsey, 2002, p. 207) and that "the most credible evidence about program effects comes through integration of multiple evaluation studies" (Lipsey, 2002, p. 194). That is, various factors involved in the evaluation of programs can cause misleading results in

any single evaluation study and we must be cautious with interpretation. However, when analyzed together, although meta-analysis cannot improve the quality of these individual studies, the power in numbers accounts for some of the methodological weaknesses that are misleading when interpreting results and improving credibility.

To this point, one of the benefits of meta-analysis is the pooling of multiple studies. Although pooling does not improve the quality of each individual study, it does reduce limitations of small sample sizes as pooling creates a larger sample from which to detect an effect and reach significance (Lipsey, 1997, 2014). Additionally, meta-analysis has the unique ability to perform moderator analysis. The use of moderator analysis allows researchers to investigate whether certain characteristics are associated with larger or smaller effect sizes, and whether this pattern exists across a set of studies; a relationship that is likely to be impossible to disentangle otherwise (Lipsey, 1997, 2003).

1.2.3. Criticisms of meta-analysis

Mixing "apples with oranges"

Although meta-analysis is largely credited for its transparency and objectivity, there is one aspect of meta-analysis that is subjective, and is often subjected to criticism. In meta-analysis, the concept of 'mixing apples and oranges' refers to the importance of grouping studies and outcomes that are comparable. As Lipsey and Wilson (2001a) state,

> because meta-analysis focuses on the aggregation and comparison of the findings of different research studies, it is necessary that those findings be of a sort that can be meaningfully compared. This means that the findings must (a) be conceptually comparable, that is, deal with the same constructs and relationships and (b) be configured in similar statistical forms (p. 2).

The issue open to criticism is the subjective nature of categorization. That is, what one person considers 'conceptually comparable' can vary from person to person (Lipsey & Wilson, 2001a, p. 3)

"Garbage in, garbage out"

Lipsey & Wilson (2001a) refer to this criticism because meta-analysis summarizes primary research and does not have the ability to improve the quality of primary studies. As such, the meta-analysis of biased studies can have an additive effect and create a cumulative bias. That is, if the sample of studies in the analysis is largely composed of methodologically weak research designs and biased studies (i.e. "garbage"), it can lead to a cumulative effect of biases and can undoubtedly lead to wrongful conclusions. Hence, the meta-analyst will get "garbage out".

1.3. Summary

In sum, despite the proliferation of evaluation research on supervision programs (as can be seen in Figure 2 (Chapter 4) from the results of the current literature search), until recently few meta-analyses have paid particular attention to the body of research as a whole and the overall effect of community-based supervision programs for youth specifically. Furthermore, the existing evidence for the effect of community-based supervision programs on juvenile recidivism is mixed. Moreover, the studies that do exist are of relatively weak scientific quality (Byrne, 1990; Corbett, 1999), indicating that more methodologically rigorous evaluations are required to capture the full effect of supervision-based practices and their effect on recidivism among youth. Despite less than ideal evidence surrounding aftercare and intensive supervision probation programs, the importance of these interventions in the modern criminal justice system cannot be understated. Considering that variations of probation are the most common sentences for youth (Bennett, 1988; Hockenberry & Puzzanchera, 2014; Lemert, 1993; Torbet, 1996), quality research is necessary to further assess and improve the interventions and find out what is working. As such, this research seeks to expand upon the existing evidence as well as address some of the methodological concerns that are prevalent in the literature on this topic. The current analysis distinguishes itself from the existing meta-analyses on the topic in two primary ways:

First, an important strength of the current analysis lies in the methodology. The current study differs from existing studies through the application of strict inclusion and

exclusion criteria (discussed in chapter 3). Specifically, in contrast to the existing studies on the topic, the current analysis focuses specifically on supervision-oriented interventions (i.e., interventions where surveillance is the primary component of the intervention), excludes programs that focus on family-based interventions, as well as programs that target a specific population of youth such as substance users, sex offenders, and offenders with mental health illness. Although the strict a priori criteria resulted in the exclusion of multiple studies, it is believed that the selection criteria are a strength of this meta-analysis, and will allow the results to be highly generalizable to the majority of young offenders. Furthermore, the present study used an extensive search strategy (see Chapter 3 for the list of search terms and grey literature that were used in the search). Due to the extensive search strategy, the authors are confident that most of the relevant literature, if not all, were identified. As such, the stringent methodology used here separates itself from existing meta-analyses on the topic.

Second, this meta-analysis is unique because it provides a bifurcated analysis of recidivism as an outcome measure. An extensive review of the literature on crime prevention programs for youth (including supervision-oriented, rehabilitation and treatment-oriented programs) shows that meta-analyses consistently use the umbrella term of recidivism as an outcome measure. As will be seen below, the current analysis bifurcates recidivistic outcome measures to demonstrate the effectiveness of each intervention towards i) alleged offenses and ii) convicted offenses; a distinction that is demonstrably important when evaluating the treatment effects of community-based supervision programs.

In the same vein, existing meta-analyses on the topic typically focus on one single intervention (e.g., either provide a meta-analysis on aftercare alone, or ISP alone, not both). A thorough examination of meta-analytic studies has shown that even when two meta-analyses claim to summarize research on the same topic, it is possible that the meta-analyses analyze two distinct data sets. To explain, meta-analyses commonly use distinct search strategies and selection criteria, so although two meta-analyses summarize the effectiveness of aftercare programs, a quick look at the methodology and included studies may reveal that the studies included in the two meta-analyses overlap only slightly. This makes it difficult to cross-compare effectiveness of interventions. As

the current study uses the same search strategy, selection criteria, and methodology for the two interventions, direct comparison of the effectiveness of interventions is more reliable and meaningful. In conclusion, the stringent methodology (i.e., exhaustive search strategy and strict selection criteria), the generalizability of the results to the 'typical' offender, as well as the unique bifurcated analyses on recidivistic outcome measures illuminates the ways in which this study differentiates itself from existing meta-analyses and underlines the importance of this study.

Chapter 2. Objectives

In recent years, correctional evaluation research has expanded vastly. With the emergence of the 'what works' era and the accompanied proliferation of evaluation studies, navigating through the sheer volume of intervention evaluations can quickly become a daunting task as the literature is replete with mixed results, methodological concerns, and various measures of effectiveness. Still, relatively little is determinately known about the effects of community-based supervision programs on juvenile recidivism.

The primary aim of this thesis is to summarize the effectiveness of community-based supervision programs on juvenile recidivism. This aim will be achieved by way of the following objectives:

- First, the review will assess the effects of aftercare/re-entry programs on juvenile recidivism compared to the recidivism rates of youth who receive 'care as usual' post-incarceration.

- Second, the review will assess the effects of intensive supervision probation programs on juvenile recidivism compared to the recidivism of youth on standard probation.

- The analysis will further explore the effectiveness of community-based supervision programs by considering the differing impact of the two programs on two types of recidivistic outcome measures: alleged offenses and convicted offenses.

Chapter 3. Methodology

This chapter presents and explains the steps and procedures used in the meta-analysis. First, the steps followed to conduct the literature review are discussed, including the selection criteria and search strategy. Second, information is provided on data collection and management; specifically, study selection procedures, data extraction, decision rules, and coding guidelines. Finally, analytical techniques are discussed, including calculation of effect sizes, procedures used to pool data, assessment of heterogeneity, sensitivity analysis, and subgroup analysis.

3.1. Selection criteria

Eligibility criteria (selection criteria that determine whether studies are to be included or excluded from the analysis) are especially important in a meta-analysis. Outlining and defining constructs to determine inclusion rather than intuitively following "[I'll] know it when I see it" (Card, 2011, p. 39) is essential to a smooth and efficient literature search. Explicit inclusion and exclusion criteria aid in establishing necessary conceptual boundaries (features of studies and/or interventions that should or should not be included in the analysis), maintaining transparency and facilitating replication (explicitly reporting all steps and all decision rules so that others can replicate the study by following the same steps), as well as minimizing bias and optimizing objectivity throughout the selection process (Card, 2011; Littell et al., 2008).

Selection criteria for this study were developed *a priori* and were adopted in part from a meta-analysis conducted by Wong, Bouchard, Bouchard, Morselli and Gravel (2013).

3.1.1. Inclusion criteria

Types of interventions

Eligible studies for inclusion were a) intensive community-based supervision and monitoring programs that target juvenile offenders, where b) supervision (as opposed to treatment) was the *primary focus* of the intervention. Interventions were specifically limited to 1) Intensive Supervision Probation (ISP) programs and 2) aftercare/re-entry programs. The two interventions can be distinguished by the following characteristics.

Interventions classified as ISP were:

- an intensive form of probation, characterized by smaller caseloads and more frequent supervisory contacts in comparison to traditional probation.

Interventions classified as aftercare/re-entry were:

- programs that had a primary focus of enhanced supervision services (small caseloads and increased supervision) for youth following incarceration or placement in a closed facility.

Although supervision/monitoring must be the primary focus of the ISP and aftercare/re-entry interventions, it need not be the only component of the intervention. In addition to intensive supervision, youth may also receive treatment services while under supervision (either from the probation officer or caseworker directly or through brokered community services).

Types of Participants

Participants were juvenile offenders who were in contact with the criminal justice system and were primarily between the ages of 12 and 18 years old. When the age range of participants was slightly above or below 12 to 18 (maximum age 21 years) but the *average age* of study participants fell in between 12 and 18 years of age, the study was included. For example, the age range in Bergseth and McDonald (2007) was 11 years to 19 years (mean 16.3 years), so was included for analysis.

Types of outcome measures

Measured outcomes were quantitative and criminogenic in nature (see section 3.1.2. below for a list of exclusions regarding 'nature' of offense). The presented data for outcome measures must also have provided sufficient numerical or graphical data to allow for computation of an effect size.

Types of studies (design)

Studies selected for inclusion were restricted to moderately rigorous comparison/control group designs, i.e., randomized controlled trials, quasi-experimental designs in which participant characteristics were matched on important variables (e.g., criminal history), or quasi-experimental designs in which the treatment and control groups were matched on at least some variables (e.g., age, sex).

This criterion was in part developed and informed by The Maryland Scale of Scientific Methods; a tool used to rank the methodological rigour of a study. The five-point ordinal scale (1 (weak) to 5 (strong)) corresponds with the scientific rigour and internal validity of a study's methodology (Sherman, Gottfredson, MacKenzie, Eck, Reuter, & Bushway, 1998). To be deemed 'appropriate' for inclusion in this study, studies must be rated as at least a level three for research design on the Maryland Scale, which identifies studies as being at least moderately rigorous in research design. See Appendix G and H for a summary of studies, including the Maryland Scale rating for each study.

Sample size

Treatment and control groups used a minimum sample of 20 subjects.

Although there are no determinate rules in terms of sample size that is considered adequate for inclusion in a meta-analysis, it is well known that studies with small sample sizes are subject to greater sampling error which may lead to inaccurate estimation (Lipsey, 2014; Littell et al., 2008). To minimize the methodological limitations associated with small sample size, samples of less than 20 participants in either the treatment group or the control group were excluded.

Setting

The program was delivered at least partially in a non-closed setting in the community.

The purpose of the research was to examine the effects of community-based supervision programs; however, the definition of what was to be included as "community-based" required specification. As discussed above, the community-based supervision component could follow the supervision component in a closed setting (i.e., prison, jail). Any *other* form of 'closed' or 'partially-closed' settings (e.g., school-based, secured-facility boot camp, or wilderness camp) that preceded the community-based supervision, or where the intervention itself was being carried out, was excluded. Further detail on this criterion is provided in section 3.1.2.

Unit of analysis

Reported on at least one individual-level outcome measure of crime.

Time frame

Studies were published between the dates January 01, 1990 and April 21, 2015.

Pooling data from studies conducted over the past 25 years was deemed sufficient to summarize the current state of community-based supervision programs. For example, boot camp and wilderness programs were very popular in the 1980s, but are no longer commonplace. Extensive research was conducted on these programs throughout that time which suggested their ineffectiveness. Extending the research beyond the last 25 years might include types of programs that are no longer being conducted or pursued and, thus, would affect the relevance of the pooled effect and generalizability to studies that are prevalent today.

Language restrictions

Primary studies were published in English.

Country where study was conducted

Studies were conducted in Canada, the United States, Australia, New Zealand, or a Western European country.

In order to maximize the generalizability of the results, studies were restricted to those that were conducted in countries that are reasonably similar in terms of types of offenders, general approach to criminal justice, and intervention types.

3.1.2. Exclusion criteria

Types of interventions

Where the primary intervention in a study was 'traditional' supervision (i.e., standard probation), the study was excluded. The reason for this is that although standard probation is an important intervention, standard probation served as the comparison group for all of the intensive supervision probation programs, and was therefore not appropriate to include.

As discussed in section 1.1.2, there exists a continuum of eight correctional options that are classified as intermediate sanctions: intensive supervision programs, monetary penalties (including fines), community service, day reporting centers, home confinement (including electronic monitoring), boot camps, day halfway houses, and aftercare/post-release supervision (Caputo, 2004). Although this thesis focuses specifically on the categories of ISP and aftercare, it is noteworthy that an aftercare component can be part of an intermediate sanctions intervention, such as a boot camp. As outlined in section 3.1.1, any intervention that preceded the community-based supervision or was carried out in a partially closed setting other than a prison or jail was excluded. This is because, as specified above, the purpose of this study is to evaluate the effect of supervision on recidivism. If some youth receive aftercare as a component of a boot camp program, there would be no way of separating the effects of the boot camp intervention from the effects of the aftercare supervision itself.

Types of participants

In order to keep the conclusions of this analysis generalizable to a broad group of juvenile offenders, very specific types of offenders such as perpetrators of domestic violence, those with serious mental health problems, substance users, sex offenders, and known gang members were excluded. As substance abusers make up a large portion of juvenile offenders, the exclusion of this population warrants specific attention. First, substance abusers and interventions that target substance users is a sizeable area of interest that is deserving of an analysis of its own. Consistently, there is a growing body of literature that is dedicated to specific population and targeted interventions (see Filges Knudsen, Svendsen, Kowalski, Benjaminsen, & Jorgensen 2015; Filges, Rasmussen, Andersen & Jorgensen, 2015; and Lindstrom, Saidj, Kowalski, Filges, Rasmuseen, & Jorgensen, 2015 for recent systematic reviews conducted on the topic).

This operationalization and selection criterion is considered a strength of this meta-analysis, and will allow the results to be highly generalizable to the majority of young offenders.

Types of outcome measures

For the purposes of this analysis and its focus on criminal recidivism, status offenses (truancy, antisocial behaviour, traffic violations) and technical probation violations were not selected. In addition, studies that measured substance use (tobacco, alcohol, illicit substances) as the only outcome measure were not considered; however, when substance use was the primary outcome measure but the study also reported other recidivistic outcome measures, the study was eligible for inclusion.

The exclusion of these outcome measures was intentional and strategic. Status offenses, technical probation violations and substance use were not included because the authors wanted to capture the effects of the intervention on criminal recidivism rather than juvenile delinquency.

Types of studies (design)

Consistent with the inclusion criteria listed above that specifies the selection of moderately rigorous study methodology in accordance with the Maryland Scale, studies were excluded when the appropriateness of the comparison group was questionable.[1]

It can be argued that for meta-analyses to be truly comprehensive and representative of the literature on a topic, the selection of studies for inclusion should not discriminate based on study design and methodology (Card, 2011, p. 27). On the one hand, strict criteria towards the inclusion of methodologically rigorous designs ensure that the study is summarizing the most rigorous literature (Lipsey & Wilson, 2001a). On the other hand, evaluations that are considered to be methodologically rigorous represent only a small portion of the literature, and, thus, including only methodologically rigorous studies would reduce the sample substantially and potentially introduce bias. That being said, methodologically weak study designs are known to overestimate study effects compared to more methodologically rigorous designs (Durlak, 2009). Although there are many strengths associated with conducting a meta-analysis, as discussed above, meta-analysis does not have the power to improve the quality of primary studies, so, if you put "garbage in", you will get "garbage out" (Lipsey & Wilson, 2001a, p. 9).

Setting

The intention of the research was to gather evidence on the effectiveness of interventions in which supervision was the primary component of the program and would not be confounded by settings that threaten the validity of 'supervision'. As stated above, 'community-based interventions' was operationalized so that the supervision component could follow a period of incarceration in a closed setting. Any other form of 'closed' or 'partially-closed' settings (i.e. school-based, secured-facility boot camp or

[1] Study designs that were excluded include single group pre-test/post-test, quasi-experimental designs for which the treatment and control group were shown not to be comparable at baseline, or studies in which attrition bias and selection bias were a substantial concern (see 20% attrition rule in section 3.4.4).

wilderness camp) that preceded the community-based supervision, or where the intervention itself was being carried out, were excluded.

Other

Studies were excluded if the outcome measures were not for the youth who participated in the program or where the primary target of the intervention was someone other than the youth him/herself (e.g., parent, family, sibling).[2]

3.2. Search strategy for identification of studies

In essence, the literature search in a meta-analysis is equivalent to the data collection process in primary research, and each individual study that is ultimately included in the analysis is a participant (Card, 2011). In the same way that a large sample of participants increases generalizability and representativeness, the same logic applies to a meta-analysis, where a large sample of studies increases representativeness and generalizability of the results from the pooled analysis (Card, 2011). Detailed step-by-step documentation of research decisions and analytic procedures is a unique feature of meta-analysis. The systematic searching of databases is important for reasons of transparency (protecting from author biases and permitting research decisions to be open to criticism) and replication (i.e., by following identical decisions and procedures, any researcher should be able replicate the search strategy and analysis, holding authors directly accountable for their research decisions) (Lipsey & Wilson, 2001a). Furthermore, conducting an exhaustive and comprehensive search of the literature will also reduce various forms of bias (sampling bias, publication bias) and

[2] For example, in some circumstances the primary focus of the program was parenting-related skills or workshops, or family-related interventions (e.g., Functional Family Therapy) where in essence, the study was evaluating the effect of the intervention on the youth, however the outcome measure was related to an entity other than the youth themselves and the youth did not participate in the program directly.

increase generalizability by ensuring that the set of studies is as representative of the general population of studies as possible (Littell et al., 2008).

3.2.1. Electronic searches

The following 20 electronic databases were searched with the terms listed below.

Academic Search Premier
Canadian Research Index
Cochrane Central Register of Controlled Trials
Cochrane Database of Systematic Reviews
Criminal Justice Abstracts
Database of Abstracts of Reviews of Effects
Education Resource Information Center (ERIC)
Medline
National Criminal Justice Reference Service (NCJRS)
Open Access Theses and Dissertations
ProQuest Dissertations and Theses Full Text
PsycARTICLES
PsycBOOKS
PsycINFO
Public Affairs Information Service (PAIS) International
Social Sciences Abstracts
Social Sciences Full Text
Social Services Abstracts
Sociological Abstracts
Web of Science

3.2.2. Grey literature search

Although the number of studies identified and selected from the electronic databases can be prolific, limiting a search to electronic databases is not sufficient for a comprehensive and unbiased search. One of the major problems with relying solely on electronic databases for a comprehensive literature search is that electronic databases

are not representative of the body of literature as a whole (see section on publication bias below). That is, even the most thorough search of electronic databases will not uncover all relevant studies (Card, 2011). Thus, the search of grey literature is pivotal to offsetting publication bias by identifying unpublished works such as technical reports, theses and dissertations, conference papers, and individual projects. In the present study, efforts toward a comprehensive literature search continued with a search of grey literature as well as an ongoing backward search.

Hand searching

The term "hand searching" refers to a manual inspection of relevant journals (either page by page, or searching table of contents) for articles that should be considered for inclusion (Littell et al., 2008). Journals believed to contain articles that were relevant to the search but not identified in the primary search of electronic databases were hand-searched[3]. Additional avenues that were hand-searched include the curriculum vitae (CVs) of pertinent authors as well as the websites of relevant organizations[4].

Backward searching

In addition to searching grey literature (hand searches), backward searching was used to identify relevant literature that was not uncovered by electronic database searches. Backward searching, also known as "reference harvesting" is a less systematic approach to finding relevant literature, but is nonetheless an effective strategy to locate relevant studies that were not identified in the search of electronic databases (Littell et al., 2008). Throughout the searching and coding stages of a meta-analysis, backward searching refers to searching out citations in previously-selected studies that were not uncovered in

[3] See Appendix F for a list of journals that were hand searched.

[4] Regrettably, the process of hand searching grey literature was not systematic and was not consistently tracked. A complete list of specific authors and websites that were searched are not included here. However, the curriculum vitae of authors whose studies were identified for inclusion in the literature search were reviewed for relevant studies. Additionally, the Office of Juvenile Justice and Delinquency Prevention (OJJDP) website was extensively searched for relevant publications.

the literature search of electronic databases, but that might be relevant to the current study. On an on-going basis, the literature reviews (and reference lists) of studies that were selected for inclusion were scanned for relevant studies that were not uncovered via the aforementioned avenues. Additionally, any meta-analysis pertaining to the subject of intensive supervision probation or aftercare/re-entry programs was retrieved and scanned for articles, reports, theses, dissertations, or individual papers that were applicable to the present study.

3.3. Search terms

Four constructs were used in the development of a comprehensive list of key search terms: (1) juveniles, (2) criminogenic outcome, (3) intervention, and (4) evaluation. Search engines use key terms to identify potentially relevant studies, and for that reason, it is important that the key terms used in the search strategy are all-encompassing. To ensure that the search strategy would capture any and all literature pertaining to evaluation studies of intermediate sanction interventions on juvenile recidivism, relevant literature was consulted to find synonyms and interchangeable terms and phrases. Additionally, Boolean operators and wildcard marks were used to broaden the search. The search terms for each construct (shown below) were developed over multiple trial and error iterations. The search strategy was applied to all 20 electronic databases and key terms were searched in Abstract. Challenges relating to varying language and developing a comprehensive literature search (particularly pertaining to construct 3) are discussed in Chapter 6.

Construct 1: Juveniles

(youth* OR juvenile* OR adolesc* OR teen*)

Construct 2: Criminogenic outcome

(crime* OR criminal* OR devian* OR violen* OR delinquen* OR offend* OR offense* OR recidiv* OR gang OR gangs)

Construct 3: Intervention

(diversion* OR probat* OR parole OR aftercare OR reentry OR surveillance OR supervis* OR deterrence OR “alternative to imprisonment” OR “alternative to incarceration” OR “alternative to detention” OR restitution OR rehabilitat* OR “intensive community program” OR “graduated sanction*” OR “intermediate sanction*” OR “shock program*” OR “shock incarceration” OR “boot camp*” OR retribution OR counseling OR mentor* OR wilderness OR “day center*” OR “day reporting” OR “reporting center” OR “early release” OR “pretrial release” OR “supervised release” OR “electronic monitoring” OR “home confinement” OR “house arrest” OR “home detention” OR “community tracking” OR “community service” OR “halfway house” OR “transitional center*” OR “community correctional center*” OR “community release center*” OR “work release” OR “electronic release” OR “work camp” OR “residential treatment” OR “residential placement*” OR “residential service*” OR “residential program*” OR “wraparound program” OR “wraparound service*”)

Construct 4: Evaluation

(evaluat* OR effect* OR impact* OR outcome*)

3.4. Data collection and analysis

3.4.1. Selection of studies

Using electronic databases and the search terms listed above, as well as the grey literature search techniques, one reviewer read through the titles and abstracts of identified hits to determine studies that were relevant and to be retrieved for further review. The search of the 20 electronic databases identified an initial 12,199 studies for

review.[5] Once titles and abstracts of studies were screened and deemed to meet initial eligibility criteria, a master list was created for relevant articles to be retrieved in full. When articles were difficult to locate, the Simon Fraser University interlibrary loan (ILL) system was used.

Once the articles were retrieved in full, two reviewers used the inclusion and exclusion criteria listed above to determine whether articles met the full list of selection criteria. If there was disagreement between reviewers concerning the inclusion or exclusion of a study, reasoning for the discrepancy was discussed until a consensus was reached. All retrieved studies were then classified into 18 different categories relating to the topic of intermediate sanctions (e.g., aftercare/re-entry, intensive supervision, boot camps with/without aftercare, wilderness programs with/without aftercare, home confinement, halfway houses, day reporting centers, residential facilities, etc.). Studies that were classified as either aftercare/re-entry or intensive probation supervision were selected for inclusion for this analysis and were coded by two independent reviewers.

3.4.2. Data extraction and management

Once consensus was reached for the inclusion of individual studies, data from each study were extracted and coded (into an Excel spreadsheet) by two independent reviewers. Data were extracted for 63 variables including general study information (e.g., publication year, publication type, location), program characteristics (e.g., number of sessions, number of weeks, family involvement, risk level), sample characteristics (e.g., sample size, participant age, gender mix), outcome measures (e.g., outcome measure coded, outcome source, follow-up period), study characteristics (e.g., research design, random assignment), and treatment group and control group outcomes (e.g.,

[5] Although this study has a narrow focus of ISP programs and aftercare/re-entry programs, the search strategy focused on the continuum of intermediate sanctions. For that reason, the number of initial hits identified here (12,199) would have been substantially less if the search strategy only used terms and phrases related to these two interventions specifically.

mean, percentage, frequency of recidivism). As per above, any disagreements between coders were discussed and resolved. See Appendix D for a comprehensive list of variables that were extracted for coding, as well as the coding guide that demonstrates the respective coding schemes for each variable.

3.4.3. Measures of treatment effect

One of the purposes of meta-analysis is to produce a single overall effect size statistic by combining the effect sizes from multiple studies. An effect size is a metric that is computed to demonstrate the direction and magnitude of an effect between two variables (Card, 2011; Littell et al., 2008). To produce a summary effect, the extracted outcome data for each study is used to calculate an effect size that tells us about the relationship between two variables. From here, effect sizes from each individual study can be pooled together (Lipsey & Wilson, 2001a; Littell et al., 2008; Wilson, 2001). Practically speaking, combining the effects of multiple studies into one summary effect allows researchers to bypass the convoluted process of sifting through individual studies to draw conclusions about an overall effect. However, the outcome data that is extracted from each individual study is rarely readily suitable for a pooled analysis. As an illustration, one study might report arrest data as a dichotomy (yes or no), another might use a continuous scale, and another might use inferential statistics. Although the outcome of arrest is consistently reported across studies, each study uses a different statistical procedure to compute the respective findings. This is inappropriate for a meta-analysis because in order for data to be pooled, data must be commensurable. Hence, in order to pool effect sizes the data from each individual study must be standardized on a common metric so that studies can be meaningfully analyzed together and carry similar interpretive meaning and value (Dunst, Hamby, & Trivette, 2004; Lipsey & Wilson, 2001a, p. 35; Sanchez-Meca, Marin-Martinez & Chacon-Moscoso, 2003).

Dichotomous data

The majority of the selected studies used dichotomous data to report outcomes. With dichotomous data there are only two possible outcomes (yes or no; 1 or 0). Effect sizes are calculated by computing the number of treatment group participants who experienced an event (e.g., whether or not the youth was arrested during the follow-up)

compared to the number of control group participants who experienced the event (Littell et al., 2008)[6].

Odds ratio with Cox logit transformation (N=51)

The odds ratio for dichotomous data is calculated by comparing the odds of the event occurring in the intervention group compared to the odds of the event occurring in the control group (Lipsey & Wilson, 2001a). Put simply, "*odds* refers to the chance that something will happen compared to the chance that it will not" (Littell et al., 2008, p. 82). In this case, *odds* refers to the odds of someone who participated in an intervention committing a new offense compared to not committing a new offense relative to those who were in the control group.[7] See Equation 1.

$$Equation\ 1:\ ES_{OR} = \frac{ad}{bc} = \frac{P_a\ P_d}{P_b\ P_c} = \frac{P_a\ \div\ P_b}{P_c\ \div\ P_d} = \frac{P_a\,(1-p_c)}{P_c\,(1-p_a)}$$

In Equation 1, a, b, c and d correspond to the raw frequency of those who recidivated and those who did not for each group. Specifically, **a** refers to the number of youth in the treatment group who *were* arrested, **d** refers to the number of youth in the treatment group who were not arrested, **b** refers to the number of youth in the comparison group who *were* arrested and **c** refers to the number of youth in the comparison group who were not arrested. The superscript **P** refers to the proportion in the relative cell (a, b, c, d) and lower case **p** refers to the proportion of persons in its relative group (a or c) that experienced a positive outcome (reduction in recidivism) (Lipsey & Wilson, 2001a).

Continuous data

For the remainder of the studies, continuous data was used to report the number of times (frequency or average) treatment group participants experienced an event (i.e.,

[6] David Wilson's online effect size calculator was used for the computation of all odds ratios

[7] Equation from Lipsey & Wilson (2001a, p.53)

number of times arrested during the follow-up period, average number of arrests during a specified follow-up period) compared to the number of times control group participants experienced the same event.

Standardized mean difference (N=4)

As described above, the standardized mean difference is used to standardize the results from the various calculations so they can be comparable across studies and pooled (Breaugh, 2003; Card, 2011; Littell et al., 2008; Wilson, 2001). As demonstrated in Equation 2, the standardized mean difference (also known as Cohen's d) is the quotient of the difference between group means and the pooled standard deviation (Card, 2011; Littell et al., 2008).[8] See Equation 2.

$$Equation\ 2\text{:}\ d = \frac{M_1 - M_2}{sd_{pooled}}$$

Consistent with the idea of commensurability discussed above, it is inappropriate to pool effect sizes that are computed based on dichotomous data with those computed based on continuous data. Here, the Cox transformation was used so that effect sizes could be pooled. Additionally, although there are multiple ways to calculate effect sizes based on the statistical information that is presented, Lipsey & Wilson (2001a) advise that when the majority of effect sizes in the meta-analytic sample use dichotomous data (as in this analysis), it is best to use the logit transformation when computing dichotomous data into effect sizes and estimating *d*, and use the standardized mean difference to estimate *d* when computing effect sizes for the remaining continuous data (p. 56).[9]

$$Equation\ 3\text{:}\ d_{Cox} = \frac{\ln(ES_{OR})}{1.65}$$

[8] Equation from Card (2011, p.90)
[9] Equation from Card (2008, p.124)

Standardized mean difference with conversions:

When means and standard deviations are not available to compute the standardized mean difference, available inferential statistics (t-value, F-ratio) may be used (Lipsey & Wilson, 2001a). The four effect sizes calculated through the methods discussed below were also transformed using Cox.

One study presented an independent t-test with unequal group sizes[10]

$$Equation\ 4:\ ES_{SMD} = t\sqrt{\frac{n_1\ +\ n_2}{n_1\ n_2}}$$

Two studies presented an independent F-ratio with unequal group sizes[11]

$$Equation\ 5: ES_{SMD} = \sqrt{\frac{F(n_1 + n_2)}{n_1\ n_2}}$$

One study presented a logistic regression coefficient and standard error

Stafford and Glassner (2012) reported their results as a logistic regression coefficient and standard error. Hand calculations were conducted to calculate the equivalent odds ratio.

$$Equation\ 6: OR = \frac{1}{\beta}$$

Interpreting effect sizes

All effect sizes were calculated as odds ratios. As odds ratios are centered around 1 (a value of 1 represents no difference between groups), the interpretation of odds ratios

[10] Equation from Lipsey & Wilson (2001a, p.174)
[11] Equation from Lipsey & Wilson (2001a, p.199)

can be difficult. For the purpose of discussion and interpretation, the odds ratios were log-transformed. In order to ensure that effect size calculations are interpreted correctly, it is necessary to clarify that the logged odds ratio (LOR) calculations are centered around a value of 0. This means that a logged odds ratio of 0 demonstrates no differential treatment impact between groups (i.e., the event is equally likely for both groups). Further, studies were coded such that a value below 0 demonstrates that the intervention favours the control group and the treatment group is more likely to experience the event (e.g., arrest, incarceration) in comparison to the control group. In addition, a logged odds ratio with a value above 0 indicates that the treatment group is less likely to recidivate (beneficial impact) relative to the odds of the comparison group (Card, 2011; Lipsey & Wilson, 2001a; Littell et al., 2008; Szumilas, 2010).

3.4.4. Decision rules

As in primary research, the basic assumption that observations are independent of each other carries similar logic in a meta-analysis, where the calculation of a mean effect assumes that all effect sizes are independent (not related or influenced by one another) (Littell et al., 2008). Non-independence may occur when multiple effect sizes are calculated from the same or overlapping samples, or, similarly, when multiple effect sizes are used from the same study (i.e., reports of different outcome measures), in which case outcome measures should be pooled separately (Littell et al., 2008). To ensure the independence of effect sizes, the researcher should follow a set of decision guidelines. The following decision rules were used for ensuring independence of effect sizes.

Multiple reports from a single study

Where multiple documents containing the same data or overlapping samples were identified, the source with the most complete and comprehensive information was used in the analysis. For example, if a journal article was subsequently published from a report but the report contained more detailed narratives, descriptions, and results, the latter was chosen for inclusion (e.g., Hennigan, Kolnick, Tian, Maxson, & Poplawski. (2010) and Hennigan, Maxson, & Zhang (2005) provided outcome data for overlapping

samples. Hennigan et al. (2010) was excluded as the 2005 report contained more complete information on the description and outcome measures of juvenile offenders).

Multiple follow up periods

In the case of multiple available follow-up observations, the closest to 1-year follow up was chosen in order to maximize commensurability among studies, *unless* attrition from the longer follow-up was above 20%, in which case a shorter follow-up was chosen in order to avoid the risk of potential attrition bias. For example, Zhang and Zhang (2005) presented results for the follow-up times: 6 months, 12 months, and 18 months. From the 6-month follow-up to the 12-month follow up, the treatment group experienced an attrition of 20% (106 youth to 85 youth). In this case, attrition from one follow-up to another presented the potential risk of the results being affected by attrition bias (the youth remaining in the program being different than the youth who dropped out), as a result, the aforementioned decision rule to choose the follow-up closest to one year was applied and data from the 6-month follow-up was selected.

Multiple outcome measures of recidivism

Where multiple outcome measures of recidivism were reported, the most general or "total" outcome measure was coded. For example, "total arrests" were preferred over subtypes such as "misdemeanour" or "felony" alone or "property offenses" or "violent offenses" alone.

Multiple experiments within one document

Where outcome data was available for multiple independent sites, where applicable, the individual sites were coded as independent and unique studies. If only combined results were presented, the combined results across the various sites were coded as a single outcome. For example, Barnoski (2003) reports on the results from 11 independent study sites within a single report. In addition to presenting the effects for each individual outcome, he also presents the combined effect of these 11 sites. Effects on the independent study sites were favoured over the total calculated effect.

3.4.5. Dealing with missing data

One of the limitations of analyzing secondary data is the inability to control the information that is reported. Where missing data could not be inferred, the information was coded as "missing".

3.4.6. Assessment of heterogeneity

In a meta-analysis, the term 'heterogeneity' refers to the lack of similarity (i.e., diversity or variation) in treatment effects across studies. That is, when heterogeneity is present among effect sizes, it is important to consider whether the overall average effect is estimating the same population effect (common mean), as the heterogeneity is suggesting that the variability is not occurring by chance, and there are other factors driving the variability of effects (Lipsey & Wilson, 2001a). Heterogeneity can come from a variety of sources, e.g., participant characteristics (age, gender, ethnicity, risk level), intervention characteristics (type of intervention received, type of supervision, frequency of contacts, number of hours logged in the program), and methodological characteristics (study design, sample size, follow-up period, outcome source). When there is substantial heterogeneity, it is prudent to investigate potential sources through subgroup analysis. Investigating specific variables might help to further explain which characteristics are related to producing higher or lower effect sizes, for example, whether effects differ depending on the methodological rigour of the evaluation design (Card, 2011; Littell et al., 2008).

Two approaches were used to assess the degree and significance of heterogeneity: the Q-statistic and I^2 statistic. The Q-statistic indicates the significance of heterogeneity. That is, "if the Q exceeds the critical x^2 value given the *df* and level of statistical significance chosen, then you conclude that the effect sizes are heterogeneous" (Card, 2011, p. 186). If the Q-statistic is significant ($p<.05$), it is suggested that the variability is caused by sources beyond sampling error alone and moderator analysis should be pursued to investigate these other potential sources of variance (Lipsey & Wilson, 2001a). Meanwhile, the I^2 statistic communicates the magnitude of heterogeneity that is present and addresses the percentage of variability that can be attributed to factors other than chance. "This index [of heterogeneity] is

interpreted as the percentage of variability among effect sizes that exists between studies relative to the total variability among effect sizes (Card, 2011, p. 188). An I^2 statistic exceeding a percentage of 50% is considered substantial and warrants further examination through moderator analysis (Card, 2011; Littell et al., 2008).

3.4.7. Data synthesis (meta-analysis)

In a meta-analysis, there are two possible approaches to modelling data. The model chosen to analyze the outcome data (e.g., a fixed-effects model or a random-effects model) is contingent on the sources of variation that are believed to cause the heterogeneity (Guolo & Varin, 2015). For instance, if heterogeneity is low and variance can be attributed to sampling error alone, a fixed effects model should be used (Borenstein & Higgins, 2013; Lipsey & Wilson, 2001a; Littell et al., 2008). However, when the effect sizes across studies are deemed to be heterogeneous and the effect may be due to more than sampling error alone, a random effects model is used (Littell et al., 2008; Wilson, 2001). Although the general literature calls for the use of a random effects model when heterogeneity is present, recent literature suggests that there are limitations to using a random effects model with small samples, in which case using a fixed effects model might be more appropriate (even when heterogeneity is present) (Guolo & Varin, 2015; Schulze, 2007). In accordance with this literature, fixed effects models were used for the primary analyses, however, due to the presence of heterogeneity, the results from random effects models are also shown.

Inverse variance weights

Fixed effects model

In a fixed effects model, studies are weighted by the inverse of their variance. That is, the weight assigned to each study is directly related to its standard error and confidence intervals (Lipsey & Wilson, 2001a, p. 36). The larger the sample size, the smaller the standardized error and the smaller the confidence intervals, which invariably leads to more precise estimates (Wilson, 2001). As larger studies (bigger N) are more precise estimates of the effect, it is argued that greater weight should be given to these studies with higher levels of precision, while smaller studies with less precise estimates (more

room for error by chance) should be given smaller weights and subsequently contribute less to the overall pooling/averaging to studies (Sterne, Egger, & Smith, 2001; Wilson, 2001). For example, considering their varying levels of precision, it would not be fair for a study with a sample of 20 to contribute to the overall pool of effects to the same degree as a study with a sample of 800. It therefore makes sense to weight each effect size by their relative sample size (the larger the N, the larger the weight) so that more precise estimates of treatment effects will contribute more to the overall average of studies rather than being potentially distorted by imprecise estimates (Durlak & Lipsey, 1991; Wilson, 2001).

Random effects model

In contrast to the fixed effects model (wherein it is assumed that variability among effect sizes is due to sampling error alone), the random effects model assumes that variability is due to factors beyond sampling error (Card, 2011; Lipsey & Wilson, 2001a). In effect, both models weight studies by inverse variance; however, the factors taken into consideration for the computation of weights is dependent on the factors that are assumed to be causing the heterogeneity (Lipsey & Wilson, 2001a). The computation of random effects weights therefore takes into account two factors: the overall estimated population variance (random variability) and the standard error for each individual study (Card, 2011; Lipsey & Wilson, 2001a). Card (2011) further explains that "studies still have the same relative ranking of weights using random- or fixed-effects models (i.e., studies with the largest weights for one hand the largest weights for the other), [however] the discrepancies in weights across studies is less for random- than for fixed-effects models" (p. 238). That is, in comparison to a fixed effects model, the weightings of each study in a random effects model will, in general, be smaller, and studies with large sample sizes will be weighted less heavily (and smaller sample sizes will be weighted more heavily).

3.4.8. Subgroup analysis and investigation of heterogeneity

Where there is substantial heterogeneity between studies, subgroup analyses can be performed. In essence, subgroup analysis allows researchers to model differences between groups by performing separate "mini" meta-analyses on variables of

interest that the researcher believes might be related to a certain effect and explain heterogeneity. That is, where varying effects between studies (heterogeneity) might be due to an underlying characteristic, such as a methodological variable, rather than the treatment intervention itself, subgroup analysis may be used. For example, if the researcher believes that studies using rigorous research designs consistently produce smaller treatment effects in comparison to studies that use weaker methodology, the researcher can conduct a mini meta-analysis on each group. From here, the researcher can evaluate a) whether there is a difference, b) if this difference is statistically significant, and c) which variable is the stronger group (rigorous versus weak methodology) when it comes to producing effects. This is important for purposes of interpretation because without investigating heterogeneity and potential moderators, we might draw false and misleading conclusions about the effects of an intervention.

There are three heterogeneity related statistics that are important when it comes to performing subgroup analysis:

Q-total = the overall heterogeneity produced from the complete set of studies.

Q-within = Qx + Qy (e.g. the Q-statistic produced by the 'weak research design' mini meta-analysis, added to the Q-statistic produced by 'strong research design' mini meta-analysis)

Q-between = Q-total minus Q-within

A significant Q-between suggests that the variable is important in moderating results. The significance of Q-between indicates that there is a difference between the effects produced by each subgroup, and this difference is due to more than sampling error (or chance) alone (Card, 2011; Lipsey & Wilson, 2001a).

In sum, each 'mini' meta-analysis will calculate a pooled treatment effect for the individual subgroups. From there, heterogeneity can be examined from the statistical measures described above. Researchers can then compare the direction, magnitude, and statistical significance of that variable to determine whether it is a significant moderator of effect (differences between groups are greater than expected by chance)

(Card, 2011). Four study characteristics were selected for closer examination: publication type, research design, follow-up period, and sample size of treatment group. Chapters 6 and 7 will provide discussion on why these four study characteristics were selected for moderator analysis while other variables were omitted.

3.4.9. Sensitivity analysis

Sensitivity analysis is performed to assess the influence of outliers on the overall mean of the pooled treatment effect. Where an unusually large effect size contributes to the overall pooled effect, this lone study may exert substantial influence on the overall mean, thereby pulling the overall mean upward (or downward) and misrepresenting the true overall effect (Lipsey & Wilson, 2001a; Wilson, 2001). By re-running analyses with and without outliers, we can assess the data's sensitivity to change (Littell et al., 2008). If the overall pooled effect size does not change despite the inclusion or exclusion of an outlier, we can be reassured that the overall pooled effect was robust and was not misrepresented by a single study that may have caused concern for potentially distorting the distribution (Littell et al., 2008; Wilson, 2001).

3.4.10. Assessment of reporting biases

Literature on publication bias suggests that published studies are more likely to report positive and/or significant results (Card, 2011; Wilson, 2009). This is highly problematic when it comes to meta-analysis. As the aim of meta-analysis is to synthesize the body of literature on a given topic, the fact that only positive and statistically significant findings are traceable and readily accessible is concerning. When authors fail to publish important data due to its direction and level of significance, even the most extensive and comprehensive literature search (and grey literature search) will not accurately capture the full body of evidence. If investigators do not take reasonable precaution in searching diverse forums for published as well as unpublished works, there is a possibility that the end result of a pooled effect size may be biased towards positive findings (Card, 2011).

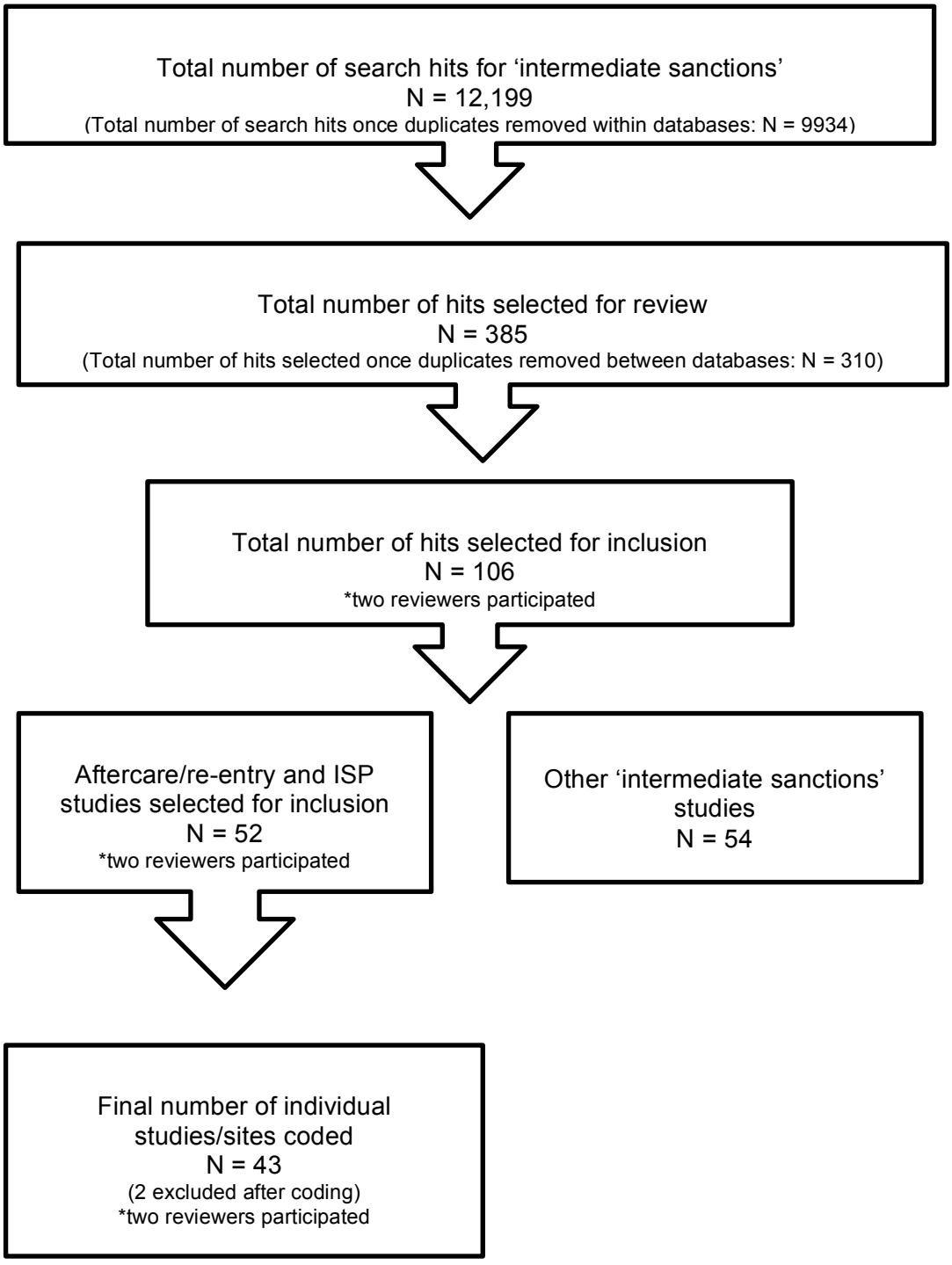
The funnel plot, a graphical method used to plot effect sizes (x-axis) and their relative standard errors (y-axis), is used to assess and detect publication bias (Littell et al., 2008). When no publication bias is detected, the shape of the plot is expected to resemble an inverted funnel (where less precise estimates (large standard errors) are expected to be widely dispersed around the mean, while more precise estimates (small standard errors) are expected to hover closely to the center line of the plot, that is, the overall mean effect across the set of studies (Littell et al., 2008; Sterne et al., 2001). In regards to examining the plot for bias, an absence of publication bias should reveal symmetry in the plot (symmetrical distribution of effect sizes on either side of the centerline of the plot). Conversely, an asymmetrical plot (a surplus of study-level data points falling on the right side of the plot compared to the left side of the plot) might be an indication of reporting bias and a potential overestimating of the overall pooled effect (Sterne et al., 2001). Furthermore, visual inspection of the funnel plot can also provide an indication of extreme or outlier effects.

Chapter 4. Results

4.1. Results of the search

The flow chart in Figure 2 presents the specific number of hits selected in each successive stage of the search strategy. A total of 12,199 hits were identified in the initial search of electronic databases. After searching the grey literature and carefully scanning the titles and abstracts of identified studies, 385 studies were selected as possibly meeting the inclusion criteria and were retrieved in full. Following the application of selection criteria to each study, 106 remained. At this time, the set of studies were classified into multiple intermediate sanctions categories; 52 of which were intensive supervision probation programs and aftercare/re-entry programs. Nine studies were excluded throughout the process of narrative coding and data extraction, and two studies (Barton & Butts, 1990 and Cillo, 2001) were excluded due to the nature of the comparison group (discussed below), resulting in a final sample of 41 individual studies/sites (from 27 studies).

Figure 2: Search strategy results



4.2. Aftercare/Re-entry

4.2.1. Characteristics of included studies

Overview of included studies

Table 1 demonstrates the characteristics of included aftercare/re-entry studies by outcome measure. Characteristics that were examined include: publication year, publication type, program delivery year, follow-up period, type of research design, gender, ethnicity, and treatment group sample size.

A total of 15 independent sites (from 10 studies) were selected for inclusion and contributed 24 effect sizes. With respect to outcome measure, recidivism included a) convicted offenses (convictions, sustained petitions, adjudications, incarceration) and b) alleged offenses (arrests, charges, referrals, court/police contacts). The distinction by outcome measure in this analysis was meant to differentiate between alleged criminal acts versus 'true'/convicted criminal acts. The majority of sites reported on both outcome measures (N=9); thirteen sites reported on alleged offenses, while 11 sites reported on convicted offenses, for a total of 24 effect sizes. Although selection criteria were open to the inclusion of studies from North America, western European countries, as well as Australia and New Zealand, all 15 independent sites included in the analysis were located in the United States. Table 1 provides a side-by-side comparison of study characteristics, by outcome measure, for studies that were included in the analysis.

Table 1: Characteristics of included studies for aftercare/re-entry

Study characteristics	N (%) Alleged offenses (N=13)	N (%) Convicted offenses (N=11)
Publication year		
1990-1994	2 (15.38%)	1 (9.09%)
1995-1999	3 (23.08%)	3 (27.27%)
2000-2004	0 (0.0%)	0 (0.0%)
2005-2009	7 (53.85%)	7 (63.64%)
2010-2014	1 (7.69%)	0 (0.0%)
Publication type		
Journal article	2 (15.38%)	0 (0.0%)
Report	11 (84.62%)	11 (100.0%)
Program delivery year		
1980-1989	2 (15.38%)	1 (9.09%)
1990-1999	6 (46.15%)	6 (54.55%)
2000-2010	5 (38.46%)	4 (36.36%)
Follow-up period [a]		
<1 yr	3 (25.0%)	0 (0.0%)
1 yr	7 (58.33%)	7 (70.0%)
>1 yr	2 (16.67%)	3 (30.0%)
Type of research design		
Randomized control trial	5 (38.46%)	4 (36.36%)
Quasi-experiment with matched comparison group	2 (15.38%)	4 (36.36%)
Quasi-experiment with weakly matched comparison	6 (46.15%)	3 (27.27%)
Sample gender mix		
All males	6 (46.15%)	6 (54.55%)
Mixed	7 (53.85%)	5 (45.45%)
Sample race/ethnicity		
Caucasian/mixed	3 (23.08%)	1 (9.09%)
Predominantly minority	10 (76.92%)	10 (90.91%)
Sample size in treatment group		
Less than 100	10 (76.92%)	8 (72.73%)
100+	3 (23.08%)	3 (27.27%)

[a]1 missing for alleged offenses; 1 missing for convicted offenses

Publication

With respect to publication type, the majority of studies were derived from non peer-reviewed sources (technical reports). Eighty-five percent of studies measuring alleged offenses were technical reports, and the remainder of the studies were from peer-reviewed journal articles (15%). Similarly, 100% of the studies that reported on convicted offenses were technical reports. With respect to year of publication, a slight majority of studies were published within the last 10 years; most were published between 2005 and 2009.

Intervention

Despite extensive efforts to code information for multiple intervention characteristics (e.g., number of sessions, length of treatment, whether participation in the program was voluntary or mandatory, whom the program interventions were delivered by, the risk level the program was targeting, whether an assessment tool was used to assess the risk level of youth, and whether parents/family were involved in the intervention), this information was in large part unreliably and inconsistently reported. As a result, the only intervention characteristic coded consistently enough to be included for analysis was 'program delivery year'. Implications of inconsistent and unreliable reporting of important information will be discussed in Chapter 6.

With respect to program delivery, the majority of programs were delivered between 1990 and 1999 (46% of studies evaluating alleged offenses and 55% of studies measuring convicted offenses). Approximately one third of the studies were evaluating programs that were delivered between 2000 and 2010 (38% and 36% respectively).

Sample

For purposes of analysis, gender was dichotomized between all-male samples and samples that were mixed (i.e., included both males and females), ethnicity/race was dichotomized between studies that were primarily composed of ethnic minorities and samples that were primarily Caucasian or near equivalent in ethnic composition, and treatment group sample size was dichotomized between studies with a sample of less

than 100 participants and studies with samples of more than 100 participants. The sample size ranged between 50 and 317 participants.

Overall, the composition of samples was similar for both alleged offenses and convicted offenses. With respect to sample size, studies were small; the majority of studies (77% and 73% respectively) used a treatment group of less than 100 participants. The vast majority of studies were primarily composed of minority youth, and all-male samples were predominant.

Design

As discussed in Chapter 3, the selection criteria specified that the comparison group of aftercare/re-entry programs must be composed of youth who received 'care as usual' or remained in custody for the completion of their sentence. One aftercare/re-entry study (Cillo, 2001) used a comparison group that consisted of youth who participated in traditional probation. Since this comparison group carries different interpretive implications than "care as usual" (and was thus not commensurable), the study was excluded.

Alleged offenses

With respect to research design, studies were evenly distributed between strong and moderately strong methodological designs. Seven studies used strong methodology (five studies (38%) used a randomized controlled trial and two studies (15%) used a quasi-experimental design with matched control group). Nearly half of the studies (46%) used a quasi-experimental design with weakly matched groups.

Convicted offenses

The majority of studies used rigorous methodology; four studies (36%) used a randomized controlled trial and four studies (36%) used a quasi-experimental design with matched control groups. Only 27% of studies employed a moderately strong design (used a quasi-experimental design with weakly matched groups).

Outcomes

As discussed in section 3.4.4., in the case of multiple follow-up observations the closest to 1-year was chosen to maximize commensurability. Length of follow-up was categorized among three intervals: less than 12 months, 12 months exactly, and more than 12 months.

The majority of the studies used relatively long follow-up periods. With respect to alleged offenses, 58% used a follow-up period of 12 months exactly and 17% used a follow-up of more than 12 months, while 25% used a follow-up period of less than 12 months. Similarly, 100% of studies measuring convicted offenses used a follow-up period of 12 months or longer.

4.2.2. Outcome 1: The effect of aftercare/re-entry on alleged offenses

4.2.2.1. Pooled effect

Aftercare/re-entry programs that presented outcome results of alleged offenses, i.e., criminal contacts, charges, arrests, etc. (N=13) were pooled together. Overall, the pooled effect for the fixed effects model yielded an estimate of 0.179 (z=2.18, p=.029), a positive and statistically significant result. This finding suggests that youth who participate in aftercare/re-entry programs are less likely to be charged for an offense or arrested upon release from custody in comparison to the control group that received 'care as usual'. The significant Q-statistic shown in Table 2 demonstrates a significantly heterogeneous sample (43.20, df=12, p<.001) and the I^2 statistic demonstrates that 78.8% of this heterogeneity can be attributed to factors beyond sampling error. Individual effect sizes, standard errors, confidence intervals and relative weights – calculated by inverse variance – can be seen for each individual study in Table 3.

Table 2: Fixed effects meta-analysis for aftercare/re-entry programs, alleged offenses

Pooled estimate	95% CI Lower	95% CI Upper	*Z*	No. of studies	*Q*-statistic	I^2
0.179	0.018	0.340	2.18, p=.029	13	43.20, df=12, p<.001	72.2%

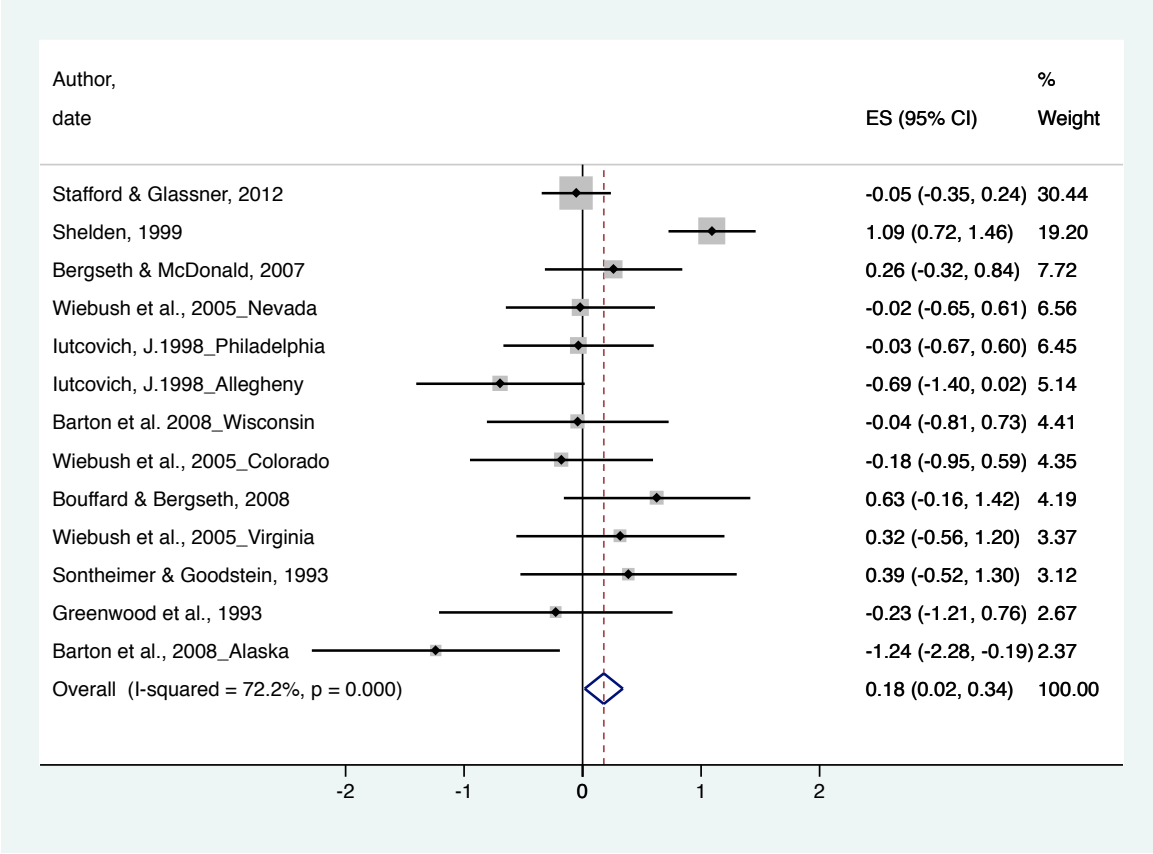
Table 3: Study-level data fixed effect model meta-analysis for aftercare/re-entry, alleged offenses

Author, year	Effect size (LOR)	Standard error (SE)	95% CI	Relative weight
Stafford & Glassner, 2012	-0.053	.149	-.345, .239	30.44
Shelden, 1999	1.093	.187	.725, 1.46	19.20
Bergseth & McDonald, 2007	0.262	.295	-.318, .841	7.72
Wiebush et al., 2005_Nevada	-0.019	.320	-.648, .610	6.56
Iutcovich & Pratt, 1998_Philadelphia	-0.034	.323	-.668, .600	6.45
Iutcovich & Pratt, 1998_Allegheny	-0.693	.362	-1.40, .018	5.14
Barton et al., 2008_Wisconsin	-0.041	.391	-.808, .726	4.41
Wiebush et al., 2005_Colorado	-0.178	.394	-.950, .594	4.35
Bouffard & Bergseth, 2008	0.629	.401	-.158, 1.41	4.19
Wiebush et al., 2005_Virginia	0.319	.448	-.559, 1.19	3.37
Sontheimer & Goodstein, 2008	0.388	.465	-.525, 1.30	3.12
Greenwood et al., 1995	-0.226	.502	-1.21, .759	2.67
Barton et al., 2008_Alaska	-1.238	.533	-2.28, -.192	2.37

Figure 3 (below) provides a visual representation of the data presented in Tables 2 and 3. The individual black dots (data points) plotted in the figure demonstrate the effect size for each individual study. The grey square around the data point demonstrates the relative weight each effect is given (inverse variance weight). Accordingly, a large square indicates that the study has a relatively large sample, a small standard error, and, thus, should produce a precise estimate of effect. The lines branching out from each square demonstrate the 95% confidence interval (CI) for that effect. The diamond and the dotted vertical line show the overall pooled effect for the

entire set of studies. The size of the diamond corresponds to the magnitude of the overall weighted pooled effect (i.e., the larger the diamond, the larger the overall pooled effect). The solid black line that is extended vertically through the middle of the plot demonstrates the 'line of no effect' (LOR=0). If either the diamond or the lines that represent the 95% CIs cross the 'line of no effect', the effect is not statistically significant. Finally, studies falling on the right side of the solid black line favour the treatment group and therefore demonstrate a positive treatment effect, and studies falling on the left favour the control group.

Figure 3: Forest plot for aftercare/re-entry, alleged offense



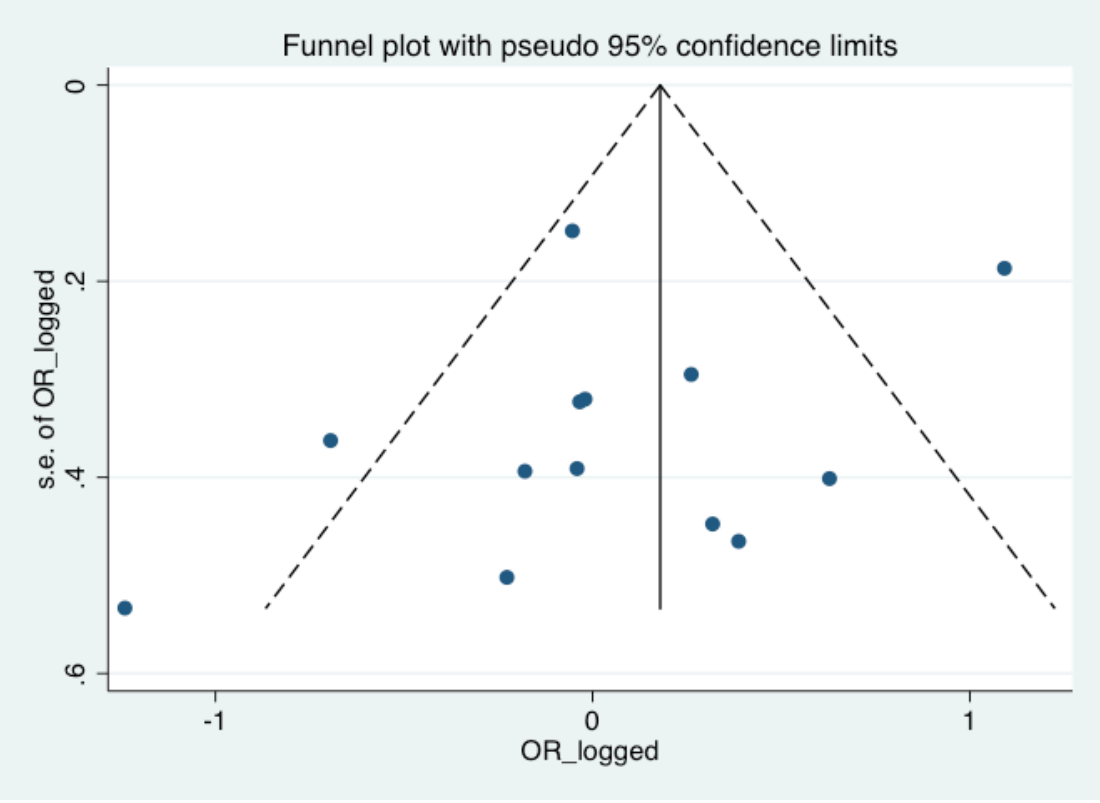
As illustrated above, logged odds ratios range from -1.238 to 1.093, demonstrating considerable heterogeneity and possible outlier effects. The plot further demonstrates that 5 studies produce a positive effect, however only 1 of these is statistically significant. E.g., there are 5 data points that fall on the right side of the plot, however only the 95% confidence intervals from Shelden (1999) do not cross the line of

no effect. Similarly, 8 studies demonstrate a negative impact, however, 4 of these studies have nearly null effects (an effect size of -.05 or less) and only 2 of these 8 are statistically significant.

4.2.2.2. Sensitivity analysis

As shown by the forest plot above, some of the data points fall far above or below the common mean effect (sometimes substantially). This raises a concern of outlier effects, wherein extreme values bias the overall mean effect by pulling it upward or downward. As discussed in section 3.4.10., although the funnel plot is typically used to identify the presence of publication bias, it may also be used for the identification of outlier effects. The funnel plot depicted in Figure 4 provides evidence of three outlier studies, demonstrated by the three data points that fall outside of the pseudo 95% confidence interval; Shelden (1999), Barton et al._Alaska (2008), and Iutcovich & Pratt_Allegheny (1998) were identified as the outlier studies.

Figure 4: Funnel plot for aftercare/re-entry, alleged offenses



While the funnel plot demonstrates the presence of outliers, it does not provide information on the degree of influence the outliers have on biasing the pooled estimate. To further examine whether the outlier studies were responsible for biasing the overall pooled effect, an influence analysis was conducted. In an iterative fashion, the influence analysis removes each study one by one and recalculates what the new pooled effect would be if that single study were to be omitted.

Table 4: Influence analysis on aftercare/re-entry studies, alleged offenses

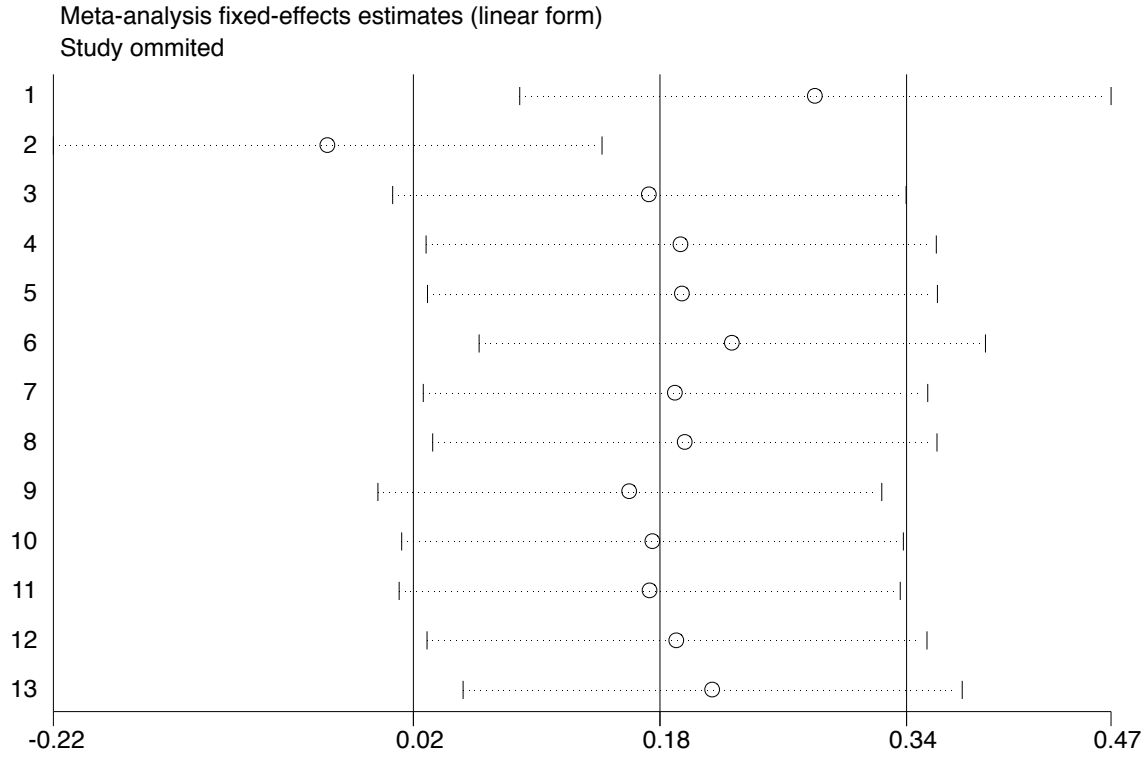
Study omitted	Effect size (LOR)	95% CI
Stafford & Glassner, 2012	.280	.087, .473
Shelden, 1999	-.038	-.217, .141
Bergseth & McDonald, 2007	.172	.004, .339
Wiebush et al., 2005_Nevada	.192	.026, .359
Iutcovich & Pratt, 1998_Philadelphia	.193	.027, .360
Iutcovich & Pratt, 1998_Allegheny	.226	.060, .391
Barton et al., 2008_Wisconsin	.189	.024, .354
Wiebush et al., 2005_Colorado	.195	.030, .360
Bouffard & Bergseth, 2008	.159	-.005, .323
Wiebush et al., 2005_Virginia	.174	.010, .338
Sontheimer & Goodstein, 2008	.172	.008, .336
Greenwood et al., 1995	.190	.026, .353
Barton et al., 2008_Alaska	.213	.050, .376
Combined	.179	.017, .340

Table 4 and Figure 5 demonstrate that Stafford and Glassner (2012) and Shelden (1999) carry substantial influence over the pooled results[12]. To explain, the table and figure illustrate that by omitting Shelden (1999) from the analysis, the overall pooled effect would shift to a coefficient of -0.038; further, the pooled effect would lose

[12] The numbered studies in Figure 5 correspond with the order of the studies in Table 4. I.e. study number 1 corresponds with Stafford and Glassner (2012), study number 2 corresponds with Shelden (1999), study number 3 corresponds with Bergseth and McDonald (2007), etc.

its significance. This is important because Shelden (1999) alone is so influential on the pooled effect that omitting it from the analysis would change the overall meaning and conclusions drawn from the analysis. In sum, the strong influence of one study on the overall effect demonstrates that the pooled effect is sensitive to change and is not robust.

Figure 5: Influence analysis for aftercare/re-entry, alleged offense



4.2.2.3. Alternative analyses: Random effects model and addressing outliers

Pooled effect

To examine the effect of outliers beyond an influence analysis, the data were re-analyzed with the outliers removed. Table 5 provides a comparison of the pooled estimate, z-score, p-values, and heterogeneity statistics when outliers are included in the analysis and when they are removed. To account for the significant heterogeneity found in the fixed effects model (with outliers), the table also provides an alternate approach to modelling the pooled analysis by using a random effects model. Consistent with recent

literature on meta-analysis data synthesis, the pooled effect from the random effects model (with outliers) is smaller than the pooled estimate (.065, z=0.38, p=.705) from the fixed effect model with outliers (0.179, z=2.18, p=.029), demonstrating that it is indeed possible that a random effects model underestimates the pooled effect when the number of studies being pooled together is less than 30.

Although the sensitivity analysis above showed three studies to be outliers (based on size of effect), it is also important to consider the weight that is assigned to each value. For example, when examining Table 3 and Figure 3 closely, we can see that, in correspondence to its large sample size and small standard error, Shelden (1999) is assigned a substantial weight (19.20%). Furthermore, we can see that Barton et al._Alaska (2008) and Iutcovich & Pratt_Allegheny (1998) are assigned small weights (2.37% and 5.14% respectively). As such, the latter two contribute minimally to the overall effect and are less of a concern in terms of extreme outlier effects that influence the pooled estimate. Thus, in the interest of keeping the set of studies as complete as possible, the analysis was re-conducted with only Shelden (1999) removed.

Table 5: Comparative meta-analyses for aftercare/re-entry programs, alleged offenses

Model	Pooled estimate	*Z*	No. of studies	*Q*-statistic	I^2
Fixed effects (with outliers)	0.179	2.18, p=.029	13	43.20, df=12, p<.001	72.2%
Random effects (with outliers)	0.065	0.38, p=.705	13	43.20, df=12, p=<.001	72.2%
Fixed effects (with outlier (Shelden) removed)	-0.038	0.42, p=.678	12	13.85, df=11, p=.241	20.6%

Consistent with the influence analysis presented in section 4.2.2.2., when Shelden (1999) is omitted from the pooled analysis, the overall estimate of effect changes in terms of directionality and statistical significance. Additionally, Table 5 demonstrates that the Q-statistic changes to a non-significant value (13.85, df=11, p=.241) and the I^2 statistic changes to 20.6%. This finding suggests that the new set of

studies is more homogeneous and is measuring similar underlying effects, where variation is likely due to sampling error alone. Furthermore, heterogeneity is no longer statistically significant, suggesting that Shelden (1999) was in large part responsible for the magnitude and significance of heterogeneity. See Appendix A for tables and figures from the full analysis with outliers omitted.

In sum, there is considerable variability in the distribution of effects, and the substantive and interpretive meaning of the analysis changes drastically depending on which model is used, demonstrating that the results are not robust, begging the question "which model produces the 'true' effect"? Due to the small number of studies, the fixed effects model was chosen for the primary analysis (Guolo & Varin, 2015; Schulze, 2007). However, due to the significant level of heterogeneity present in the sample, an argument can be made for using a random effects (RE) model. Guolo and Varin (2015) and Schulze (2007) state that the precision of the RE model is dependent on the number of studies included in the analysis. As such, when a small number of studies ($n<30$) is used in a meta-analysis, the random effects model is conservative, lacks power to detect effects, and underestimates the pooled estimate; in which case, the fixed effects model might be a more accurate representation (Schulze, 2007). As can be seen in Table 5, when using the random effects model for the analysis, the pooled estimate is substantially smaller and is no longer statistically significant.

Comparatively, the overall pooled estimate from the fixed effects model (with outliers removed) is substantially different when Shelden (1999) is removed from the analysis[13], suggesting that the analysis of the full set of studies might not be an accurate representation of the body of literature due to the distortion caused by the outlier. By virtue of being an outlier, it is assumed that Shelden (1999), in some way, is substantively different that the rest of the studies included in the sample (and thus

[13] Note that when there is zero heterogeneity, the random effects model and fixed effects model are identical. As the level of heterogeneity in the fixed effects model (with outliers removed) is not statistically significant, the full set of result for the random effects model is not shown here.

warrants exclusion). However, consistent with existing meta-analyses on the topic, there is significant variability between studies with respect to program effectiveness, and Card (2011) argues that it is misguided to not consider all of the available evidence, i.e., it is important to account for all studies. Moreover, as can be seen in the complete analysis of the fixed effects model with outliers removed (Appendix A), the redistribution of inverse variance weights carries the potential of creating new outliers. With the ever-present potential of new outlier effects and influential studies with each iterative process, continuing to remove outlying and influential studies could ultimately result in the exclusion of multiple studies. Although it is important to recognize the variation between models, and understand the advantages and disadvantages of each model, as the primary aim of this study is to summarize as much of the available literature as possible, the fixed effects model (with outliers) is the best model to represent the data.

4.2.2.4. Subgroup analysis

Subgroup analysis (also referred to as moderator analysis) was conducted on the full set of studies (i.e., outlier (Shelden, 1999) included in the analysis) to investigate the moderating effect of three study-level characteristics. The analysis examined the effect of (1) research design – strong methodological designs (RCT or quasi-experimental designs with researcher-matched comparison groups) compared to studies that employed a quasi-experimental design with weakly matched control groups, (2) long follow-up periods (12 months or more) compared to short follow-up periods (less than 12 months), and (3) sample size of the treatment group (more than 100 participants versus less than 100 participants). As demonstrated in Table 6 (below), the analysis indicates that all of the variables examined here are significant moderator variables.

Table 6: Fixed effects moderator analysis for aftercare/re-entry, alleged offense

Study characteristic	Effect size and Q-statistics
Research design (N=13)	
Strong (RCT or QE matched) (N=7)	*LOR* = -0.087, z=0.55, p=0.584
Moderate (QE weakly matched) (N=6)	*LOR* =0.277 , z=2.88, p=0.004*
Between-studies heterogeneity	Q_B =3.84 ~χ^2_1, p=0.050*
Within-studies heterogeneity	Q_W = 39.36 ~χ^2_{11}, p<0.001
Follow-up period (N=12)	
Long (12 months or more) (N=9)	*LOR* =-0.136 , z=,1.34 p=0.180
Short (less than 12 months) (N=3)	*LOR* = 0.390, z=1.84, p=0.066
Between-studies heterogeneity	Q_B =5.01 ~χ^2_1, p=0.025*
Within-studies heterogeneity	Q_W = 8.84 ~χ^2_{10}, p=0.547
Treatment group sample size (N=13)	
Less than 100 (N=10)	*LOR* =-0.031 z=0.25, p=0.806
100+ (N=3)	*LOR* = 0.342, z=3.12, p=0.002*
Between-studies heterogeneity	Q_B = 5.04~χ^2_1, p=0.021*
Within-studies heterogeneity	Q_W = 38.16~χ^2_{11}, p<0.001

First, with respect to the moderating effect of research design, Table 6 demonstrates a statistically significant Q-between (Qb=3.84, p=.05), suggesting that research design is an important factor in moderating the results. From this, we can deduce that the effect sizes produced by strong research designs are significantly different than the effect sizes produced by moderately strong research designs, and this difference is due to more than chance alone. The statistically significant logged odds ratio (LOR) for moderate research designs (LOR=.277, z=2.88, p<.05) is indicative of weakly matched designs being related to stronger and statistically significant effects (or conversely, strong research designs related to negative effects). This is consistent with the literature, which suggests that weak methodology is related to the overestimation of effects (Weisburd, 2000, 2010). This moderator analysis confirms that whenever possible, methodologically rigorous research designs should be used.

Second, a statistically significant Q-between for follow-up period (Qb=5.01, p<.05) suggests that length of follow-up period is an important moderator of effects, and that short follow-up periods (less than 12 months) are related to stronger effects (LOR=.390, z=1.84, p=.066), perhaps suggesting that, over time, treatment effects dissipate.

Finally, a significant Q-between statistic for treatment group sample size (Qb=5.34, z=3.12, p<.05) illustrates that sample size is a significant moderator of treatment effects. The statistically significant LOR for large sample sizes (100 or more participants) demonstrates that large sample sizes are related to stronger and statistically significant effects (LOR=.342, z=3.12, p<.05). This makes sense because larger sample sizes have more statistical power to detect an effect when there is one present.

4.2.3. Outcome 2: The effect of aftercare/re-entry on convicted offenses

4.2.3.1. Pooled effect

Table 7 demonstrates that the pooled effect for aftercare/re-entry on convicted offenses (adjudication, conviction, incarceration) (N=11) is negative and non-significant (LOR=-.029, z=.27, p=.784), suggesting that there is no difference between the aftercare and comparison groups when it comes to convicted offenses. The non-significant Q-statistic demonstrates a fairly homogeneous sample (14.16, df=10, p=.166). Consistently, the I^2 statistic shows that 29.4% of the heterogeneity that is present can be attributed to factors beyond sampling error (chance).

Table 7: Fixed effects meta-analysis for aftercare/re-entry, convicted offenses

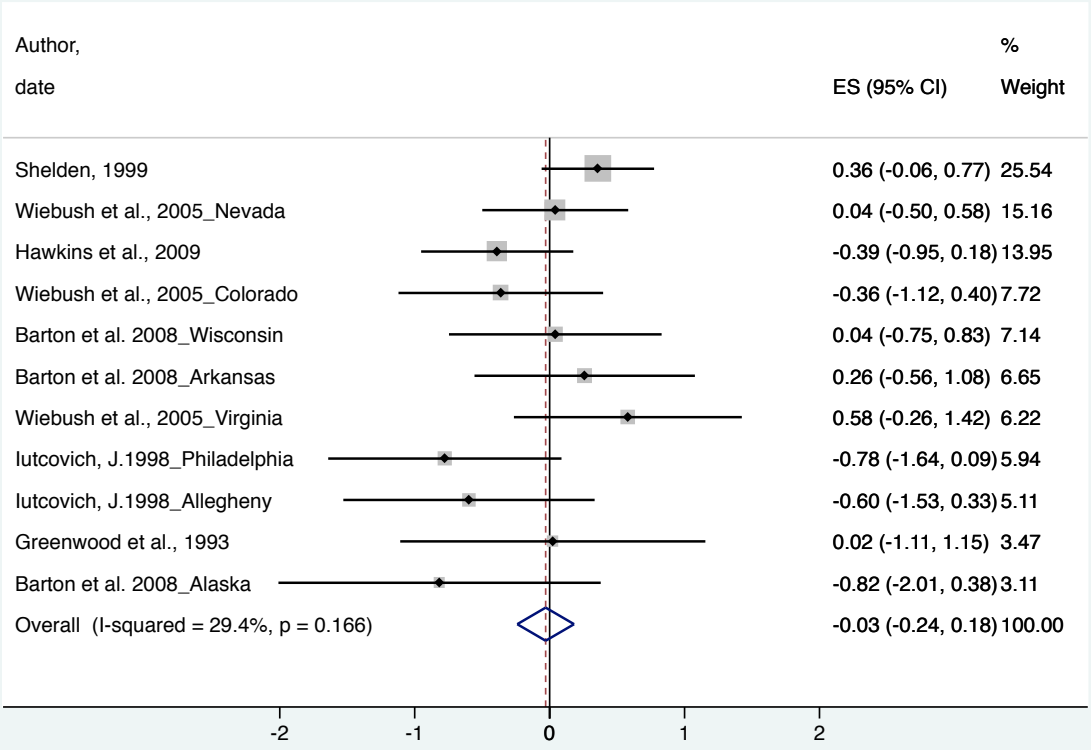
Pooled estimate	95% CI Lower	95% CI Upper	Z	No. of studies	Q-statistic	I^2
-0.029	-0.240	0.181	0.27, p=.784	11	14.16, df=10, p=.166	29.4%

Table 8: Study-level data fixed effect model meta-analysis for aftercare/re-entry, convicted offenses

Author, year	Effect size (LOR)	Standard error (SE)	95% CI	Relative weight
Shelden, 1999	0.357	.212	-.059, .774	25.54
Wiebush et al., 2005_Nevada	0.041	.275	-.499, .582	15.16
Hawkins et al., 2009	-0.389	.287	-.952, .175	13.95
Wiebush et al., 2005_Colorado	-0.362	.386	-1.12, .396	7.72
Barton et al., 2008_Wisconsin	0.042	.402	-.746, .830	7.14
Barton et al., 2008_Arkansas	0.260	.416	-.557, 1.07	6.65
Wiebush et al., 2005_Virginia	0.580	.430	-.265, 1.42	6.22
Iutcovich & Pratt, 1998_Philadelphia	-0.776	.441	-1.64, 0.08	5.94
Iutcovich & Pratt, 1998_Allegheny	-0.598	.475	-1.52, .333	5.11
Greenwood et al., 1993	0.023	.576	-1.10, 1.15	3.47
Barton et al., 2008_Alaska	-0.816	.609	--2.01, .379	3.11

Again, the forest plot (Figure 6) illustrates a visual representation of the data presented in Tables 7 and 8. The plot demonstrates that 6 studies favoured the treatment group (i.e., aftercare produced a desired effect of reduced recidivism in the treatment group) and 5 studies favoured the control group, however, these effects were very small. Logged odds ratios ranged from -0.816 to 0.580.

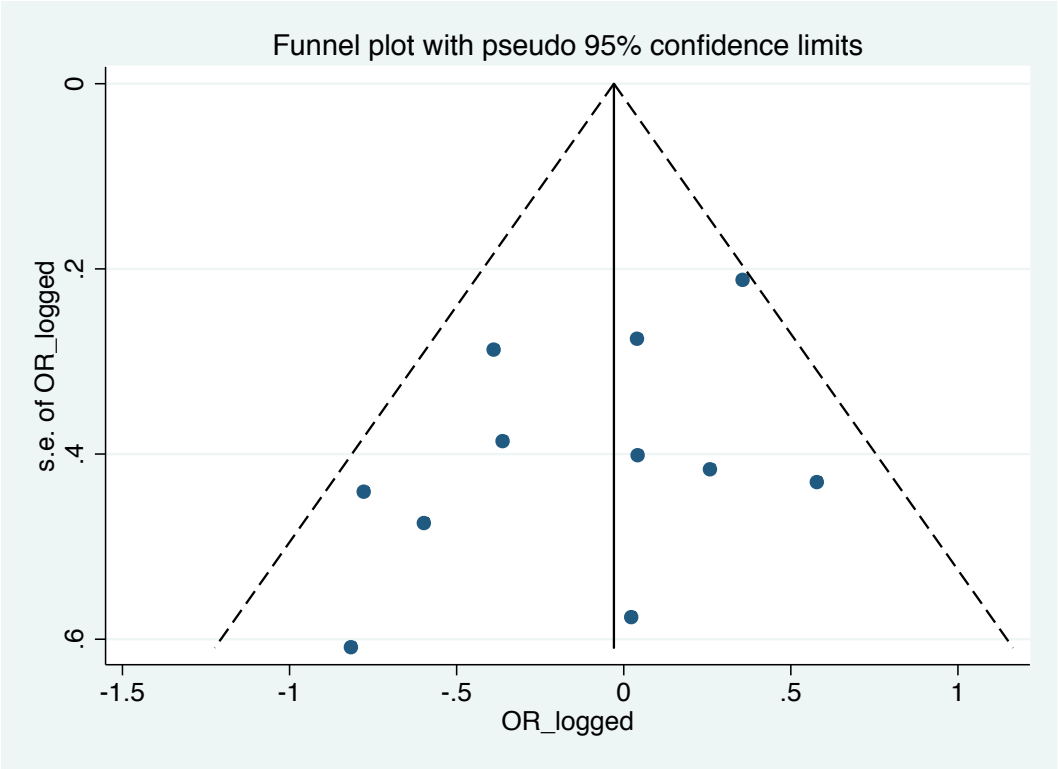
Figure 6: Forest plot for aftercare/re-entry, convicted offenses



4.2.3.2. Sensitivity analysis

Consistent with the absence of extreme effects in the forest plot above, no outlier effects are identified in the funnel plot. Furthermore, the funnel plot is symmetrical, indicating that publication bias is not a threat.

Figure 7: Funnel plot for aftercare/re-entry, convicted offenses

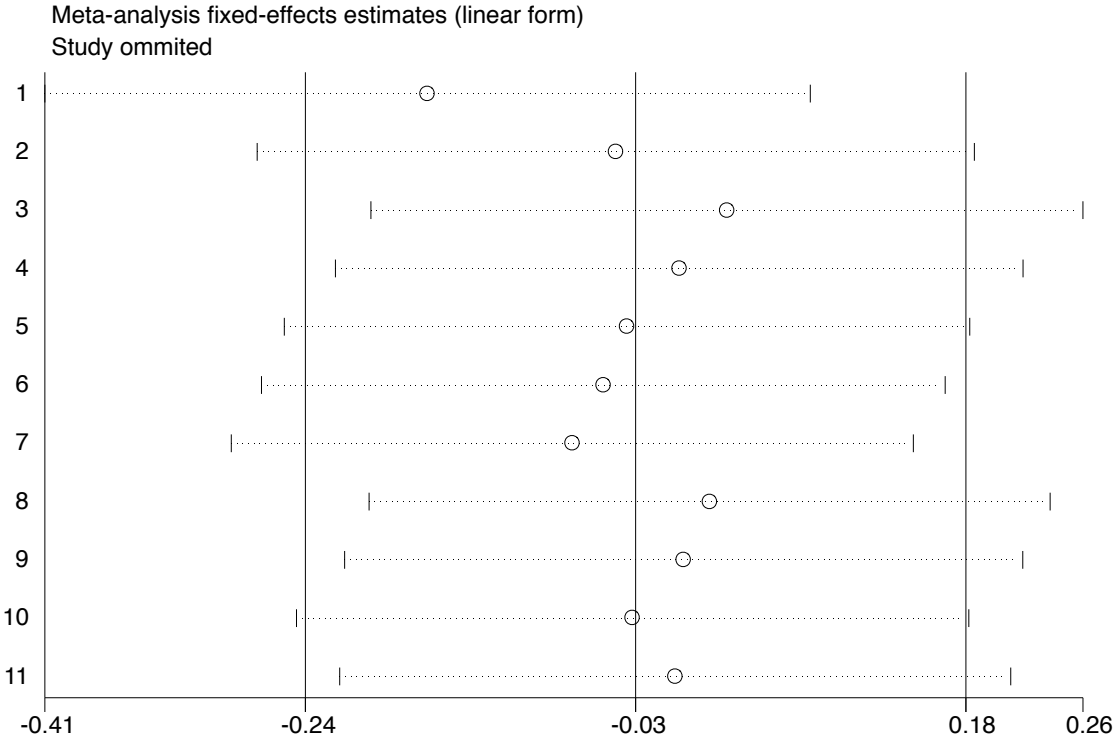


As in the previous analysis on alleged offenses, the influence analysis for convicted offenses (Table 9 and Figure 8) demonstrates that Shelden (1999) carries considerable influence on the overall pooled effect. Due to weighting each study by inverse variance, the effect size produced by Shelden (1999) exerts substantial influence. Consistently, Table 9 shows that if Shelden (1999) were omitted from the analysis, the overall pooled effect would shift to a coefficient of -0.162, substantially affecting the pooled estimate of aftercare/re-entry programs on convicted offenses.

Table 9: Influence analysis for aftercare/re-entry, convicted offenses

Study omitted	Effect size (LOR)	95% CI
Shelden, 1999	-.162	-.406, .081
Wiebush et al., 2005_Nevada	-.042	-.270, .186
Hawkins et al., 2009	.028	-.198, .255
Wiebush et al., 2005_Colorado	-.001	-.220, .217
Barton et al., 2008_Wisconsin	-.034	-.253, .183
Barton et al., 2008_Arkansas	-.050	-.267, .167
Wiebush et al., 2005_Virginia	-.069	-.287, .147
Iutcovich & Pratt, 1998_Philadelphia	.017	-.199, .234
Iutcovich & Pratt, 1998_Allegheny	.001	-.214, .217
Greenwood et al., 1993	-.031	-.245, .183
Barton et al., 2008_Alaska	-.004	-.218, .209
Combined	-.029	-.239, .181

Figure 8: Influence analysis for aftercare/re-entry, convicted offenses



4.2.3.3. Alternative analyses: Random effects model and addressing outliers

As illustrated in Figure 7, there were no outliers in the pooled analysis. Furthermore, as demonstrated in Table 10, the differences between the fixed effects model and the random effects model are very small (i.e., the pooled estimate is still considered a 'small' effect, it does not change in significance, nor in directionality).

Table 10: Comparative meta-analyses for aftercare/re-entry programs, convicted offenses

Model	Pooled estimate	*Z*	No. of studies	*Q*-statistic	I^2
Fixed effects	-0.029	0.27, p=.784	11	14.16, df=10, p=.166	29.4%
Random effects	-0.074	0.55, p=.585	11	14.16, df=10, p=.166	29.4%

*no outliers

The influential effect of Shelden (1999) on the overall estimate (as demonstrated in Figure 8) is worthy of consideration. Together, a study with a heavy weighting combined with a large effect size has the ability to bias the overall effect. To explain further, if an influence analysis was not conducted, conclusions drawn from the analysis would suggest that the effect of aftercare/re-entry programs on convictions and incarceration is slightly negative or null (i.e., no effect), however, the influence analysis provides insight towards the susceptibility to change and whether the findings are robust.

4.2.3.4. Subgroup analysis

Table 11: Fixed effects moderator analysis for aftercare/re-entry, convicted offenses

Study characteristic	Effect size and Q-statistics
Research design (N=11)	
Strong (RCT or QE matched) (N=8)	*LOR* = -0.069, z=0.51, p=0.607
Moderate (QE weakly matched) (N=3)	*LOR* =0.040 , z=0.22, p=0.822
Between-studies heterogeneity	Q_B =0.24 ~χ^2_1, p=0.624
Within-studies heterogeneity	Q_W = 13.92 ~χ^2_9, p=0.125
Follow-up period (N=10)	
Long (more than 12 months) (N=3)	*LOR* = -0.523 , z=2.44, p=0.015*
Short (12 months) (N=7)	*LOR* =0.021 , z=0.13, p=0.893
Between-studies heterogeneity	Q_B =4.26 ~χ^2_1, p=0.039*
Within-studies heterogeneity	Q_W = 5.45 ~χ^2_8, p=0.709
Treatment group sample size (N=11)	
Less than 100 (N=8)	*LOR* =-0.161 z=1.01, p=0.314
100+ (N=3)	*LOR* =0.079 , z=0.55, p=0.585
Between-studies heterogeneity	Q_B =1.24 ~χ^2_1, p=0.265
Within-studies heterogeneity	Q_W = 12.92~χ^2_9, p=0.166

The subgroup analysis shown above (Table 11) investigated the effect of (1) research design, (2) follow-up period, and (3) treatment group sample size as potential moderators variables. Length of follow-up period was the only study characteristic shown to be a significant moderating variable. The statistically significant Q-between (Qb=4.26, p<.05) indicates that the effect sizes produced from studies with longer follow-up periods are significantly different than studies with shorter follow-up periods, and that this difference is due to more than chance alone. The significant p-value for the long follow-up periods (more than 12 months) (LOR=-.523, z=2.44, p<.05) demonstrates that longer follow-ups are related to negative treatment effects. This finding is consistent with the moderator analysis conducted for alleged offenses, and is intuitively interesting because it suggests that the impact of aftercare/re-entry programs may not have a lasting effect over time when it comes to reducing recidivism.

4.3. Intensive supervision probation (ISP)

4.3.1. Characteristics of included studies

This section presents the characteristics of intensive supervision probation programs, followed by the pooled analyses for alleged and convicted offenses. Table 12 highlights the general characteristics of the included studies by outcome measure. The descriptive section is followed by pooled analyses, sensitivity analyses and subgroup analyses for each outcome measure.

Overview of included studies

A total of 26 individual sites (from 15 studies) were selected for inclusion, and contributed 31 effect sizes. As previously discussed, studies were separated based on outcome measure (alleged offenses or convicted offenses). Again, the effect sizes for these 26 sites are not independent of each other as multiple sites reported on both alleged offenses and convicted offenses (N=5). Twelve sites reported on alleged offenses, and 19 sites reported on convicted offenses, for a total of 31 effect sizes. Furthermore, as opposed to the gender distribution in aftercare/re-entry programs, 100% of ISP studies were composed of samples that were mixed (included both males and females). Table 12 offers a side-by-side comparison of characteristics of included studies for intensive supervision probation programs and alleged and convicted offenses.

Table 12: Characteristics of included studies for intensive supervision probation

Study characteristics	N (%)	N (%)
	Alleged offenses (N=12)	Convicted offenses (N=19)
Publication year		
1990-1994	2 (16.67%)	2 (10.53%)
1995-1999	1 (8.33%)	0 (0.0%)
2000-2004	6 (50.0%)	14 (73.68%)
2005-2009	3 (25.0%)	3 (15.79%)
Publication type		
Journal article	6 (50.0%)	4 (21.05%)
Report	6 (50.0%)	15 (78.95%)
Program delivery year		
1980-1989	1 (9.09%)	2 (10.53%)
1990-1999	6 (54.55%)	13 (68.42%)
2000-2010	4 (36.36%)	4 (21.05%)
Location		
United States	11 (91.67%)	16 (84.21%)
England	1 (8.33%)	3 (15.79%)
Follow-up period [a]		
<1 yr	3 (25.0%)	0 (0.0%)
1 yr	5 (41.67%)	4 (22.22%)
>1 yr	4 (33.33%)	14 (77.78%)
Type of research design		
Randomized control trial	10 (83.33%)	15 (78.95%)
QE (matched comparison)	1 (8.33%)	3 (15.79%)
QE (weakly matched)	1 (8.33%)	1 (5.26%)
Sample gender mix [b]		
All males	0 (0.0%)	0 (0.0%)
Mixed	10 (100.0%)	7 (100.0%)
Sample race/ethnicity [c]		
Caucasian/mixed	4 (40.0%)	5 (71.43%)
Predominantly minority	6 (60.0%)	2 (28.57%)
Sample size in treatment group		
Less than 100	5 (41.67%)	9 (47.37%)
100+	7 (58.33%)	10 (52.63%)

[a]Convicted offenses: 1 missing
[b]Alleged offenses: 2 missing; convicted offenses: 12 missing
[c]Alleged offenses: 2 missing; convicted offenses: 12 missing

Publication

Alleged offenses

The publication type of included studies was evenly distributed between journal articles (50%) and technical reports (50%). With respect to year of publication, interestingly, 75% of studies were published between 2000 and 2009; however no studies were identified post-2009. Similar to aftercare/re-entry, the majority of selected studies were conducted in the United States (N=11, 92%), however 1 study (8%) was conducted in England.

Convicted offenses

The majority of studies were published as technical reports (N=15; 79%), while 4 (21%) studies were published as journal articles. Similar to the studies on alleged offenses, the majority of studies were published between 2000 and 2009 (89%), and the large majority of studies (N=16; 84%) were conducted in the United States, while 3 studies (16%) were conducted in England.

Intervention

Similar to the efforts made for coding the aftercare/re-entry programs, information on intervention-specific characteristics was unreliably and inconsistently reported and the only intervention characteristic coded consistently enough to be included for analysis was 'program delivery year'. Interventions were most commonly delivered in the 1990s, with relatively few programs delivered in the 1980s and post-2000.

Sample

As with aftercare/re-entry programs, ethnicity and treatment group sample size were dichotomized. With respect to treatment group sample size, the distribution of treatment was near equivalent (between samples of more than 100 participants or less than 100 participants) for both alleged offenses and convicted offenses. It is important to note that sample size for each individual study varied substantially and ranged between 41 and 1108 participants. Consistent with the studies evaluating aftercare/re-entry programs, samples of studies that measured alleged offenses were primarily composed of ethnic

minority youth (60%), however, for the studies that evaluated convicted offenses, the majority used samples that were either primarily Caucasian or near equivalent in ethnic composition (71%).

Design

Selection criteria in Chapter 3 specified that studies evaluating intensive supervision probation programs were required to have standard/traditional probation as the comparison group. As a result, one study (Barton & Butts, 1990) was excluded because the comparison group was composed of youth who were incarcerated.

Interestingly, the vast majority of studies (for both alleged offenses and convicted offenses) used a randomized controlled trial (83% and 79% respectively), while only a small percentage employed quasi-experimental designs.

Outcomes

The majority of the studies used relatively long follow-up periods; 75% of studies that measured alleged offenses used a follow-up that was at least 12 months (12 months (42%) and more than 12 months (33%)). Similarly, for studies that measured convicted offenses, all of the studies used at least 12 months for follow-up: 14 studies reported outcome data from a follow-up period longer than 12 months (78%) and 4 studies (22%) used a follow-up of exactly 12 months. No studies had a follow up of less than one year.

4.3.2. Outcome 1: The effect of intensive supervision probation on alleged offenses

4.3.2.1. Pooled effect

Intensive supervision probation programs that presented outcome results on alleged offenses, i.e., criminal contact, charges, arrests, etc. (N=12) were pooled together. Overall, the estimate for the fixed effects model yielded a pooled effect of 0.088 (z=1.25, p=.210), a very small and non-significant effect. This finding suggests that there are no significant differences between youth in the treatment and comparison groups with respect to alleged offenses. The significant Q-statistic presented in Table 12

demonstrates a significantly heterogeneous sample (30.86, df=11, p=.001) and the I^2 statistic indicates that 64.4% of this heterogeneity can be attributed to factors beyond sampling error. Table 14 shows the effect size, standard error, confidence intervals and relative weight for each individual study.

Table 13: Fixed effects meta-analysis for intensive supervision probation, alleged offenses

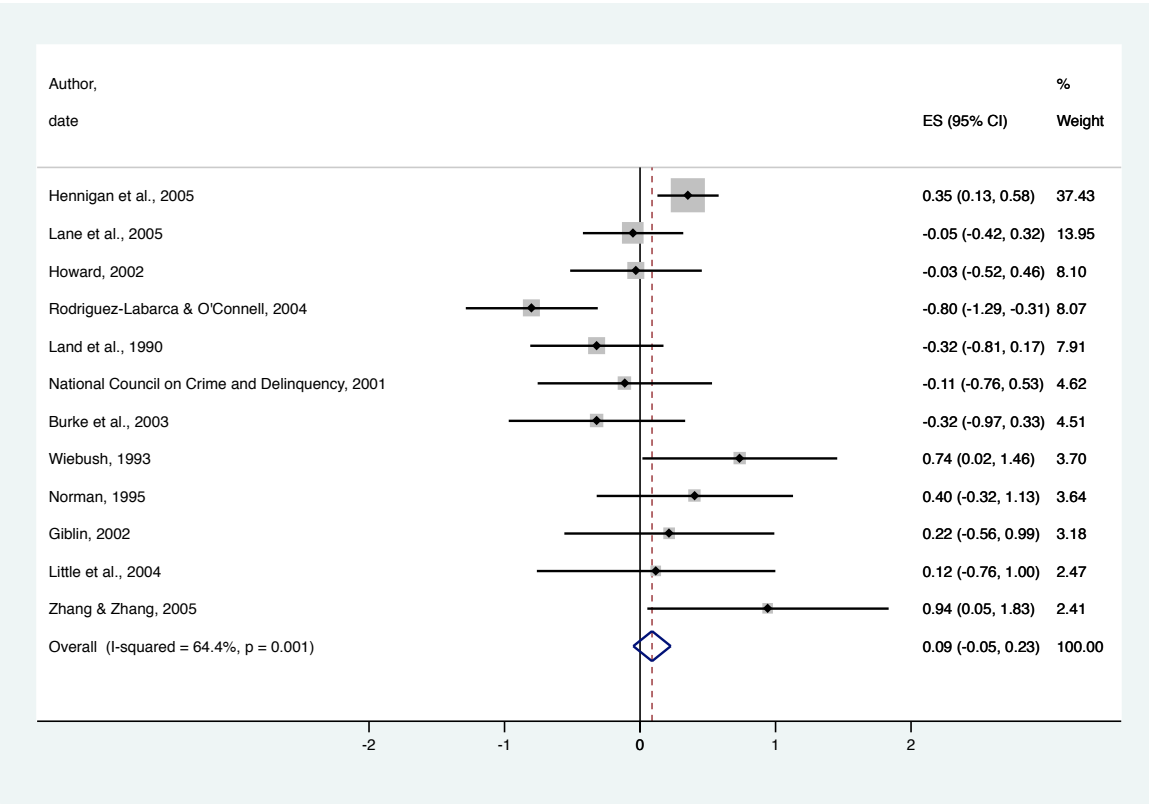
Pooled estimate	95% CI Lower	95% CI Upper	*Z*	No. of studies	*Q*-statistic	I^2
0.088	-0.050	0.227	1.25, p=.210	12	30.86, df=11, p=.001	64.4%

Table 14: Study-level data fixed effects meta-analysis for intensive supervision probation, alleged offenses

Author, year	Effect size (LOR)	Standard error (SE)	95% CI	Relative weight
Hennigan et al., 2005	0.354	.115	.128, .580	37.43
Lane et al., 2005	-0.051	.188	-.421, .319	13.95
Howard, 2002	-0.030	.247	-.516, .456	8.10
Rodriguez-Labarca & O'Connell, 2004	-0.799	.248	-1.28, -.312	8.07
Land et al., 1990	-0.319	.250	-.810, .173	7.91
National Council on Crime and Delinquency, 2001	-0.112	.328	-.755, .531	4.62
Burke et al., 2003	-0.318	.332	-.969, .333	4.51
Wiebush, 1993	0.736	.366	.018, 1.45	3.70
Norman, 1995	0.405	.369	-.320, 1.12	3.64
Giblin, 2002	0.216	.395	.558, .991	3.18
Little et al., 2004	0.119	.449	-.761, .999	2.47
Zhang & Zhang, 2005	0.944	.454	.054, 1.83	2.41

Again, the forest plot below illustrates the information provided in Tables 13 and 14. Here, we can see that the pooled logged odds ratio (shown by the diamond and broken vertical line) favours the treatment group; a small but positive effect. Moreover, 6 studies favoured the treatment group, and 6 studies favoured the control group. Finally, across studies, logged odds ratios ranged from -1.238 to 1.093, demonstrating considerable heterogeneity and possible outlier effects.

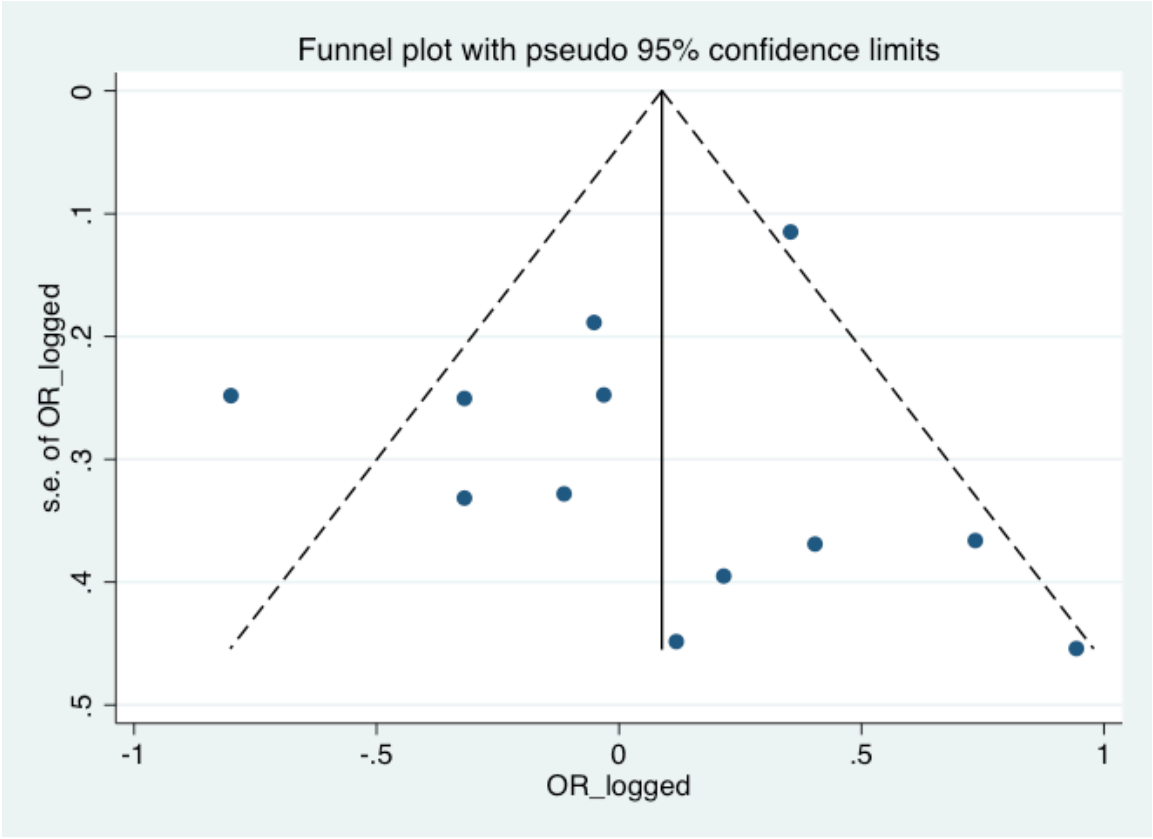
Figure 9: Forest plot for intensive supervision probation, alleged offenses



4.3.2.2. Sensitivity analysis

A sensitivity analysis was conducted to investigate the presence of extreme outlier effects that potentially bias the pooled estimate. The funnel plot below (Figure 10) identifies two outlier studies (two data points that fall outside of the pseudo 95% confidence interval); these were identified as the studies by Hennigan et al. (2005) and Rodgriquez-Labarca and O'Connell (2004).

Figure 10: Funnel plot for intensive supervision probation, alleged offenses



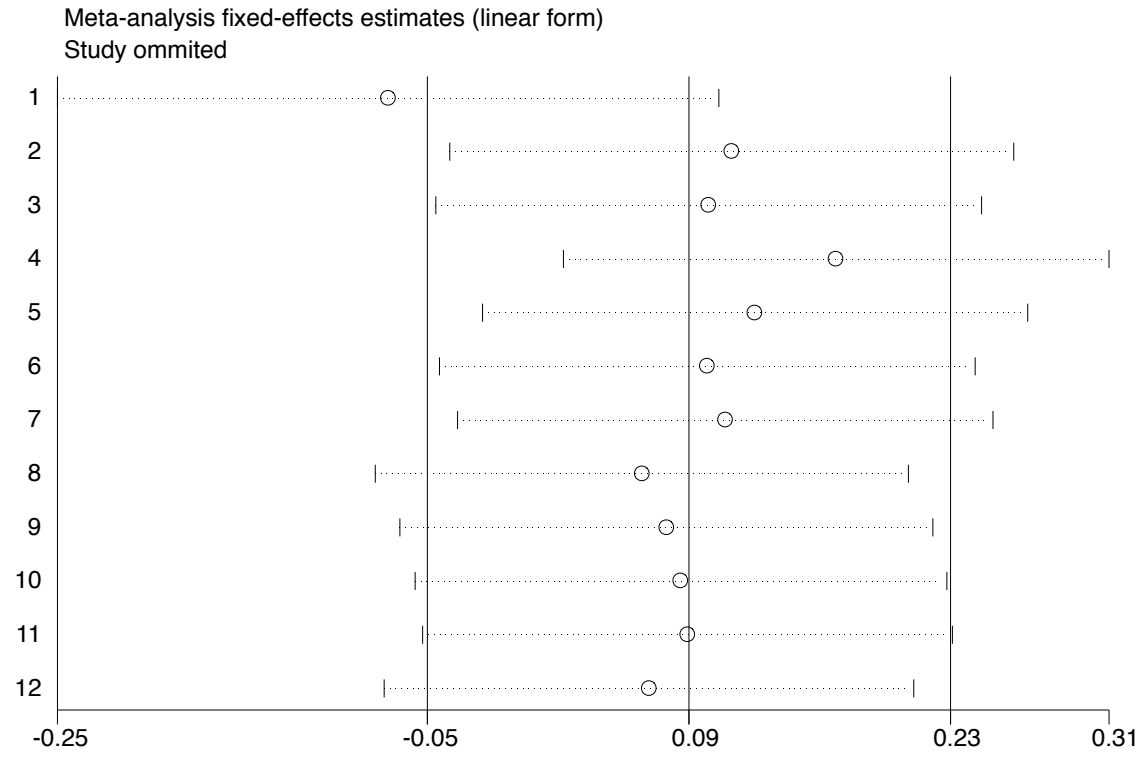
A second sensitivity analysis (influence analysis) was conducted to examine the influence of the outlier studies on the overall pooled effect. In an iterative fashion, the influence analysis removes each study one by one and recalculates what the new pooled effect would be if that single study were to be omitted.

Table 15: Influence analysis for intensive supervision probation, alleged offenses

Study omitted	Effect size (LOR)	95% CI
Hennigan et al., 2005	-.070	-.245, .104
Lane et al., 2005	.111	-.037, .260
Howard, 2002	.098	-.045, .243
Rodriguez-Labarca & O'Connell, 2004	.166	.022, .310
Land et al., 1990	.123	-.020, .267
National Council on Crime and Delinquency, 2001	.098	-.043, .239
Burke et al., 2003	.107	-.033, .249
Wiebush, 1993	.063	-.077, .204
Norman, 1995	.076	-.064, .217
Giblin, 2002	.084	-.056, .224
Little et al., 2004	.087	-.052, .227
Zhang & Zhang, 2005	.067	-.072, .207
Combined	.088	-.049, .226

The influence analysis conducted here illustrates the importance of considering both size of effect, as well as relative weight, when analyzing outlier effects. To explain further, we can see in Figure 10 that there are two outliers; the slight outlier (LOR=.35, SE=.16) corresponds to Hennigan et al. (2005) while the more obvious outlier corresponds to Rodriguez-Labarca and O'Connell (2004) (LOR= -.799, SE=.25). However, while Rodriguez-Labarca and O'Connell (2004) is the largest outlier it terms of effect size, as can be seen in Figure 11 below, the weighting assigned to Hennigan et al. (2005) (37.43%) makes this study a substantially more influential study when it comes to biasing the overall pooled effect. Moreover, when Hennigan et al. (2005) is removed from the analysis, the pooled estimate shifts from 0.088 to -0.071, further demonstrating the importance of considering both size of effect and weighting when analyzing outliers.

Figure 11: Influence analysis for intensive supervision probation, alleged offenses



4.3.2.3. Alternative analyses: Random effects model and addressing outliers

Pooled effect

To examine the effect of outliers beyond the influence analysis, the meta-analysis was re-conducted with Hennigan et al. (2005) and Rodriguez-Labarca and O'Connell (2004) omitted from the analysis (fixed effect with outliers removed), as well as with a random effects model. Table 16 provides a comparison of the pooled estimate, z-score, p-values, and heterogeneity statistics for the fixed effects model when outliers are included in the analysis and when they are removed (0.088, z=1.25, p=.210; 0.037, z=0.39, p=.186), as well as for the random effects model (0.047, z=0.35, p=.723). Notably, in comparison to the fixed effects model (with outliers) the pooled estimate for the random effects model is smaller, demonstrates a smaller z-score, as well as a substantially larger p-value. With respect to the three models presented here, there is

considerably less variation among the pooled estimates in comparison to previous analyses. Moreover, the pooled estimates consistently demonstrate positive and non-significant findings.

Table 16: Comparative meta-analyses for intensive supervision probation, alleged offenses

Model	Pooled estimate	*Z*	No. of studies	*Q*-statistic	I^2
Fixed effects (with outliers)	0.088	1.25, p=.210	12	30.86, df=11, p=.001	64.4%
Random effects (with outliers)	0.047	0.35, p=.723	12	30.86%, df=11, p=.001	64.4%
Fixed effects (with outliers removed)	0.037	0.39, p=.697	10	12.50, df=9, p=.186	28.0%

Table 16 shows that when Hennigan et al. (2005) and Rodriguez-Labarca and O'Connell (2004) are excluded from the pooled analysis, the overall estimate of effect changes slightly in magnitude and the pooled estimate remains non-significant. More interestingly, the Q-statistic changes to a non-significant value (12.50, df=9, p=.186) and the I^2 statistic changes to 28.0%. This finding suggests that the new set of studies is more homogeneous and there is less variation in effect size across studies. See Appendix B for tables and figures from the full analysis with outliers omitted.

4.3.2.4. Subgroup analysis

Table 17: Fixed effects moderator analysis for intensive supervision probation, alleged offenses

Study characteristic	Effect size and Q-statistics
Publication type (N=12)	
Journal article (N=6)	*LOR* = 0.082, z=0.67, p=0.502
Report (N=6)	*LOR* = 0.092, z=1.06, p=0.289
Between-studies heterogeneity	Q_B = 0.0~χ^2_1, p=1.00
Within-studies heterogeneity	Q_W =30.86 ~χ^2_{10}, p<0.001
Follow-up period (N=12)	
Long (more than 12 months) (N=4)	*LOR* =-0.008 , z=0.06, p=0.954
Short (12 months or less) (N=8)	*LOR* = 0.126, z=1.51, p=0.130
Between-studies heterogeneity	Q_B = 0.72~χ^2_1, p=0.396
Within-studies heterogeneity	Q_W = 30.14 ~χ^2_{10}, p<0.001
Sample size in treatment group (N=12)	
Less than 100 (N=5)	*LOR* = 0.127 z=0.82, p=0.410
100+ (N=7)	*LOR* =0.078 , z=0.99, p=0.324
Between-studies heterogeneity	Q_B =0.08 ~χ^2_1, p=0.777
Within-studies heterogeneity	Q_W =30.78 ~χ^2_{10}, p<0.001

Table 17 shows the results of the fixed effects moderator analysis for the full set of intensive supervision probation (alleged offenses) studies (i.e., outlier studies Hennigan et al. (2005) and Rodriguez-Labarca and O'Connell (2004) are included). Subgroup analysis was performed to examine publication type (journal article or report), follow-up period (long versus short), and treatment group sample size (less than 100 versus more than 100) as possible variables that moderate results. For each of the variables examined, the Q-between statistic was non-significant, suggesting that the identified study variables do not play a role in moderating results. As the literature on publication bias states that positive and significant results are more likely to be published in peer-reviewed articles, the detection of no difference is somewhat surprising, however, it is noteworthy that the technical reports used exceptionally rigorous research designs (83.33% were randomized controlled trials). Reasons for why these methodologically rigorous research studies were not published in peer-reviewed journals should be examined further.

4.3.3. Outcome 2: The effect of intensive supervision probation on convicted offenses

4.3.3.1. Pooled effect

Intensive supervision probation programs that presented results for convicted offenses, i.e., adjudication, convictions, incarceration etc. (N=19) were pooled together. The overall estimate for the fixed effects model yielded a pooled effect of -.117 (z=1.64, p=1.01). Similar to the pooled analysis for alleged offenses, the pooled analysis for convicted offenses suggests that there is no difference between the treatment and control groups. The significant Q-statistic presented in Table 18 demonstrates a significantly heterogeneous sample (43.17, df=18, p=.001) and the I^2 statistic indicates that 58.3% of the heterogeneity can be attributed to factors beyond sampling error. Table 19 shows the effect size, standard error, confidence intervals and relative weight for each individual study.

Table 18: Fixed effects meta-analysis for intensive supervision probation, convicted offenses

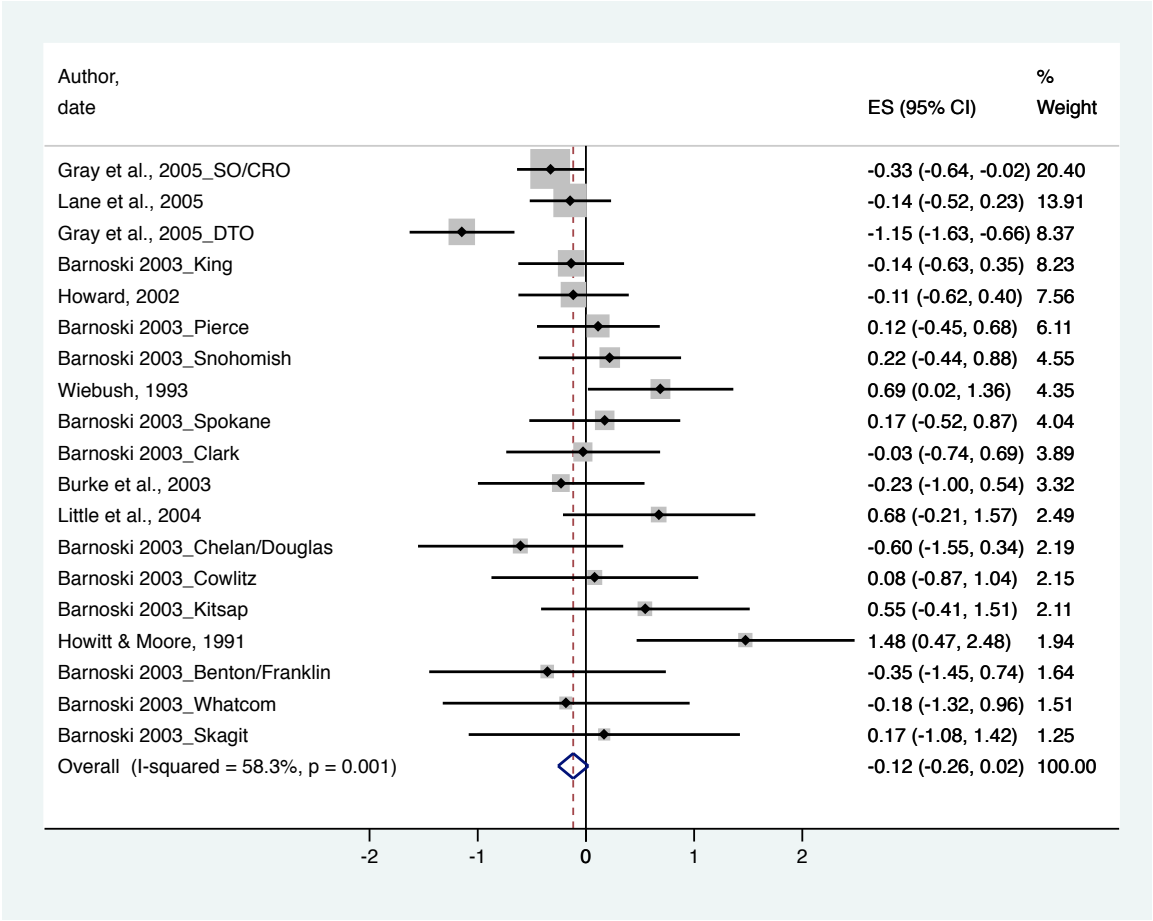
Pooled estimate	95% CI Lower	95% CI Upper	*Z*	No. of studies	Q-statistic	I^2
-0.117	-0.258	0.023	1.64, p=.101	19	43.17, df=18, p=.001	58.3%

Table 19: Study-level data fixed effects meta-analysis for intensive supervision probation, convicted offenses

Author, year	Effect size (LOR)	Standard error (SE)	95% CI	Relative weight
Gray et al., 2005_SO/CRO	-0.326	.158	-.636, -.016	20.40
Lane et al., 2005	-0.143	.191	-.519, .233	13.91
Gray et al., 2005_DTO	-1.145	.247	-1.63, -.661	8.37
Barnoski, 2003_King	-0.136	.249	-.625, .353	8.23
Howard, 2002	-0.114	.260	-.624, .396	7.56
Barnoski, 2003_Pierce	0.115	.289	-.452, .683	6.11
Barnoski, 2003_Snohomish	0.222	.335	-.435, .879	4.55
Wiebush, 1993	0.690	.342	.018, 1.36	4.35
Barnoski, 2003_Spokane	0.174	.355	-.523, .872	4.04
Barnoski, 2003_Clark	-0.025	.362	-.736, .686	3.89
Burke et al., 2003	-0.228	.392	-.998, .542	3.32
Little et al., 2004	0.677	.453	-.212, 1.56	2.49
Barnoski, 2003_Chelan/Douglas	-0.604	.483	-1.52, .345	2.19
Barnoski, 2003_Cowlitz	0.082	.487	-.873, 1.03	2.15
Barnoski, 2003_Kitsap	0.550	.492	-.415, 1.51	2.11
Howitt & Moore, 1991	1.476	.513	.469, 2.48	1.94
Barnoski, 2003_Benton/Franklin	-0.355	.557	-1.44, .739	1.64
Barnoski, 2003_Whatcom	-0.182	.582	-1.32, .959	1.51
Barnoski, 2003_Skagit	0.170	.639	-1.08, 1.42	1.25

Figure 12 (below) shows that for the entire set of studies, the logged odds ratios were considerably heterogeneous and ranged from -1.145 to 1.476. The plot further demonstrates inconsistent treatment effectiveness; 9 studies favoured the treatment group, while 10 studies favoured the control group, and few of these reached statistical significance.

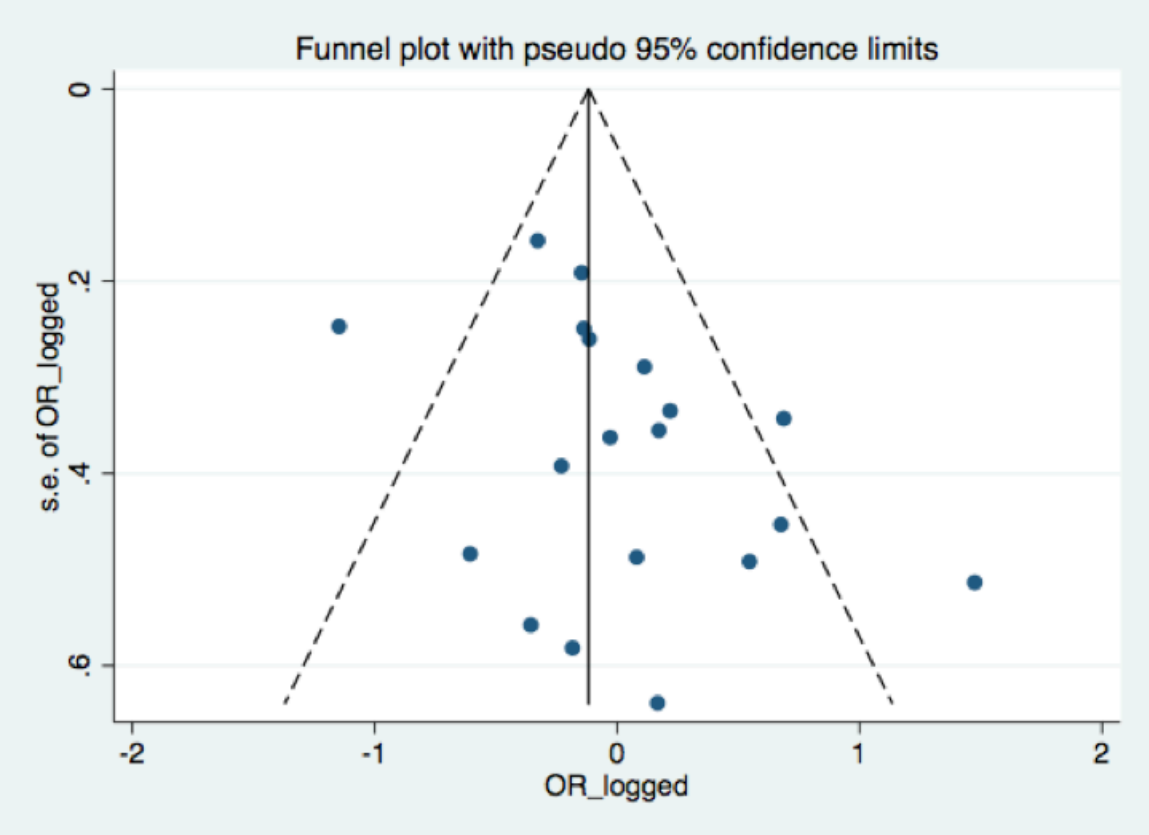
Figure 12: Forest plot for intensive supervision probation, convicted offenses



4.3.3.2. Sensitivity analysis

As indicated by the forest plot above, some of the plotted data points (effect sizes) extend above or below the pooled estimate, suggesting that some of the data points may be significant outliers and bias the overall mean effect. Visual inspection of the funnel plot in Figure 13 demonstrates that three studies fall outside of the pseudo 95% confidence interval and can therefore be considered outlier effects that possibly distort the pooled estimate.

Figure 13: Funnel plot for intensive supervision probation, convicted offenses



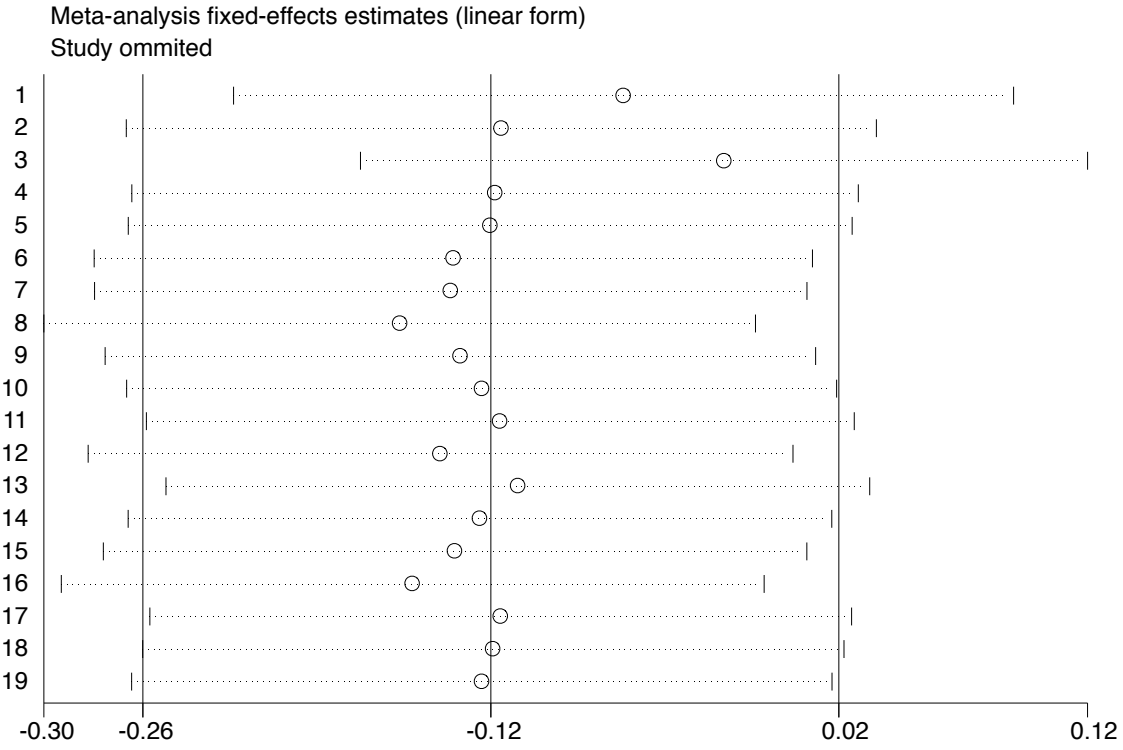
The three outliers were identified as Gray et al._DTO (2005), Howitt and Moore (1991), and Wiebush (1993). As the forest plot provides indication of the presence of outlier effects, the influence analysis provides insight to the degree of influence exerted by each outlier.

Table 20: Influence analysis for intensive supervision, convicted offenses

Study omitted	Effect size (LOR)	95% CI
Gray et al., 2005_SO/CRO	-.063	-.220, .093
Lane et al., 2005	-.113	-.264, .038
Gray et al., 2005_DTO	-.023	-.169, .123
Barnoski 2003_King	-.115	-.261, .030
Howard, 2002	-.117	-.263, .028
Barnoski 2003_Pierce	-.132	-.277, .012
Barnoski 2003_Snohomish	-.133	-.276, .010
Wiebush, 1993	-.154	-.297, -.010
Barnoski, 2003_Spokane	-.129	-.272, .013
Barnoski, 2003_Clark	-.121	-.264, 0.21
Burke et al., 2003	-.113	-.256, 0.29
Little et al., 2004	-.137	-.279, .004
Barnoski, 2003_Chelan/Douglas	-.106	-.248, .035
Barnoski, 2003_Cowlitz	-.121	-.263, .020
Barnoski, 2003_Kitsap	-.131	-.273, .009
Howitt & Moore, 1991	-.148	-.290, -.007
Barnoski, 2003_Benton/Franklin	-.113	-.254, .028
Barnoski, 2003_Whatcom	-.116	-.257, .024
Barnoski, 2003_Skagit	-.120	-.262, .020
Combined	-.117	-.257, .022

Graphically, we can see in Figure 14 that study number 1 (corresponding with the Gray et al._SO/CRO, 2005) and study number 3 (Gray et al._DTO, 2005) carry considerable influence on the overall pooled effect. That is, if either of the studies were omitted from the analysis, the pooled estimate would change substantially. Consistently, Table 20 shows that by omitting Gray et al._DTO (2005) DTO, the pooled effect would change to -0.023 and omitting Gray et al._SO/CRO (2005) would change the pooled effect to -0.064, a considerable departure from the inclusive pooled effect of -0.117.

Figure 14: Influence analysis for intensive supervision probation, convicted offenses



4.3.3.3. Alternative analyses: Random effects model and addressing outliers

Pooled effect

To examine the influence of outlier effects further, the analysis was re-conducted without the three outliers identified in Figure 13 (i.e., Gray et al._DTO, 2005, Howitt & Moore, 1991, and Wiebush, 1993), as well as with a random effects model. Table 21 demonstrates a comparison of the fixed effects model meta-analysis, with and without outliers (-0.117, z=1.64, p=.101; -0.094, z=1.21, p=.226). Although the pooled estimates change only slightly, it is worth noting that the Q-statistic changes to a non-significant value (p=.778) and the I^2 statistic changes to 0.0%, indicating that the new set of studies is homogeneous and that any differences between the remaining studies can be attributed to chance or random noise. See Appendix C for tables and figures from the full analysis with outliers omitted. Similar to above, the random effects (with outliers) presents a substantially smaller pooled estimate than the fixed effects model (with

outliers) (-0.015, z=0.12, p=.001), again suggesting that when the number of studies being pooled together is small, the random effects model underestimates the pooled effect. However, it is worth noting that all three models consistently demonstrate a negative and non-significant finding.

Table 21: Comparative meta-analyses for intensive supervision probation, convicted offenses

Model	Pooled estimate	*Z*	No. of studies	*Q*-statistic	I^2
Fixed effect (with outliers)	-0.117	1.64, p=.101	19	43.17, df=18, p=.001	58.3%
Random effect (with outliers)	-0.015	0.12, p=.904	19	43.17, df=18, p=.001	58.3%
Fixed effect (with outliers removed)	-0.094	1.21, p=.226	16	10.63, df=15, p=.778	0.0%

4.3.3.4. Subgroup analysis

Table 22 illustrates the importance of study characteristics in moderating the results for intensive supervision probation, convicted offenses (full set of studies; N=19). As seen below, the fixed effect moderator analysis demonstrated that all four of the identified variables (publication type, research design, follow-up period, and treatment group sample size) are important when it comes to significantly moderating results.

Table 22: Fixed effects moderator analysis for intensive supervision probation, convicted offenses

Study characteristic	Effect size and Q-statistics
Publication type (N=19)	
Journal article (N=4)	*LOR* = 0.245, z=1.63, p=0.103
Report (N=15)	*LOR* = -0.224 , z=2.75, p=0.006*
Between-studies heterogeneity	Q_B = 7.52 ~χ^2_1, p=0.006*
Within-studies heterogeneity	Q_W = 35.65 ~χ^2_{17}, p=0.005
Research design (N=19)	
Strong (RCT) (N=15)	*LOR* = -0.021 , z=0.24, p=0.814
Moderate (QE matched or QE weakly matched) (N=4)	*LOR* = -0.296, z=2.45, p=0.014*
Between-studies heterogeneity	Q_B =3.37 ~χ^2_1, p=0.066
Within-studies heterogeneity	Q_W = 39.80 ~χ^2_{17}, p=0.001
Follow-up period (N=18)	
Long (more than 12 months) (N=14)	*LOR* = 0.056, z=0.60, p=0.550
Short (12 months) (N=4)	*LOR* = -0.450, z=3.96, p=0.000*
Between-studies heterogeneity	Q_B =11.82 ~χ^2_1, p<0.001*
Within-studies heterogeneity	Q_W =21.54 ~χ^2_{16}, p=0.159
Sample size in treatment group (N=19)	
Less than 100 (N=9)	*LOR* = 0.352 z=2.18, p=0.029*
100+ (N=10)	*LOR* = -0.232, z=2.91, p=0.004*
Between-studies heterogeneity	Q_B =10.54 ~χ^2_1, p=0.001*
Within-studies heterogeneity	Q_W = 32.63~χ^2_{17}, p=0.013

The subgroup analysis for publication type demonstrates a significant Q-between statistic (Qb=7.52, p<.01). The statistically significant logged odds ratio for technical reports (LOR=-.224, z=2.75, p<0.01) suggests that reports are significantly related to negative effects (i.e. favouring the control group). Consulting Table 11 to make sense of this finding, we can see that for intensive supervision probation, convicted offenses, 15 of the studies were RCTs, 13 of which were published in technical reports. This makes sense because, as Durlak (2009) suggests, effect size estimates are directly related to research methodology; rigorous methodology is typically associated with smaller effect sizes, while weaker methodological designs overestimate effects. Considering the intuitiveness of this finding, however, it is surprising that the result for research design

(RCT versus quasi-experimental) is contradictory. The Q-between statistic for research design (Qb=3.37, p<0.1) suggests that quasi-experimental research designs are related to significantly negative effects (LOR=-.296, z=2.45, p<.05). That is, the direction and magnitude of an effect size can be tempered by whether the researchers used a randomized controlled design or a quasi-experimental design. Lipsey (2003) provides a plausible explanation that this finding might be due to the moderator variable's relationship with other moderator variables. To explain further, Shadish, Matt, Navaro, and Phillips (2000) suggest that, due to self-selection biases associated with quasi-experimental designs (which are largely absent in randomized controlled trials by virtue of the randomization of participants), the groups are not equivalent in quasi-experimental designs, which might therefore cause slightly biased effects and the relationship we observe through the subgroup analyses.

With respect to follow-up period, the analysis demonstrates that this variable is important in moderating results (Qb=11.82, p<.001) and that shorter follow-up periods are significantly related to estimates of negative effect sizes (LOR=-.450, z=3.96, p<0.001). Finally, treatment group sample size is also an important variable in moderating results (Qb=10.54, p=.001), in that studies with large sample sizes (more than 100 participants) have a larger effect than do smaller studies (LOR=-.232, z=2.91, p<.01). This is somewhat counterintuitive considering that it is expected that studies with larger samples have higher statistical power to detect effects. However, it is possible that the low statistical power associated with smaller samples is associated with overestimating the size of effect.

Chapter 5. Discussion

Altogether, community-based supervision programs are promising when it comes to reducing juvenile recidivism, as there is some evidence towards the deterring effect of supervision. However, the future success of these interventions is largely contingent on the findings of further research on the circumstances under which these interventions are effective at reducing recidivism. As a summary, the pooled analyses demonstrated varying degrees of program effectiveness, indicating that the direction and magnitude of community-based supervision-oriented programs effectiveness, overall, are not consistent and vary from site to site. Moreover, the pooled analyses yielded conflicting results with respect to the effect of treatment intervention on alleged offenses and convicted offenses. Most interestingly, however, a clear pattern emerged from the varying models and analyses (demonstrated in Table 23 below); the effect of the interventions (both aftercare/re-entry and ISP) on alleged offenses favoured the treatment group, while the effect of the interventions on convicted offenses favoured the treatment group. More specifically, the results suggest that in comparison to 'supervision as usual', youth receiving intensive supervision (either through intensive probation or aftercare/re-entry) are less likely to be arrested or charged for an offense, but are more likely to be convicted of an offense and/or incarcerated.

Table 23: Summary of pooled analyses

Intervention	Model	Pooled estimate	Z
Aftercare, alleged offenses	Fixed effects (with outliers)	0. 179	2.18, p=.029*
	Random effects (with outliers)	0.065	0.38, p=.075
	Fixed effects (with outliers removed)	-0.038	0.42, p=.678
Aftercare, convicted offenses	Fixed effects	-0.029	0.27, p=.784
	Random effects	-0.074	0.55, p=.585
ISP, alleged offenses	Fixed effects (with outliers)	0.088	1.25, p=.210
	Random effects (with outliers)	0.047	0.35, p=.723
	Fixed effects (with outliers removed)	0.037	0.39, p=.697
ISP, convicted offenses	Fixed effects (with outliers)	-0.117	1.64, p=.101
	Random effects (with outliers)	-0.015	0.12, p=.904
	Fixed effects (with outliers removed)	-0.094	1.21, p=.226

With respect to aftercare/re-entry programs, the pooled analysis for alleged offenses demonstrated an overall estimate of 0.179 (z=2.18, p=.029), suggesting that youth who participate in aftercare/re-entry are significantly less likely to be arrested or charged with an offense in comparison to youth who receive 'care as usual' upon release from incarceration. Conversely, the analysis for convicted offenses demonstrated a pooled estimate of -0.029 (z=0.27, p=.784), suggesting no differential treatment impact between the two groups on convicted offenses.

A similar pattern emerged in the pooled analyses for intensive supervision probation programs. The pooled estimate for alleged offenses demonstrated an overall effect of 0.088 (z=1.25, p=.210), whereas the pooled analysis for convicted offenses demonstrated an overall estimate of -0.117 (z=1.64, p=1.01), however these effects were not statistically significant. The finding that this pattern was consistent across intervention types is telling of the importance in making a distinction when it comes to recidivism as an outcome measure. Additionally, it provokes consideration toward the importance of not combining "apples and oranges" when it comes to recidivism measures. Future research should investigate this pattern further.

Considering the variability that is demonstrated between the pooled effects for each model, the results should be taken with caution as to which is the 'true' effect. First, with respect to the fixed effects model (with outliers), for aftercare/re-entry, the inclusion of Shelden (1999) creates the threat of a distorted pooled estimate. Second, although the argument can be made for the use of a random effects model, according to Guolo and Varin (2015) and Schulze (2007), due to the small number of studies included in the analysis, it is likely that the results from the random effects model is an imprecise measurement of the pooled effect. Third, the fixed effects model (with outlier removed) is heterogeneous, and the influence analysis shows that Shelden (1999) exerts considerable influence over the pooled effect, arguably warranting exclusion from the analysis. However, in light of the history of heterogeneous effects for aftercare/re-entry and ISP programs and the range of effect sizes demonstrated here, Card's (2011) argument for the importance of analyzing the full set of studies, is valid if the research wishes to summarize and capture the effect of the available literature.

The statistically significant difference between the treatment and control group for aftercare/re-entry programs on alleged offenses is an interesting finding. In light of the criticisms of intermediate sanctions discussed in section 1.1.3. and the concerns towards net-widening, this is a favourable result and provides insight towards the importance of aftercare/re-entry programs for youth. The results suggest that youth who receive post-incarceration intensive supervision (via aftercare/re-entry) are less likely to be arrested compared to those who do not receive such treatment. As such, the results suggest that

close monitoring and frequent contacts do not lead to unnecessarily higher levels of detection, but rather serve as an effective deterrent.

Although the results for convicted offenses suggests that the differences between the treatment group and control group could be due to chance ($p>.05$) the negative coefficient for convicted offenses (i.e., the treatment group is more likely to be convicted or incarcerated in comparison to the control group) is intuitively interesting. First, it is interesting to think about why aftercare/re-entry programs produce a different result depending on whether 'alleged offenses' (e.g., arrests, charges) and convicted offenses (e.g., adjudications and incarceration) are used as the outcome measure. Also, it is important to consider an alternative explanation; that a higher likelihood of conviction might also be attributed to the intensity of the supervision, where the strict conditions of compliance ensure that when/if there is any sort of violation (no matter the severity of offense), it will result in revocation and conviction/incarceration. As discussed in section 1.1.3, Caputo (2004) states that "[w]hen [offenders] fail as a result of rule violations, their punishments are typically more severe…" (Caputo, 2004, p. 189), suggesting that this failure is possibly more attributable to strict enforcement and revocation rather than criminal activity (Caputo, 2004; Merrington, 2006).

Subgroup analysis

Although methodological characteristics were not consistently shown to moderate results, the subgroup analyses presented here showed demonstrable importance in examining the moderating effect of methodological variables. Specifically, research design, length of follow-up period, sample size of treatment group, and publication type were all shown to be significant moderating variables. These findings suggest that methodological factors should be carefully considered in evaluation design, and rigorous methodology should be used whenever possible.

Lipsey (1997) discusses that in a typical outcome evaluation, only 25% of the treatment effect is attributed to differences in characteristics of the individual programs, whereas 50% of variance in effect sizes can be attributed to methodological characteristics. That is, it is possible that the size of an effect has less to do with the actual characteristics and effectiveness of the intervention than the methodological

differences and the associated moderating abilities of these variables (Lipsey & Wilson, 2001c). Lipsey and Wilson (2001b) note that “[g]iven the smaller average magnitude of the eta-squared and its smaller variability relative to that for constructs and operationalizations, it appears that the "who" may be less important than the "what" and "how" of measurement” (p. 420). As the findings demonstrated herein suggest that methodological characteristics are significant moderators of the results, analyses should be interpreted with caution as there is no guarantee that the effectiveness of the intervention is producing the effect, or whether the size of an effect might be more attributable to the ‘how’ of the study rather than the ‘what’.

As length of follow-up was the most consistent moderator across analyses, it warrants some discussion. Although the findings that shorter follow-up periods are related to stronger and statistically significant effects makes sense in terms of program effects wearing off over time, the subgroup analysis for ISP convicted offenses suggested the opposite result; that short follow-up periods were related to smaller effects. Parker et al. (2014) offer a plausible explanation for why this might be. Parker et al. (2014) argue that in community-based supervision programs, “recidivism” tends to start being measured at day 1 of supervision. This means that treatment effects hardly have the chance to take effect before the youth is considered to have failed (Parker et al., 2014). Although interventions are not expected to have an immediate impact, the effectiveness of the program starts to be measured immediately. Taking this into consideration, long follow-up periods might be better estimates to capture the true effect of the program on recidivism. This, along with the taking into account the ‘danger time’ discussed above (where youth in aftercare are more likely to recidivate immediately following their release), it might not be surprising that youth have high recidivism rates at short follow-ups.

Although only four variables were selected for further investigation into moderating effects, it is important to make note of why other variables were not selected for analysis. First, as will be discussed in the following paragraph, although examining the moderating effects of intervention characteristics would have been ideal, and would have allowed for valuable insight towards components of programs that contribute to beneficial treatment impacts, information was not reported consistently, and could thus

not be coded for analytic purposes. Second, it is important that the variables included in moderator analysis have a reasonably equal distribution in terms of number of studies in either category. For example, although it would have been intuitively interesting to know if interventions were more effective on males compared to females, or minorities versus Caucasians, the samples in studies were fairly one-sided, and could not be selected for subgroup analysis.

To expand on the point made above, although the subgroup analyses shown in Chapter 4 demonstrate the importance of examining methodological characteristics as significant moderator variables, this analysis was not able to provide insight towards the moderating effects of intervention characteristics. All things considered, it is important for future research to examine intervention characteristics as moderator variables, and a source of insight for why treatment effectiveness is consistently inconsistent. There are a number of intervention characteristics that could affect the effectiveness of an intervention. Although it was not possible to examine the moderating effects of intervention characteristics here, it is worth mentioning certain variables for consideration: treatment intensity and dosage (Merrington, 2006; Taxman, 2008), level of supervision and enforcement (Taxman, 2008), lack of financial and community resources (Petersilia, 1993), poorly trained staff (Gendreau et al., 2004), inadequate or inconsistent implementation (Clear & Hardyman, 1990; Corbett, 1999; Taxman, 2008), and follow-up period (Parker et al., 2014).

With respect to program effectiveness, acknowledging and examining the effect of moderator variables must not be understated and greater effort needs to be put forth by authors to consistently report details on intervention characteristics so that these characteristics can be examined further. It is also important for these potential moderator variables to be consistently reported so that the effect of individual characteristics treatment regimes can be examined and evidence-based policies and practices can be developed. Furthermore, consistent reporting of intervention characteristics is important so that the effectiveness of interventions are not reduced to statements about programs being holistically effective or ineffective. If information on moderator variables are consistently reported and subsequently analyzed, greater insight can be gained with respect to why evidence on supervision-oriented studies is

inconsistent across studies. As such, although the findings suggest that there is no differential impact on recidivism between intensive supervision probation and traditional probation, further efforts should be made towards examining the circumstances under which more supervision is better, and the circumstances under which more supervision is detrimental.

To summarize, in line with existing literature on the topic, the results do not present consistently favourable effects for either aftercare/re-entry or intensive supervision probation programs for youth. However, this does not diminish the importance of the new analyses. In contrast to the existing meta-analyses on the topic, this study provides important insight to the differing effect an intervention can have on different recidivistic outcome measures, and introduces a new realm that is worthy of further investigation. Furthermore, due to the strict inclusion criteria, and the exclusion of studies that focus on specific populations of offenders, the results demonstrated here are highly generalizable to the majority of offenders. Additionally, this study demonstrated that some methodological variables are significant moderators to the results, particularly the length of follow-up. The analyses herein also provide insight toward the influential effect of outlier effects, demonstrating the importance of sensitivity analyses and the transparent reporting of the distribution of individual effect sizes.

Chapter 6. Limitations

6.1. Search strategy

In a meta-analysis, the identified studies are highly dependent on the search terms. Thus, it is imperative that the search strategy is comprehensive because an incomplete search strategy may result in important studies being excluded. One of the challenges with conducting a comprehensive literature search on the topic of intermediate sanctions was the variability in language, i.e., unstandardized language and terms being used interchangeably for similar interventions. Evidence of this can be seen in the search strategy shown above (construct 3), and the range of terms and phrases that were required to fully capture relevant literature. Unstandardized language made it difficult to feel confident that all facets of the topic were being searched and that all relevant studies were being located. To ensure that the literature search was exhaustive and comprehensive with respect to 'intermediate sanctions', the literature on the topic was extensively examined to identify terms and phrases that were commonly used to refer to the various forms of intermediate sanction interventions. Once the search construct was finalized, the search was tested through an iterative process of adding and removing key terms, and applying wildcards and Boolean operators. As can be seen from this search, it is evident that standardized key words for this literature do not exist.

On a similar note, a comprehensive literature search that consists of broad and encompassing terms comes with the caveat of a vast number of hits that are detected by the search. Further to this point, as can be seen in Figure 2, there was a large number of false positive hits (the number of initial hits from across all of the electronic databases that were identified as relevant compared to the number of studies that were selected for review). This also raises the possibility that there is a number of false negative studies (studies that were relevant to the meta-analysis but were not identified by the search strategy) that were missed, a caveat that is ever present in meta-analytic research.

While the possibility of unidentified studies exists, computerized database searches combined with an extensive search of the grey literature remains to be the technique that provides the most comprehensive search. As such, we can say with a strong degree of certainty that the sample of studies used in the analyses is, at very least, representative of the population of studies on the topic, if not the full population.

6.2. Selection criteria

One aspect of a meta-analysis that is often subject to criticism is the selection criteria. As previously mentioned, although the strict exclusion criteria used in the current study is considered a strength of this research, it is also a limitation. First, it can be argued that the strict exclusion criteria prohibit a full summary of the evidence. That is, the set of studies included here are not truly comprehensive and representative of the literature as a whole on ISP and aftercare/re-entry programs, and, in that sense, may be biased. Similarly, studies that focused on very specific types of offenders were excluded. In essence, limiting inclusion to maximize generalizability is contradictory, however, these very specific populations of offenders (substance user, issues with mental health) are indeed very specific populations that require very specific interventions. However, it can be argued that by excluding these specific populations, the results do not summarize the full body of evidence.

A second limitation of the selection criteria is the exclusion of 'technical violations' as an outcome measure. At the outset, technical violations were excluded as a measure of recidivism because the analysis was intended to capture 'true' criminal activity, rather than simple breaches of probation. However, as demonstrated in the literature, taking into consideration the strict compliance conditions and high rates of program failure and the associated volume of technical violations is important to understanding the complete picture of effectiveness for supervision-based interventions. Although the distinction of recidivism between alleged offenses and convicted offenses offers somewhat of a proxy measure, including outcomes that specifically measured technical violations could have provided valuable insight.

6.3. Commensurability

Another limitation is the commensurability of the interventions (or lack of commensurability). Ideally, the study would have allowed for a pooled analysis of the whole set of 55 effect sizes, however, upon close examination, the two types of interventions (aftercare and ISP) differed consistently on the nature of the comparison groups and were consequently not commensurable. Similarly, although meta-analyses typically pool together various measures of recidivism, it is important to recognize the variability among recidivistic outcome measures. Furthermore, it is important to consider that, at least under certain circumstances, pooling these outcome measures might not be appropriate and might constitute 'mixing apples and oranges'. Upon close examination of outcome measures, it was noted that the outcome measures ranged from consequential recidivism (i.e., incarceration) to rather inconsequential recidivism (i.e., contact with police), thus the set of studies were bifurcated once more. In light of attention to commensurability and concerns raised in regards to the fundamental differences between comparison groups and outcome measures, what would be considered a moderately sized sample for a meta-analysis (N=55) was divided into four sets of studies. The limitation that arises from this is that studies with small samples are likely to be vulnerable to bias and overestimate effects.

6.4. Recidivism as the sole measure of effectiveness

Finally, it is important to understand the limitation of using recidivism as the only measure of program 'success' or 'effectiveness'. As discussed above, correctional interventions are multi-dimensional and offenders face a series of challenges. Moreover, while reducing criminal behaviour is the ultimate end goal of correctional programming, interventions focus on an array of skills in aim of addressing multiple risk factors. That being said, although reducing criminal behaviour is the end goal, it is not the only goal, and interventions use a variety of strategies to achieve this aim, all of which should be considered when evaluating whether a program is effective.

Chapter 7. Conclusion and recommendations

The findings from the 4 analyses provide important insight toward the effectiveness of community-based supervision programs on reducing juvenile recidivism. Although the mixed evidence found herein is consistent with the existing literature on the topic, the current study differentiates itself from previous meta-analyses through the bifurcation of recidivism outcome measures, and the pattern that was observed across intervention type, outcome measure, and effectiveness. Numerous points can be taken from this study.

First, the results suggest that when it comes to measuring the effect of aftercare/re-entry and ISP interventions, there is a difference between whether alleged offenses (criminal contact, charges, arrest) or convicted offenses (adjudication, conviction, incarceration) were used as the outcome measure. More specifically, the findings for aftercare/re-entry on alleged offenses suggests that youth who participate in aftercare programs are significantly less likely to be charged for an offense or arrested upon release from custody in comparison to the control group that received 'care as usual' post-release. With respect to the pattern across intervention type(s), this finding is indicative of the importance of making a distinction when it comes to recidivism, the importance of not combining "apples and oranges" when it comes to recidivism measures and, wherever possible, investigated recidivism through separate outcome measures.

Second, with respect to the effect of aftercare/re-entry on alleged offenses, the interventions demonstrated a significant reduction among treatment group participants, providing support for the deterring effect of supervision. On the other hand, with respect to ISP, although rigorous methodology was used in both analyses (~80% of studies were randomized controlled trials), the results suggest that there is no significant difference in the pooled effect between the treatment and control groups. This finding is indicative

that intensive forms of probation are not necessary for effective supervision. This is an important finding, and one that is worthy of serious consideration, as strict enforcement and conditions of compliance that is associated with some forms of intensive supervision perpetuate a cycle of criminal activity.

Third, of particular importance to the findings of this study was the influential effect of outlier studies on the reliability of the pooled effects. As such, the results should be interpreted cautiously. The presence of outliers showed demonstrable influence in changing the interpretive meaning of the results i.e., one outlier study carried so much influence over the entire pooled effect that the exclusion of that study could change the directionality (positive to negative) and magnitude of the overall estimate, thereby suggesting that the results are not robust. However, outlier effects are still important to include in the analyses as, the fact of the matter is, that community-based supervision programs are heterogeneous when it comes to effectiveness.

Fourth, subgroup analyses showed methodological variables to be important moderators to the results. Specifically, research design, length of follow-up period, sample size of treatment group, and publication type were all shown to be significant moderating variables. Notably, across analyses "length of follow-up" was a consistent moderator, suggesting that the length of follow-up selected by the researcher should be given considerable thought, as it plays a decisive role in the ultimate outcome effect measured in the study. Furthermore, due to the "danger time" associated with aftercare/re-entry, and the suggestion that youth on intensive supervision might have high recidivism rates at short follow-ups because the measurement of "failure" starts at day 1 of supervision (when, inevitably, the intervention has not yet had a demonstrable impact), researchers might be wise to assess recidivism outcomes for longer follow-up periods.

Finally, the evaluation and improvement of community-based supervision programs for juvenile offenders should be of utmost importance to correctional research. Considering the mixed evidence (demonstrated here and in previous research), and the varying degree of program effectiveness, one could certainly argue that the inconsistent evidence is suggestive that, overall, these interventions do not reliably reduce recidivism

and should therefore be replaced with new interventions. On the other hand, an argument could also be made that, before drawing definitive conclusions with respect to the effectiveness of community-based supervision programs, yet more research on the topic of community-based supervision programs is required. For example, moderator analysis on intervention characteristics and interviews with key stakeholders may advance our understanding with respect to the circumstances under which these interventions are related to decreases in criminal activity.

Lastly, a few recommendations can be drawn from the current research and applied to future research to better fill in the gaps of the effectiveness of community-based supervision programs for youth:

1. The need for more rigorous research designs is a pressing issue. Although intensive supervision programs included in the current meta-analysis demonstrated reasonably rigorous methodology (over 80% were randomized controlled trials), the same pattern was not observed for aftercare/re-entry.

2. Standardized reporting guidelines: As discussed above, numerous intervention and sample characteristics could not be analyzed due to missing information. In order to draw more conclusive evidence towards the circumstances under which interventions are most successful (i.e., the presence or absence of certain intervention characteristics) and for whom these programs are most effective, this information needs to be consistently and reliably reported. Of utmost importance should be the detailed reporting of sample characteristics and intervention characteristics so future moderator analyses can investigate the characteristics that are related to strong (or weak) results.

3. Although the results from the moderator analyses of the current study demonstrates the importance of rigorous methodology, the moderating effect of intervention characteristics should be examined closely before drawing definitive conclusions towards the effectiveness of community-based supervision programs.

4. Measures other than recidivism alone should be included as important measures of 'success'. As demonstrated here, as well as in the existing literature, community-based supervision programs have produced what many consider to be less than impressive results in terms of reducing recidivism (Petersilia, 1993). In the evaluation of criminal justice interventions, success is most commonly measured through a variation of recidivistic measures (arrest, incarceration, conviction, severity, desistance, time to failure) (King, 2014). Although this is a critical performance measure in any criminal justice related evaluation, there is emerging reluctance towards this definition of success (Mead, 2005).

   In light of the generally inconsistent evidence towards the effectiveness of ISP and aftercare/re-entry on recidivism, perhaps 'success' is too stringent a criterion (Petersilia, 1993). With respect to aftercare/re-entry programs in particular, program interventions target multiple facets and are a collaborative effort between multiple agencies, thus, there may be more to the term 'success' than simply remaining crime free. In order to capture a meaningful measure of 'success', researchers should take into account various program components and outcome measures (King, 2014; Petersilia, 1993).

As the current research allows for cautiously optimistic conclusions towards the effect of community-based supervision in reducing recidivism among youth, these recommendations and the call for specific research efforts will help build on current knowledge and fill in the gaps of existing research. If these recommendations are followed, future research will better inform policymakers towards the circumstances under which community-based supervision programs for youth are effective at reducing recidivism. Finally, evidence-based decision-making can be used to inform and further improve the justice system's approach to supervising and rehabilitating juvenile offenders.

References

*denotes studies that are included in the meta-analysis

Altschuler, D. M., Armstrong, T. L., & MacKenzie, D. L. (1999). Reintegration, Supervised Release, and Intensive Aftercare. Juvenile Justice Bulletin. Retrieved from http://www.ncjrs.gov/pdffiles1/175715.pdf

Aos, S., Phipps, P., Barnoski, R., & Lieb, R. (2001). The comparative costs and benefits of programs to reduce crime. Version 4.0. Olympia, WA: Washington State Institute for Public Policy. Retrieved from http://www.wsipp.wa.gov/ReportFile/756/Wsipp_The-Comparative-Costs-and-Benefits-of-Programs-to-Reduce-Crime-v-4-0_Full-Report.pdf

Armstrong, T. L. (1990). Introduction. In J. Smykla & W. Selke (Eds.), *Intermediate sanctions: Sentencing in the 1990s* (1-25). Cincinnati, OH: Anderson Publishing Co.

Austin, J. (2001). Prisoner reentry: Current trends, practices, and issues. *Crime & Delinquency, 47*(3), 314-334. doi: 10.1177/0011128701047003002

Baird, C. (1991). Intensive supervision programs for high-risk juveniles: Critical issues of program evaluation. In T.L. Armstrong (Ed.) *Intensive interventions with high risk youths: Promising approaches in juvenile probation and parole* (295-315). Cincinnati, OH: Anderson Publishing Co.

*Barnoski, R. P. (2003). *Evaluation of Washington State's 1996 Juvenile Court Program for High-risk, First-time Offenders: Final Report*. Washington State Institute for Public Policy. Retrieved from http://pgn-stage.wsipp.wa.gov/ReportFile/835/Wsipp_Evaluation-of-Washingtons-1996-Juvenile-Court-Program-Early-Intervention-Program-for-High-risk-First-time-Offenders-Final-Report_Full-Report.pdf

Barton, W. H., & Butts, J. A. (1990). Viable options: Intensive supervision programs for juvenile delinquents. *Crime & Delinquency*, *36*(2), 238–256. http://doi.org/10.1177/0011128790036002004

*Barton, W. H., Jarjoura, G. R., & Rosay, Andre B. (2008). *Evaluation of the Boys & Girls Club of America: Targeted re-entry initative: Final report*. Retrieved fromhttp://justice.uaa.alaska.edu/research/2000/0411.targeted_reentry/0411.05.targeted_reentry.pdf

Bennett, L. A. (1988). Practice in search of a theory: The case of intensive supervision—An extension of an old practice or a new approach?. *American Journal of Criminal Justice*, *12*(2), 293-310. doi: 10.1007/BF02888940

*Bergseth, K. J., & McDonald, T. D. (2007). Reentry services: An evaluation of a pilot project in Clay County, MN. *Fargo, ND: North Dakota State University Department of Criminal Justice and Political Science*. Rerieved from http://www.claycountycollaborative.org/meetings/files/RSPFinalReport2007.pdf

Borenstein, M., & Higgins, J. P. T. (2013). Meta-analysis and subgroups. *Prevention Science*, *14*(2), 134–143. http://doi.org/10.1007/s11121-013-0377-7

*Bouffard, J. A., & Bergseth, K. J. (2008). The impact of reentry services on juvenile offenders' recidivism. *Youth Violence and Juvenile Justice*, *6*(3), 295–318. http://doi.org/10.1177/1541204007313384

Breaugh, J. A. (2003). Effect size estimation: Factors to consider and mistakes to avoid. *Journal of Management, 29*(1), 79-97. doi: 10.1177/014920630302900106

*Burke, C., Keaton, S., & Pennell, S. (2003). Addressing the gender-specific needs of girls: An evaluation of San Diego's WINGS program--Final report. *San Diego, CA: San Diego Association of Governments (SANDAG)*. Retrieved from http://sandiegohealth.org/sandag/publicationid_899_3916.pdf

Byrne, J. M. (1990). The future of intensive probation supervision and the new intermediate sanctions. *Crime & Delinquency, 36*(1), 6-41. doi: 10.1177/0011128790036001003

Byrnes, M., Macallair, D., & Shorter, A. (2002). Aftercare as Afterthought: Reentry and the California Youth Authority. *San Francisco, CA: Center on Juvenile and Criminal Justice*. Retrieved from https://www.opensocietyfoundations.org/sites/default/files/aftercare.pdf

Caputo, G. A. (2004). *Intermediate sanctions in corrections*. Denton, TX: University of North Texas Press.

Card, N. A. (2011). *Applied meta-analysis for social science research*. New York, NY: The Guilford Press.

Carlson, K. D., & Schmidt, F. L. (1999). Impact of experimental design of effect size: Findings from the research literature on training. *Journal of Applied Psychology, 84*(6), 851-862. Retrieved from http://www.management.pamplin.vt.edu/directory/Articles/Carlson5.pdf

Cillo, G. C. (2001). Evaluation of a theory-based transitional aftercare program for court-adjudicated adolescents. (dissertation). New York, NY: Fordham University

Clear, T. R., & Hardyman, P. L. (1990). The new intensive supervision movement. *Crime & Delinquency, 36*(1), 42-60. doi: 10.1177/0011128790036001004

Clouser, M. (1996). Intensive probation: An alternative to placement. *Pennsylvania Progress*, *3*(1), 1-6. Retrieved from http://www.ncjj.org/PDF/pa_Progress/1996/jan96.pdf

Corbett, R. (1999). Juvenile probation on the eve of the next millennium. *Federal Probation, 63*(2), 78-87.

Cullen, F. T., & Gendreau, P. (2000). Assessing correctional rehabilitation: Policy, practice, and prospects. *Criminal Justice*, *3*(109-175). Retrieved from https://www.ncjrs.gov/criminal_justice2000/vol_3/03d.pdf

Cullen, F. T., & Gendreau, P. (2001). From nothing works to what works: Changing professional ideology in the 21st century. *The Prison Journal*, *81*(3), 313–338. http://doi.org/10.1177/0032885501081003002

Dunst, C., Hamby, D., & Trivette, C. (2004). Guidelines for calculating effect sizes for practice-based research syntheses. *Centerscope, 3*(1), 1-10. Retrieved from http://www.courseweb.unt.edu/gknezek/06spring/5610/centerscopevol3no1.pdf

Durlak, J. A. (2009). How to select, calculate, and interpret effect sizes. *Journal of Pediatric Psychology, 34*(9), 917-928. doi: 10.1093/jpepsy/jsp004

Durlak, J., & Lipsey, M. (1991). A practitioner's guide to meta-analysis. *American Journal of Community Psychology, 19*(3), 291-332. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/1892130

Fairfull, A. (2013). *Young people: the experience of transition from custody to community* (Doctoral dissertation, University of Liverpool).

Farrington, D. (2003). A short history of randomized experiments in criminology: A meager feast. *Evaluation Review, 27*(3), 218-227. doi: 10.1177/0193841X03252482

Farrington, D. P. (2006). Methodological quality and the evaluation of anti-crime programs. *Journal of Experimental Criminology, 2*, 329-337.http://doi: 10.1007/s11292-006-9012-y

Farrington, D. and Welsh, B., (2005). Randomized experiments in criminology: What have we learned in the last two decades? *Journal of Experimental Criminology, 1*(1), 9–38. doi: 10.1007/s11292-004-6460-0

Filges, T., Knudsen, A.-S. D., Svendsen, M. M., Kowalski, K., Benjaminsen, L., & Jorgensen, M. K. (2015). Cognitive-behavioural therapies for young people in outpatient treatment for non-opioid drug use: A systematic review. *Campbell Systematic Reviews, 11*(3)*,* 10.4073/csr.2015.3

Filges, T., Rasmussen, P. S., Andersen, D., & Jorgensen, A. K. (2015). Multidimensional family therapy (MDFT) for young people in treatment for non-opioid drug use: A systematic review*. Campbell Systematic Reviews, 11*(8), 10.4073/csr.2015.8

Garfield, E. (1991). Meta-analysis and the metamorphosis of the scientific literature review. *Essays of an Information Scientist: Science Reviews, Journalism Inventiveness and Other Essays, 14*, 170-176. Retrieved from http://www.garfield.library.upenn.edu/essays/v14p170y1991.pdf

Gendreau, P. (1981). Treatment in corrections: Martinson was wrong! *Canadian Psychology, 22*(4), 332-338. Retrieved from http://psycnet.apa.org/index.cfm?fa=buy.optionToBuy&id=1982-33462-001

Gendreau, P., French, S. A., & Gionet, A. (2004). What works (what doesn't work): The principles of effective correctional treatment. *Journal of Community Corrections, 13*, 4-30. Retrieved from http://cp.co.hancock.oh.us/wp-content/uploads/2011/09/Gendreau-French-Gionet-2004-What-works-and-what-doesnt.pdf

Gendreau, P., & Ross, R. (1979). Effective correctional treatment: Bibliotherapy for cynics. *Crime & Delinquency,* 25 (October): 463–489. doi: 10.1177/001112877902500405

Gendreau, P., & Ross, R. (1987). Revivification of rehabilitation: Evidence from the 1980s. *Justice Quarterly* 4 (September): 349–407. doi: 10.1080/07418828700089411

*Giblin, M. J. (2002). Using police officers to enhance the supervision of juvenile probationers: An evaluation of the Anchorage CAN program. *Crime & Delinquency*, *48*(1), 116–137. http://doi.org/10.1177/0011128702048001005

*Gray, E., Taylor, E., Roberts, C., Merrington, S., Fernandez, R., & Moore, R. (2005). *The Intensive Supervision and Surveillance Programme: the Final Report*, London: Youth Justice Board. Retrieved from http://yjbpublications.justice.gov.uk/Resources/Downloads/ISSP%20the%20final%20report.pdf

Greenwood, P. (1994). What works with juvenile offenders: A synthesis of the literature and experience. *Federal Probation, 58*(4), 63-72. Retrieved from http://www.rand.org/pubs/reprints/RP388.html

*Greenwood, P. W., Deschenes, E., & Adams, J. L. (1993). *Chronic juvenile offenders: Final results from The Skillman Aftercare Experiment*. Rand Corporation. Retrieved from https://www.rand.org/content/dam/rand/pubs/monograph_reports/2006/MR220.pdf

Griffin, P. (2004). Aftercare: The Sequel. *Pennsylvania Progress*, *11*(1), 1–11. Retrieved from http://www.ncjj.org/PDF/december2004.pdf

Guolo, A., & Varin, C. (2015). Random-effects meta-analysis: The number of studies matters. *Statistical Methods in Medical Research, 0*(0), 1-19. doi: 10.1177/0962280215583568

*Hawkins, S. R., Lattimore, P. K., Dawes, D., & Visher, C. A. (2009). Reentry experiences of confined juvenile offenders: Characteristics, service receipt, and outcomes of juvenile male participants in the SVORI Multi-site Evaluation. *Research Triangle Park, NC: RTI*.. Retrieved from https://ncjrs.gov/pdffiles1/nij/grants/230423.pdf

Hennigan, K., Kolnick, K., Tian, T., Maxson, C., & Poplawski, J. (2010). Five year outcomes in a randomized trial of a community-based multi-agency intensive supervision juvenile probation program. U.S. Department of Justice. Retrieved from https://www.ncjrs.gov/pdffiles1/nij/grants/232621.pdf

*Hennigan, K., Maxson, C. L., & Zhang, S. (2005). Self reported outcomes in a randomized trial of a community-based multi-agency intensive supervision juvenile probation program. *National Criminal Justice Reference Service*, 1–58. Retrieved from https://www.ncjrs.gov/pdffiles1/Digitization/212433NCJRS.pdf

Hockenberry, S., & Puzzanchera, C. (2014). Delinquency Cases in Juvenile Court, 2011. Juvenile Offenders and Victims National Report Series Fact Sheet. Office of Juvenile Justice and Delinquency Prevention. Retrieved from http://www.ojjdp.gov/pubs/248409.pdf

*Howard, L., Misch, G., Burke, C., & Pennell, S. (2002). San Diego county probation department's Repeat Offender Prevention Program: Final evaluation report. *San Diego, CA: SANDAG, San Diego's Regional Planning Agency*. Retrieved from http://sandiegohealth.org/crime/publicationid_753_1432.pdf

Howell, J. C., & Lipsey, M. W. (2004). A practical approach to evaluating and improving juvenile justice programs. *Juvenile and Family Court Journal*, winter, 35-48. doi: 10.1111/j.1755-6988.2004.tb00095.x

*Howitt, P. S., & Moore, E. A. (1991). The efficacy of intensive early intervention: An evaluation of the Oakland County probate court early offender program. *Juvenile and Family Court Journal*, *42*(3), 25–36. doi: 10.1111/j.1755-6988.1991.tb00892.x

*Iutcovich, J. M., & Pratt, D. J. (1998). *A final report for: Assessment of aftercare services provided to delinquent youth*. Pennsylvania Commission on Crime and Delinquency.

James, C., Stams, G. J. J., Asscher, J. J., De Roo, A. K., & van der Laan, P. H. (2013). Aftercare programs for reducing recidivism among juvenile and young adult offenders: A meta-analytic review. *Clinical Psychology Review*, *33*(2), 263-274. doi: http://dx.doi.org/10.1016/j.cpr.2012.10.013

Kellum, S. A. (2006). Aftercare when there is no aftercare. *College of Juvenile Justice & Psychology*, *1*(1), 51–58. Retrieved from http://works.bepress.com/sharlette_a_kellum_gilbert/1/

King, R., & Elderbroom, B. (2014). Improving recidivism as a performance measure. Urban Institute Justice Policy Center Brief. Retrieved from http://hint-magazine.com/wp-content/uploads/2014/10/413247-improving-recidivism.pdf

*Land, K. C., McCall, P. L., & Williams, J. R. (1990). Something that works in juvenile justice: An evaluation of the North Carolina court counselors' intensive protective supervision randomized experimental project, 1987-1989. *Evaluation Review*, *14*(6), 574–606. doi.org/10.1177/0193841X9001400602

*Lane, J. (2005). Evaluating an experimental intensive juvenile probation program: Supervision and official outcomes. *Crime & Delinquency*, *51*(1), 26–52. doi.org/10.1177/0011128704264943

Langan, P. (1994). Between prison and probation: Intermediate sanctions. *Science, 264*(5160), 791-793. Retrieved from http://www.jstor.org/stable/2884018

Lemert, E. M. (1993). Visions of social control: Probation considered. *Crime & Delinquency, 39*(4), 447-461. doi: 10.1177/0011128793039004003

Lindstrom, M., Saidj, M., Kowalski, K., Filges, T., Rasmussen, P., & Klint Jørgensen, A. (2015). Family behavior therapy (FBT) for young people in treatment for illicit non-opioid drug use: A systematic review. *Campbell Collaboration Reviews, 11*(9), 10.4073/csr.2015.9

Lipsey, M. W. (1997). What can you build with thousands of bricks? Musings on the cumulation of knowledge in program evaluation. *New Directions for Evaluation, 76*(winter), 7-23. doi: 10.1002/ev.1084

Lipsey, M. W. (2000). Meta-analysis and the learning curve in evaluation practice. *American Journal of Evaluation, 21*(2), 207-212. doi: 10.1177/109821400002100208

Lipsey, M. W. (2002). Meta-analysis and program outcome evaluations. *Socialvetenskaplig tidskrift nr*, 2-3, 194-208. Retrieved from http://svt.forsa.nu/Documents/Forsa/Documents/Socialvetenskaplig%20tidskrift/Artiklar/2002/Meta-Analysis%20and%20Program%20Outcome%20Evaluation%20av%20Mark%20W%20Lipsey.PDF

Lipsey, M. W. (2003). Those confounded moderators in meta-analysis: Good, bad, and ugly. Annals of the American Academy of Political and Social Science, 587, 69-81. Retrieved from http://www.jstor.org/stable/1049948

Lipsey, M. W. (2006). Improving the evaluation of anticrime programs: There's work to be done. Journal of Experimental Criminology, 2, 517-527. doi: 10.1007/s11292-006-9022-9

Lipsey, M. W. (2009). The primary factors that characterize effective interventions with juvenile offenders: A meta-analytic overview. *Victims & Offenders*, *4*(2), 124–147. http://doi.org/10.1080/15564880802612573

Lipsey, M. W. (2014). Interventions for juvenile offenders: A serendipitous journey. *Criminology & Public Policy*, *13*(1), 1-14. doi:10.1111/1745-9133.12067

Lipsey, M. W., & Wilson, D. B. (2001a). Practical meta-analysis. *Applied Social Research Methods Series: Volume 49.* Thousand Oaks, CA: Sage Publications.

Lipsey, M. W., & Wilson, D. B. (2001b). The role of method in treatment effectiveness research: Evidence from meta-analysis*. Psychological Methods, 6*(4), 413-429. doi: 10.1037//1082-989X.6.4.413

Lipsey, M. W., & Wilson, D. B. (2001c). The way in which intervention studies have 'personality' and why it is important to meta-analysis. *Evaluation & the Health Professsions, 24*(3), 236-254. doi: 10.1037//1082-989X.6.4.413

Littell, J. H., Corcoran, J., & Pillai, V. (2008). Systematic reviews and meta-analysis. New York: Oxford University Press.

*Little, M., Kogan, J., Bullock, R., & Van Der Laan, P. (2004). ISSP: An experiment in multi-systemic responses to persistent young offenders known to children's services. *British Journal of Cri*minology, 44(2), 225–240. doi: 10.1093/bjc/44.2.225

Martinson, R. (1974). What works? –questions and answers about reform. *The Public Interest, 0(*35), 22-54. Retrieved from http://www.nationalaffairs.com/doclib/20080527_197403502whatworksquestionsandanswersaboutprisonreformrobertmartinson.pdf

Mead, B. (2005). Is there a measure of probation success? *Federal Probation, 69*(2), 3-5.

Merrington, S. (2006). Is more better? The value and potential of intensive community supervision. *Probation Journal*, *53*(4), 347–360. doi:10.1177/0264550506069360

*National Council on Crime and Delinquency. (2001). Evaluation of the RYSE Program: Alameda County Probation Department. Retrieved from http://www.nccdglobal.org/sites/default/files/publication_pdf/ryse-program.pdf

*Norman, M. D. (1995). *An evaluation of the Utah second district juvenile court intensive supervision probation program*. Utah State Juvenile Court.

Palmer, T. (1975). Martinson revisited. *Journal of Research in Crime and Delinquency* 12 (July): 133–152. doi: 10.1177/002242787501200206

Parker, R., Bush, J., & Harris, D. (2014). Important methodological issues in evaluating community-based interventions. *Evaluation Review*, *38*(4), 295–308. doi.org/10.1177/0193841X14542610

Petersilia, J. (1990). Conditions that permit intensive supervision programs to survive. *Crime & Delinquency, 36*(1), 126-145. doi: 10.1177/0011128790036001009

Petersilia, J. (1999). A decade of experimenting with intermediate sanctions: What have we learned? *Corrections Management Quarterly, 3*(3), 19-27. Retrieved from http://b.3cdn.net/crjustice/d53cccf8d18b67cd1d_p3m6bt2dc.pdf

Petersilia, J. (2004). What works in prisoner reentry? Reviewing and questioning the evidence. *Federal Probation, 68*(2), 4-8. Retrieved from http://www.researchgate.net/publication/238678771_What_Works_in_Prisoner_Reentry_Reviewing_and_Questioning_the_Evidence

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

Petrosino, A., Turpin Petrosino, C., & Buehler, J. (2003). “Scared Straight” and other juvenile awareness programs for preventing juvenile delinquency. *Campbell Systematic Reviews*, 2, 1-38. Doi: 10.4073/csr.2004.2

Pratt, T. C., Franklin, T. W. & Gau, J. M. (2011). *Key ideas in criminology and criminal justice*. Sage.

Redondo, S., Sanchez-meca, J., & Garrido, V. (1999). The influence of treatment programmes on the recidivism of juvenile and adult offenders: An european meta-analytic review. *Psychology, Crime & Law*, *5*(3), 251–278. http://doi.org/10.1080/10683169908401769

*Rodriguez-Labarca, J., & O'Connell, J. P. (2004). Delaware's serious juvenile offender program: An evaluation of the first two years of operation. State of Delaware Statistical Analysis Center. Retrieved from http://cjc.delaware.gov/pdf/sjo-evaluation.pdf

Sánchez-Meca, J., Marín-Martínez, F., & Chacón-Moscoso, S. (2003). Effect-size indices for dichotomized outcomes in meta-analysis. *Psychological Methods*, *8*(4), 448–467. http://doi.org/10.1037/1082-989X.8.4.448

Sarver, C. M., Molloy, J. K., & Butters, R. P. (2012). Utah cost of crime – intensive supervision (juveniles): Technical report. Utah Criminal Justice Center, University of Utah. Retrieved from http://ucjc.utah.edu/wp-content/uploads/ISP_Juv_Tech_updateformat.pdf

Schulz, R. (2007). Current methods for meta-analysis: Approaches, issues, and developments. *Journal of Psychology, 215*(2), 90-103. doi: 10.1027/0044-3409.215.2.90

Seiter, R., & Kadela, K. (2003). Prisoner reentry: What works, what doesn't, and what's promising. *Crime and Delinquency,* 49(3), 360-388. doi: 10.1177/0011128703253761

Shadish, Matt, G. E., Navarro, A. M., & Phillips, G. (2000). The effect of psychological therapies under clinically representative conditions: A meta-analysis. *Psychological Bulletin*, *126*(4), 512–529. doi: 10.1037//0033-2909.126.4.512

*Shelden, R. G. (1999). *Detention Diversion Advocacy: An Evaluation*. Juvenile Justice Bulletin. US Department of Justice, Office of Justice Programs, Office of Juvenile Justice and Delinquency Prevention. Retrieved from http://eric.ed.gov/?id=ED436605

Shelden, R. G. (2004). The imprisonment crisis in America: Introduction. *Review of Policy Research*, *21*(1), 5–12. doi: 10.1111/j.1541-1338.2004.00054.x

Sherman, L. W., Gottfredson, D. C., MacKenzie, D. L., Eck, J., Reuter, P., & Bushway, S.D. (1998). Preventing crime: What works, what doesn't, what's promising. National Institute of Justice Research in Brief. U.S. Department of Justice. Retrieved from https://www.ncjrs.gov/pdffiles/171676.PDF

Solomon, A. L. (2008). Putting public safety first: 13 Strategies for successful supervision and reentry (policy brief), 7. The PEW Center on the States. Retrieved from http://www.urban.org/research/publication/putting-public-safety-first-13-strategies-successful-supervision-and-reentry-policy-brief

*Sontheimer, H., & Goodstein, L. (1993). An evaluation of juvenile intensive aftercare probation: Aftercare versus system response effects. *Justice Quarterly*, *10*(2), 197–227. doi: 10.1080/07418829300091791

*Stafford, M., & Glassner, S. (2012). Evaluation of children's aftercare re-entry experience (CARE) for the Texas department of juvenile justice. School of Criminal Justice: Texas State University San Marcos. Youth Reentry and Reintegration Comprehensive Report. Retrieved from https://www.tjjd.texas.gov/publications/reports/2012ReentryReintegrationReport.pdf

Sterne, J. A., Egger, M., & Smith, G. D. (2001). Investigating and dealing with publication and other biases. M. Egger, G. D. Smith, D. G. Altman (Eds.) *Systematic Reviews in Health Care: Meta-Analysis in Context, Second Edition*, 189-208.

Stone, S. B., & Fulton, B. A. (1990). Achieving public safety through rehabilitation and reintegration: The promise of a new ISP. In J. Smykla & W. Selke (Eds.), *Intermediate sanctions: Sentencing in the 1990s* (115-134). Cincinnati, OH: Anderson Publishing Co.

Szumilas, M. (2010). Explaining odds ratios. *Journal of the Canadian Academy of Child and Adolescent Psychiatry*, *19*(3), 227.

Taxman, F. S. (2002). Supervision--Exploring the dimensions of effectiveness. *Federal Probation, 66*(2), 14-31.

Tonry, M. (1990). Stated and latent functions of ISP. *Crime & Delinquency, 36*(1), 174-191. doi: 10.1177/0011128790036001012

Tonry, M. (1998). Evaluating intermediate sanction programs. In J. Petersilia (Ed.), Community corrections: Probation, parole, and Intermediate Sanctions (79-96). Oxford University Press.

Torbet, P. (1997). *Juvenile probation: The workhorse of the juvenile justice system*. US Department of Justice, Office of Justice Programs, Office of Juvenile Justice and Delinquency Prevention. Retrieved from http://www.ncjj.org/pdf/desktop.pdf#page=13

Travis, J., & Petersilia, J. (2001). Reentry reconsidered: A new look at an old question. *Crime & Delinquency, 47*(3), 291-313. doi: 10.1177/0011128701047003001

Wakefield, S., & Uggen, C. (2010). Incarceration and stratification. *Annual Review of Sociology, 36,* 387-406. doi:10.1146/annurev.soc.012809.102551

Weaver, R. D., & Campbell, D. (2014). Fresh start: A meta-analysis of aftercare programs for juvenile offenders. *Research on Social Work Practice*, 25(2), 201-212, http:// 10.1177/1049731514521302

Weisburd, D. (2000). Randomized experiments in criminal justice policy: Prospects and problems. *Crime & Delinquency, 46*(2), 181-193. doi: 10.1177/0011128700046002003

Weisburd, D. (2010). Justifying the use of non-experimental methods and disqualifying the use of randomized controlled trials: Challenging folklore in evaluation research in crime and justice. *Journal of Experimental Criminology*, *6*(2), 209–227. http://doi.org/10.1007/s11292-010-9096-2

Weisburd, D., Lum, C. M., & Petrosino, A. (2001). Does research design affect study outcomes in criminal justice? *Annals of the American Academy of Political and Social Science, 578,* 50-70. Retrieved from http://www.jstor.org/stable/1049867 .

Wells, E. (2009). Uses of meta-analysis in criminal justice research: A quantitative review. *Justice Quarterly, 26*(2), 268-293. doi:10.1080/07418820802119984

*Wiebush, R. G. (1993). Juvenile intensive supervision: The impact of felony offenders diverted from institutional placement. *Crime & Delinquency, 39*(1), 68-89. doi: 10.1177/0011128793039001005

*Wiebush, R. G., Wagner, D., McNulty, B., Wang, Y., & Le, T. N. (2005). Implementation and outcome evaluation of the Intensive Aftercare Program. National Council on Crime and Delinquency. US Department of Justice, Office of Justice Programs, Office of Juvenile Justice and Delinquency Prevention. Retrieved from http://files.eric.ed.gov/fulltext/ED484683.pdf

Wilson, D. B. (2001). Meta-analytic methods for criminology. *Annals of the American Academy of Political and Social Science, 578*, 71-89. doi:10.1177/000271620157800105

Wilson, D. B. (2009). Missing a critical piece of the pie: Simple document searchstrategies inadequate for systematic reviews. *Journal of Experimental Criminology, 5*(4), 429-440. http://doi:10.1007/s11292-009-9085-5

Wilson, D. B., MacKenzie, D. L., & Mitchell, F. N. (2005). Effects of correctional boot camps on offending. *Campbell Systematic Reviews*, 6, 1-42. doi: 10.4073/csr.2005.6

Wong, J., Bouchard, J., Bouchard, M., Morselli, C., & Gravel, J. (2014). Effectiveness of Community-Based Crime Prevention Programs for At-Risk Youth (Public Safety Canada).

*Zhang, S. X., & Zhang, L. (2005). An experimental study of the Los Angeles county repeat offender prevention program: Its implementation and evaluation. *Criminology & Public Policy*, *4*(2), 205–236. doi: \\server05\product\C\CPP\4-2\CPP201.txt

Appendix A.

Aftercare/re-entry, alleged offense

Figure A1: Forest plot for aftercare/re-entry, alleged offense (N=1 outlier removed)

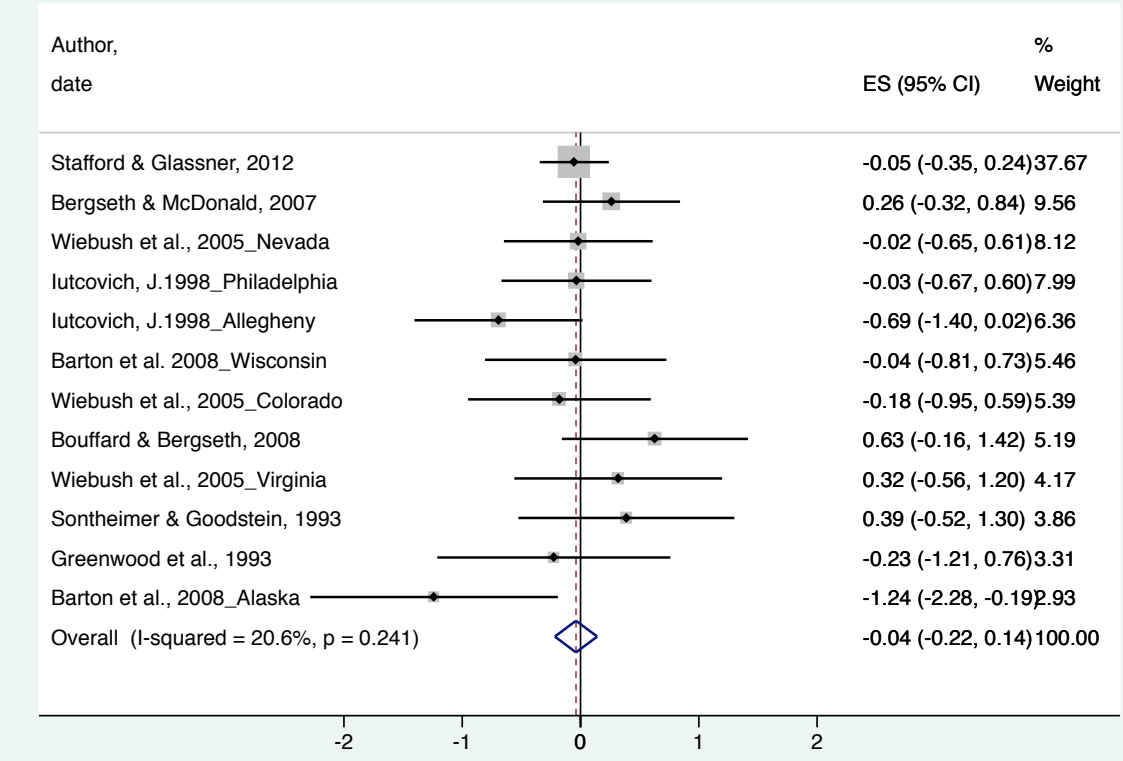


Figure A2: Funnel plot for aftercare/re-entry, alleged offenses (N=1 outlier removed)

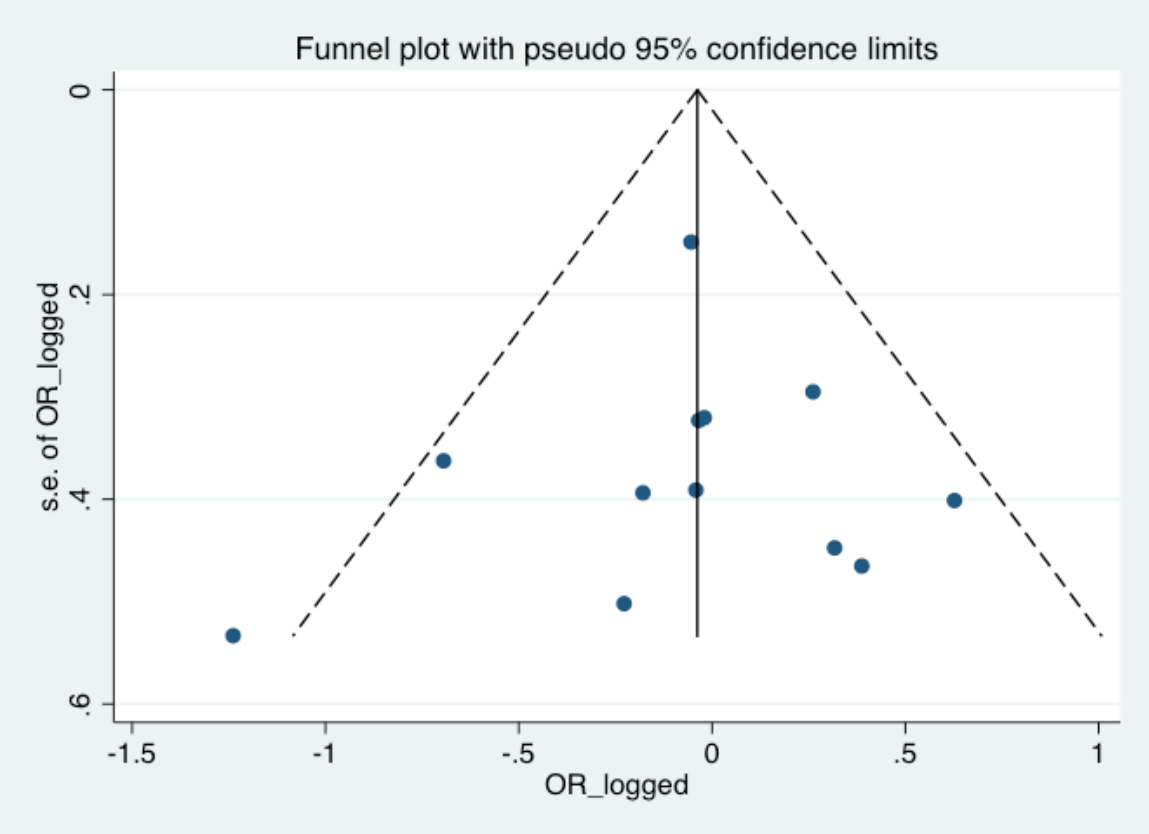
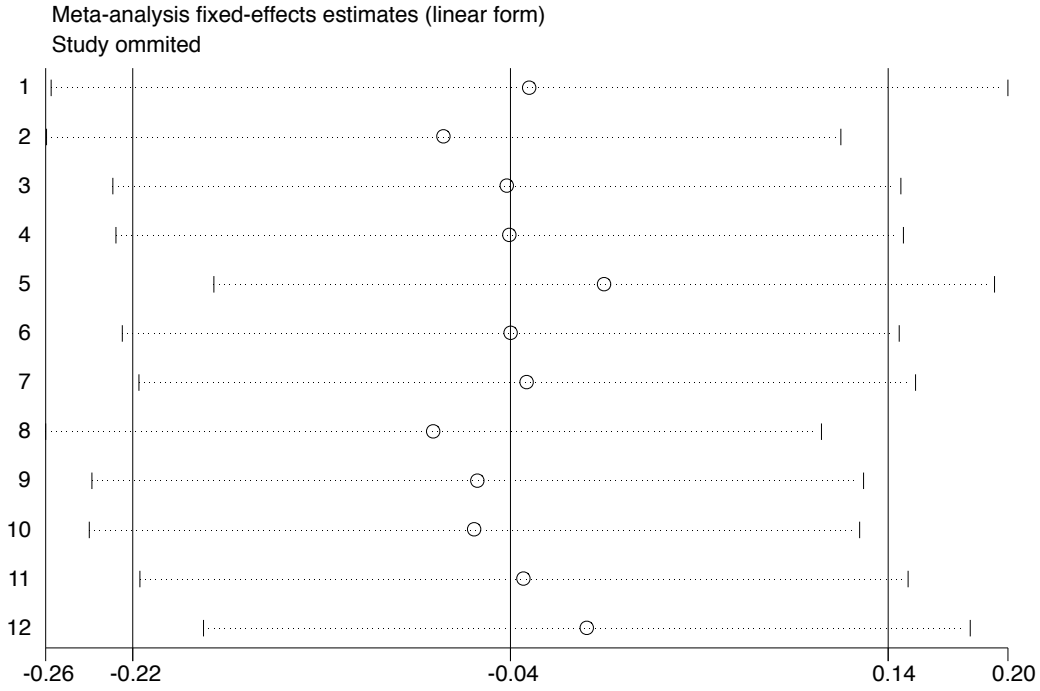


Table A1: Influence analysis for aftercare/re-entry, alleged offenses (N=1 outlier removed)

Study omitted	Effect size (ES)	95% Conf. Interval	
Stafford & Glassner, 2012	-.028	.087	.473
Bergseth & McDonald, 2007	-.069	-.217	.141
Wiebush et al., 2005_Nevada	-.039	.004	.339
Iutcovich & Pratt, 1998_Philadelphia	-.038	.026	.359
Iutcovich & Pratt, 1998_Allegheny	.006	.027	.360
Barton et al., 2008_Wisconsin	-.037	.060	.391
Wiebush et al., 2005_Colorado	-.030	.024	.354
Bouffard & Bergseth, 2008	-.074	.030	.360
Wiebush et al., 2005_Virginia	-.053	-.005	.323
Sontheimer & Goodstein, 2008	-.055	.010	.338
Greenwood et al., 1995	-.031	.008	.336
Barton et al., 2008_Alaska	-.001	026	.353
Combined	-.038	-.217	.141

Figure A3: Influence analysis (with Shelden (1999) removed)



Appendix B.

ISP, alleged offenses

Figure B1: Forest plot for intensive supervision probation, alleged offenses (N=2 outliers removed)

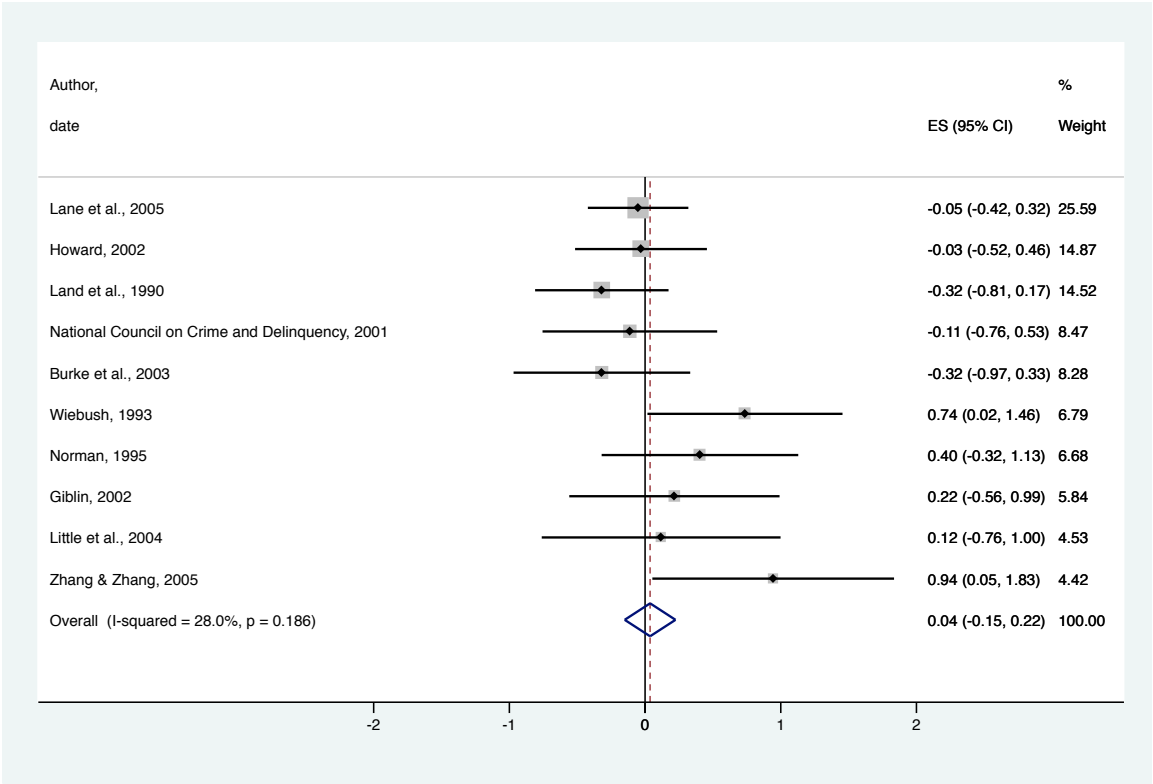


Figure B2: Forrest plot for intensive supervision probation, alleged offenses (N=2 outliers removed)

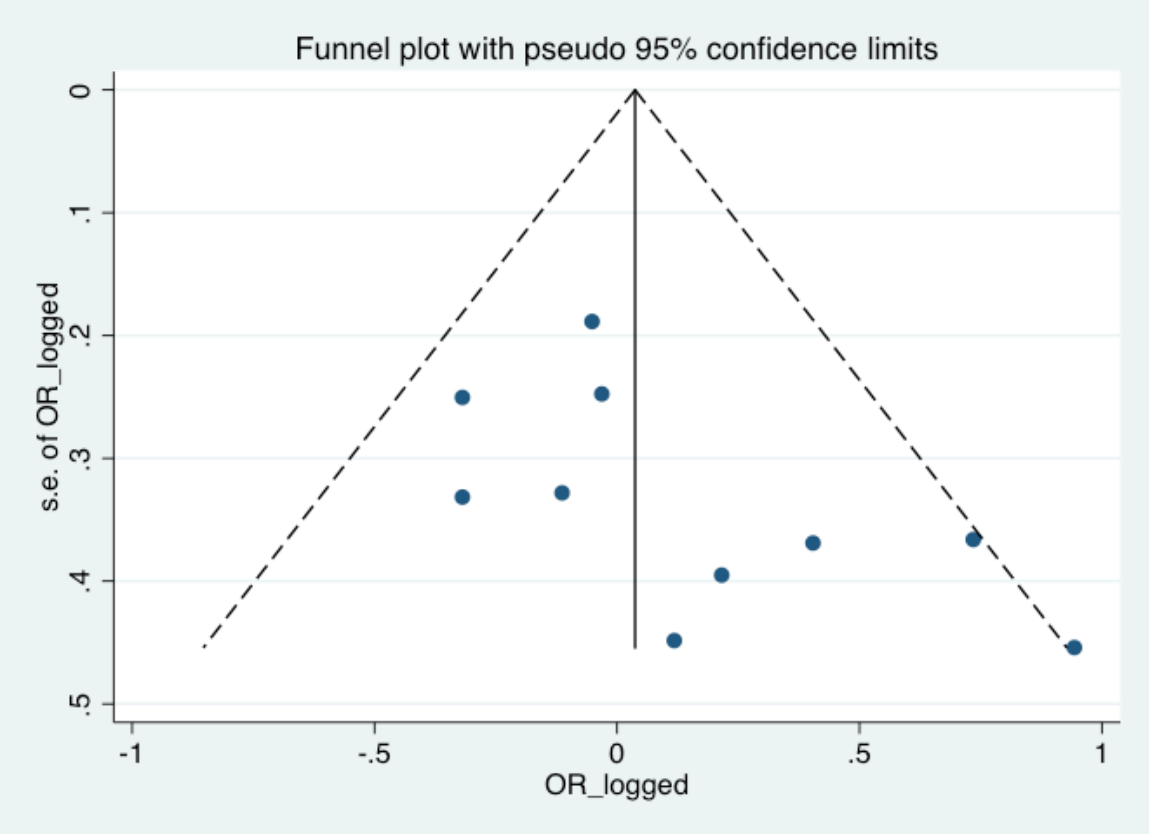
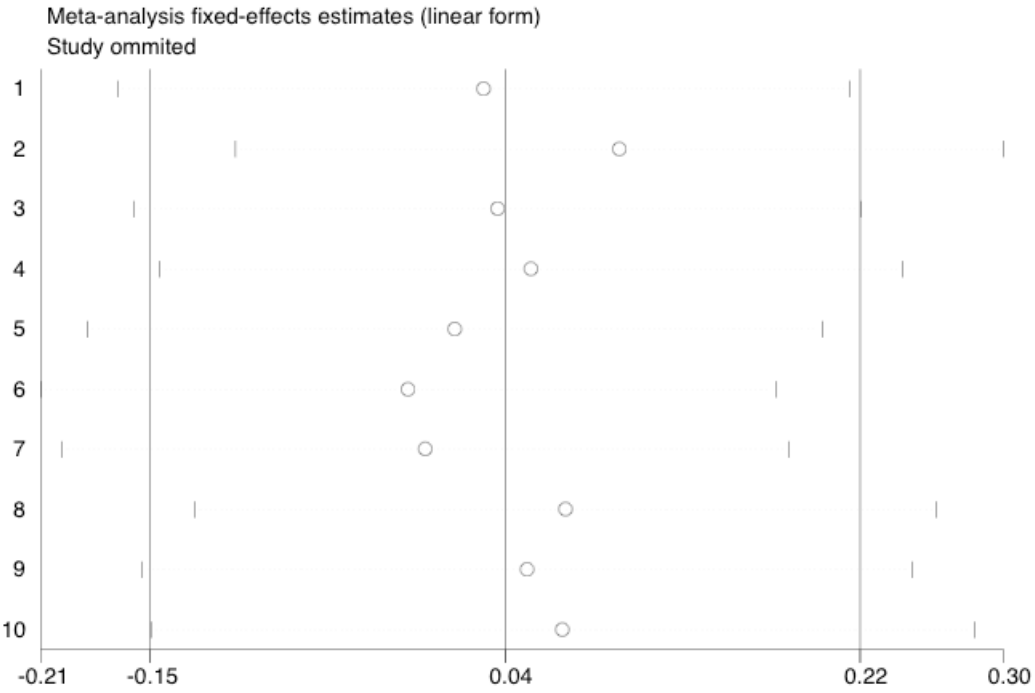


Table B1: Influence analysis with Hennigan et al. (2005) and Rodriguez-Labarca & O'Connell (2004) removed from analysis

Study omitted	Effect size (ES)	95% Conf. Interval	
Lane et al., 2005	.067	-.149	.284
Howard, 2002	.048	-.154	.251
Land et al., 1990	.097	-.104	.300
National Council on Crime and Delinq., 2001	.051	-.144	.246
Burke et al., 2003	.069	-.126	.264
Wiebush, 1993	-.013	-.207	.180
Norman, 1995	.010	-.182	.204
Giblin, 2002	.026	-.166	.219
Little et al., 2004	.033	-.158	.225
Zhang & Zhang, 2005	-.004	-.196	.186
Combined	.037	-.150	.224

Figure B3: Influence analysis for intensive supervision probation, alleged offenses (N=2 outliers removed)



Appendix C.

ISP, convicted offenses

Figure C1: Forest plot for intensive supervision probation, convicted offenses (N=3 outliers removed)

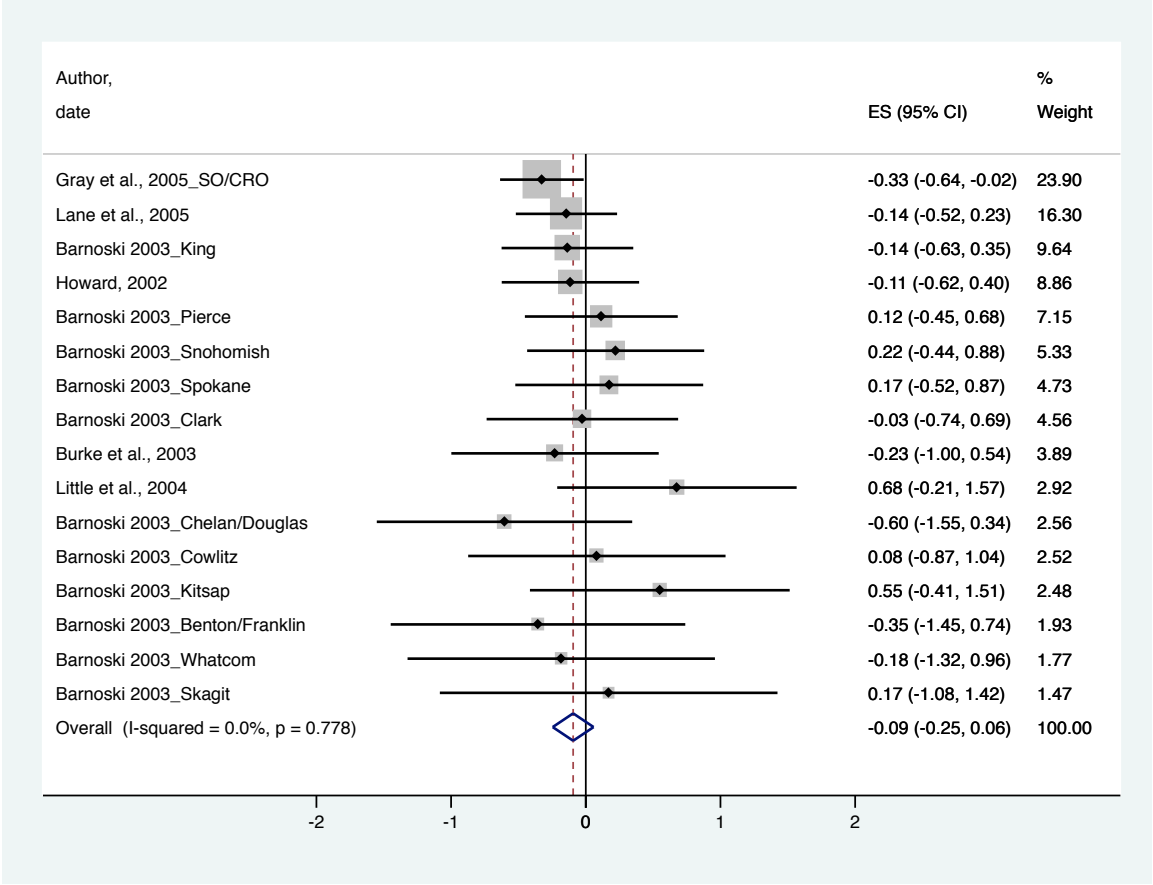


Figure C2: Funnel plot for intensive supervision probation, convicted offenses (N=3 outliers removed)

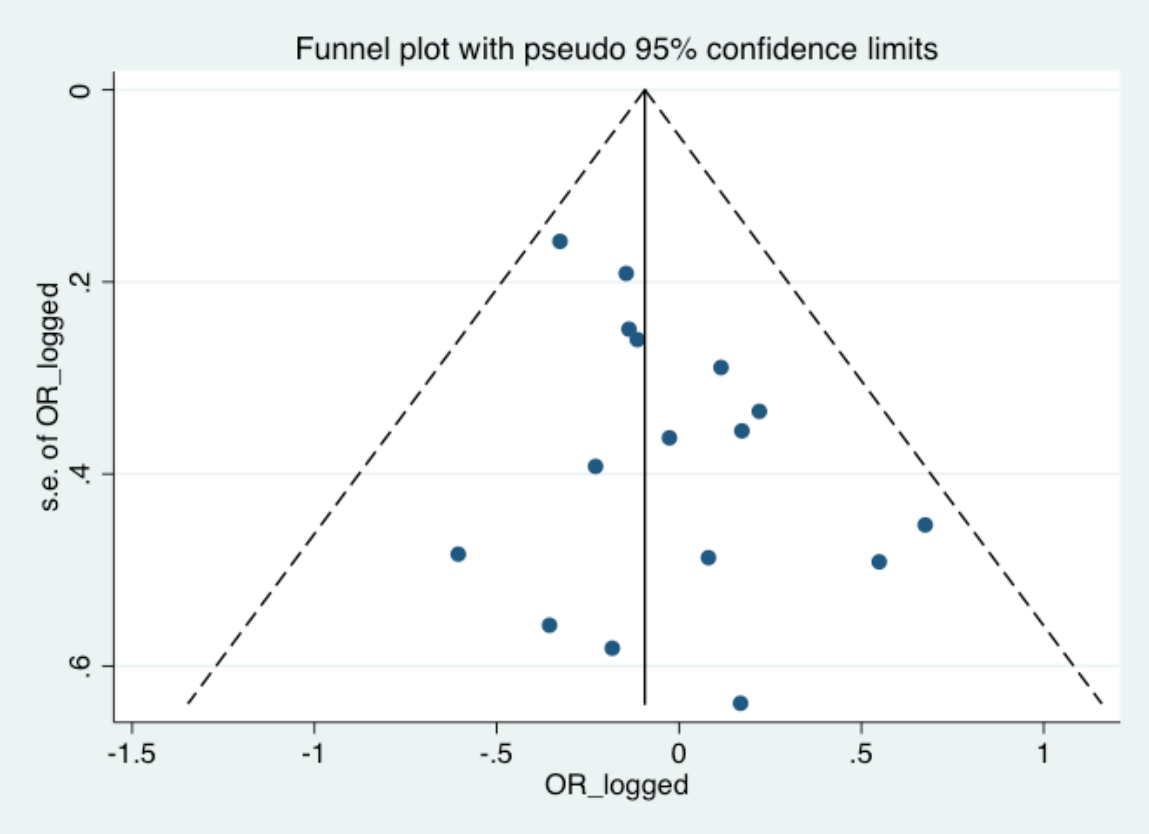
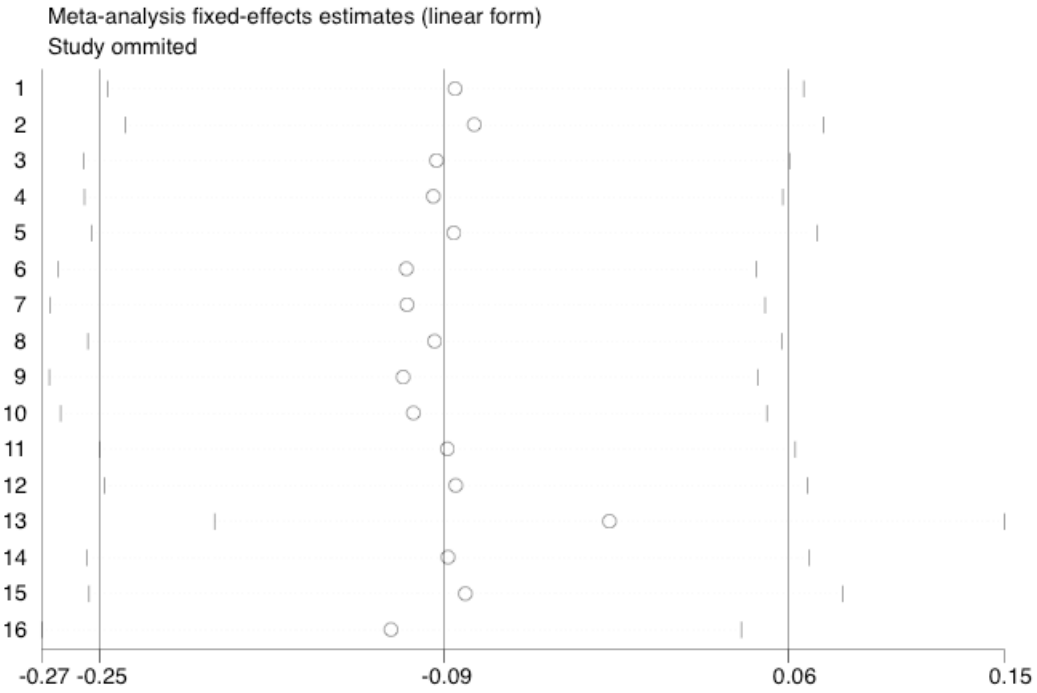


Table C1: Influence analysis for intensive supervision probation, convicted offenses (N=3 outliers removed; with Gray et al., 2005_DTO, Howitt & Moore, 2005, and Wiebush, 1993 removed from analysis)

Study omitted	Effect size (ES)	95% Conf. Interval	
Gray et al., 2005_SO/CRO	-.020	-.194	.153
Lane et al., 2005	-.084	-.250	.081
Barnoski 2003_King	-.089	-.248	.070
Howard, 2002	-.091	-.250	.067
Barnoski 2003_Pierce	-.109	-.267	.047
Barnoski 2003_Snohomish	-.111	-.267	.044
Barnoski 2003_Spokane	-.107	-.262	.048
Barnoski 2003_Clark	-.097	-.252	.058
Burke et al., 2003	-.088	-.243	.066
Little et al., 2004	-.116	-.270	.037
Barnoski 2003_Chelan/Douglas	-.080	-.234	.073
Barnoski 2003_Cowlitz	-.098	-.252	.055
Barnoski 2003_Kitsap	-.110	-.263	.043
Barnoski, 2003_Benton/Franklin	-.088	-.241	.064
Barnoski 2003_Whatcom	-.092	-.245	.060
Barnoski 2003_Skagit	-.097	-.250	.055
Combined	-.093	-.245	.057

Figure C3: Influence analysis for intensive probation supervision, convicted offenses (N=3 outliers removed)



Appendix D.

Coding Form

General study characteristics	
** denotes characteristics that must be present to be included in the analysis*	
Author (date)	No specification
Outcome number	No specification (used to keep track of number of outcomes recorded when multiple were presented)
Approach to treatment	No supervision Diversion Probation or parole*
Intervention philosophy	Surveillance* Deterrence Restorative Counseling and its variants Skill building programs Multiple coordinated services*
Type of intervention	Intensive supervision probation Aftercare/Re-entry
Subtype (primary)	Diversion without mediation Mediation Teen court Mentoring Counseling Behavioural/emotional Academic/vocational Case management* Multimodal*
Secondary services	Diversion without mediation Mediation Teen court Mentoring Counseling Behavioural/emotional Academic/vocational Case management Multimodal
Program delivery year	No specification

Publication year	No specification
Publication type	Journal article Book chapter Technical report Dissertation/thesis
Peer review	No Yes
Location	North America Europe Australia/New Zealand
Program characteristics	
Number of sessions	No specification
Number of weeks	No specification
Setting	Community only Custody and community component
Delivered by	Probation officers Counselor Program staff Students Volunteers Other Case manager Police Various community service providers
Supervision headed by a probation officer?	No Yes
Parent/family involvement?	No Yes
Risk assessment tool used?	No Yes
Participation mandatory?	No Yes
Type of offender/risk level	Low Medium-High
Sample characteristics	
Sample size	<100 >100
N treatment group	No specification
N comparison group	No specification

Participant age	No specification
Approximate mean age	No specification
Gender mix	All female Near equivalent Mostly male All male
Predominant ethnicity[14]	White/Caucasian Mixed Minority
Unit of assignment	Individual Group
Outcomes adjusted for pre-test differences?	No Yes
Direction and magnitude of initial differences between treatment and control group	No specification
Attrition from treatment and control groups on recidivism outcome	No specification
Outcome measure	
Outcome	Court contact/police contact/arrest Incarceration/institutionalization/conviction Self-report delinquency
Variable	No specification
Measurement	Dichotomous Continuous
Outcome source	Official records Self report
Follow up period	No specification

[14] The decision rule was made that when either group (those identifying as minority or white) was over 60% of the entire group of participants, that group was determined to be the predominant ethnic group (see Appendix E of Lipsey & Wilson, 2001a).

Study characteristics	
Research design	RCT QE w/matched comparison group QE w/weakly matched comparison group
Random assignment?	No Yes
Matched control group?	No Yes
Pretest?	No Yes
Quality of implementation: any problems noted	No problems Minor problems Major problems
Quality of intervention: level of researcher involvement	Research only Planning/supervising intervention Delivering program directly
Not applicable	99
Missing data	999

Appendix E.

Overview of included studies

*Note that studies in each measure of recidivism are not independent of each other

Study characteristic	ISP N (%) Total N = 31	Aftercare N (%) Total N = 24
Publication year		
1990-1994	4 (12.90%)	3 (12.50%)
1995-1999	1 (3.23%)	6 (25.00%)
2000-2004	20 (64.52%)	0 (0.00%)
2005-2009	6 (19.35%)	14 (58.33%)
2010-2014	0 (0.00%)	1 (4.17%)
Publication type		
Journal article	10 (32.26%)	2 (8.33%)
Report	21 (67.74%)	22 (61.67%)
Location		
United States	27 (87.10%)	24 (100.0%)
England	4 (12.90%)	0 (0.0%)
Program delivery year		
1980-1989	4 (12.90%)	3 (12.50%)
1990-1999	18 (58.06%)	12 (50.00%)
2000-2010	9 (29.03%)	9 (37.50%)
Type of research design		
Randomized control trial	25 (80.65%)	9 (37.50%)
Quasi-experiment with matched comparison group	4 (12.90%)	6 (25.00%)
Quasi-experiment with weakly matched comparison group	2 (6.45%)	9 (37.50%)
Outcome measure		
Criminal contact/Arrest	12 (38.71%)	11 (45.83%)
Conviction/Incarceration	19 (61.29%)	13 (54.17%)
Follow-up period		
<1 yr	3 (9.68%)	3 (12.50%)
1 yr	9 (29.03%)	14 (58.33%)
>1 yr	18 (58.06%)	5 (20.83%)
missing	1 (3.23%)	2 (8.33%)

Sample size in treatment group		
Less than 100	14 (45.16%)	18 (75.00%)
100+	17 (54.84%)	6 (25.00%)
Sample gender mix		
All male	0 (0.00%)	12 (50.00%)
Mostly male	13 (41.94%)	12 (50.00%)
Nearly equivalent	0 (0.00%)	0 (0.00%)
Mostly female	1 (3.23%)	0 (0.0%)
All female	3 (9.68%)	0 (0.0%)
missing	14 (45.16%)	0 (0.0%)
Sample race/ethnicity		
Predominantly Caucasian	3 (9.68%)	0 (0.0%)
mixed	6 (19.35%)	4 (16.67%)
Predominantly minority	8 (25.81%)	20 (83.33%)
missing	14 (45.16%)	0 (0.0%)
Total sample size		
<100	7 (22.58%)	5 (20.83%)
>100	24 (77.42%)	19 (79.17%)

Appendix F.

Grey literature hand searched journals

American Journal of Evaluation
Annals of the American Academy of Political and Social Science
Canadian Journal of Criminology
Canadian Journal of Program Evaluation
Crime and Delinquency
Criminal Justice and Behavior
Criminal Justice Review
Criminology
Criminology & Public Policy
Evaluation (SAGE)
Evaluation and Program Planning
Evaluation Review
Federal Probation
Journal of Experimental Criminology
Journal of Quantitative Criminology
Journal of Research Crime and Delinquency
Justice Quarterly

Appendix G.

Description of studies included in meta-analysis (aftercare/re-entry)

Study	Program name (location)	Study design (Maryland Scale rating)	Sample size	Follow-up	Outcome measure(s) reported	Outcome
Barton et al., 2008 (3 sites)	The Boys and Girls Club of America (Alaska)	QE with matched comparison group (4)	148	12 months	Arrest (alleged)	Favoured control group
					Conviction	
	The Boys and Girls Club of America (Arkansas)		172		Conviction	Favoured treatment group
	The Boys and Girls Club of America (Wisconsin)		131		Arrest (alleged)	Favoured control group
					Conviction	Favoured treatment group
Bergseth & MacDonald, 2007	The Reentry Services Project (RSP) (Clay County, MN)	QE with weakly matched comparison group (3)	184	12 months	Criminal contact (alleged)	Favoured treatment group
Bouffard & Bergseth, 2008	n/a	QE with weakly matched comparison group (3)	112	6 months	Criminal contact (alleged)	Favoured treatment group
Greenwood et al., 1993	The Skillman Intensive Aftercare Program Initiative (Detroit, MI)	Randomized controlled trial (5)	99	12 months	Arrest (alleged)	Favoured control group
					Conviction	Favoured treatment group

Hawkins et al., 2009	Serious and Violent Offender Re-entry Initiative (SVORI)	QE with matched comparison group (4)	276	15 months	Incarceration (convicted)	Favoured control group
Iutcovich & Pratt, 1998 (2 sites)	Philadelphia Opportunities Industrialization Center, Youth Advocacy Plus Program (OIC-YAPP) (Philadelphia, PA)	QE with weakly matched comparison group (3)	173	18 months	Arrest (alleged)	Favoured control group
				12 months	Incarceration (convicted)	
	Abraxas Non-Residential Care (Abraxas NRC) (Allegheny county, PA)		130	18 months	Arrest (alleged)	Favoured control group
				12 months	Incarceration (convicted)	
Shelden, 1999	Detention Diversion Advocacy Project (DDAP) (San Francisco, CA)	QE with weakly matched comparison group (3)	542	n/a	Petition to court (alleged)	Favoured treatment group
					Incarceration (convicted)	
Sontheimer & Goodstein, 1993	Intensive Aftercare Probation (IAP) (Philadelphia, PA)	Randomized controlled trial (5)	75	6 months	Arrest (alleged)	Favoured treatment group
Stafford & Glassner, 2012	The Children's Aftercare Re-Entry Experience (CARE) (San Antonio, TX)	QE with weakly matched comparison group (3)	850	12 months	Arrest (alleged)	Favoured treatment group
Wiebush et al., 2005 (3 sites)	Intensive Aftercare Program (IAP) (Colorado)	Randomized controlled trial (5)	118	12 months	Arrest (alleged)	Favoured control group
					Conviction	
	Intensive Aftercare Program (IAP) (Nevada)		220		Arrest (alleged)	Favoured control group
					Conviction	Favoured treatment group
	Intensive Aftercare Program (IAP) (Virginia)		97		Arrest (alleged)	Favoured treatment group
					Conviction	

Appendix H.

Description of studies included in the meta-analysis (ISP)

Study	Program name (location)	Study design (Maryland Scale rating)	Sample size	Follow-up	Outcome measure(s) reported	Outcome
Barnoski, 2003 (11 sites)	Early Intervention Program (EIP) (Washington state)	Randomized controlled trial (5)	Ranged from 61 to 294	36 months	Conviction	Favoured control group at 5 locations; favoured treatment group at 6 locations
Burke et al., 2003	Working to Insure and Nurture Girls' Success or WINGS (San Diego, CA)	Randomized controlled trial (5)	416	12 months	New charge (alleged)	Favoured control group
					Sustained petition (conviction)	Favoured control group
Giblin, 2002	The Anchorage Coordinated Agency Network (CAN) (Anchorage, AK)	Randomized controlled trial (5)	155	7 months	New charge (alleged)	Favoured treatment group
Gray et al., 2005_SO/CRO	Intensive Supervision and Surveillance Programme (ISSP) (England & Wales)	QE with matched comparison group (4)	1561	12 months	Conviction	Favoured control group
Gray et al., 2005_DTO	Intensive Supervision and Surveillance Programme (ISSP) (England & Wales)	QE with matched comparison group (4)	667	12 months	Conviction	Favoured control group
Hennigan et al., 2005	The Youth and Family Accountability Model (YFAM) (Los Angeles, CA)	Randomized controlled trial (5)	1318	12 months	Arrest (alleged)	Favoured treatment group
Howard, 2002	The Repeat Offender Prevention Program (ROPP) (San Diego, CA)	Randomized controlled trial (5)	327	12 months	Sustained petition (conviction)	Favoured control group
Howitt & Moore, 1991	The Early Offender Program (EOP)	QE with matched comparison group (4)	82	n/a	Adjudication (conviction)	Favoured treatment group

Land et al., 1990	North Carolina Court Counselor's Intensive Protective Supervision Project (IPSP)	Randomized controlled trial (5)	174	18 months	Court referral (alleged)	Favoured control group
Lane et al., 2005	South Oxnard Challenge Project (SOCP) (Los Angeles county, CA)	Randomized controlled trial (5)	462	18 months	Sustained petition (conviction)	Favoured control group
Little et al., 2004	Intensive Supervision and Support (ISSP) (England)	Randomized controlled trial (5)	55	24 months	Arrest (alleged) Conviction	Favoured treatment group
National Council on Crime and Delinquency, 2001	Reaffirming Young Sister's Excellence (R.Y.S.E.) (Alameda County, California)	Randomized controlled trial (5)	427	6 months	Arrest (alleged)	Favoured control group
Norman, 1995	The Utah Second District Juvenile Court Intensive Supervision Probation Program	Randomized controlled trial (5)	133	12 months	Arrest (alleged)	Favoured treatment group
Rodriguez-Labarca & O'Connell, 2004	Delaware's Serious Juvenile Offender program (SJO)	QE with matched comparison group (4)	336	12 months	Arrest (alleged)	Favoured control group
Wiebush, 1993	Lucas County Intensive Supervision Unit (ISU) (Toledo, Ohio)	QE with weakly matched comparison group (3)	168	18 months	Charged (alleged) Adjudication (conviction)	Favoured control group
Zhang & Zhang, 2005	The Los Angeles County Repeat Offender Prevention Program (ROPP)	Randomized controlled trial (5)	204	6 months	Arrest (alleged)	Favoured treatment group