University of Cincinnati

Date: 3/28/2016

I, Jennifer L. Lux, hereby submit this original work as part of the requirements for the degree of Doctor of Philosophy in Criminal Justice.

It is entitled:

Assessing the Effectiveness of Multisystemic Therapy: A Meta-Analysis

Student's name: **Jennifer L. Lux**

This work and its defense approved by:

Committee chair: Francis Cullen, Ph.D.

Committee member: Cheryl Lero Jonson, Ph.D.

Committee member: Edward Latessa, Ph.D.

Committee member: Paula Smith, Ph.D.

19819

Assessing the Effectiveness of Multisystemic Therapy:
A Meta-Analysis

A Dissertation Submitted to the
Division of Research and Advanced Studies
of the University of Cincinnati

In Partial Fulfillment of the Requirements of the Degree of

Doctorate of Philosophy (Ph.D.)

In the School of Criminal Justice
of the College of Education, Criminal Justice, and Human Services

2016

by

Jennifer L. Lux
B.S. Butler University, 2008
M.S. University of Cincinnati, 2009

Dissertation Committee: Francis T. Cullen, Ph.D. (Chair)
Edward J. Latessa, Ph.D.
Paula Smith, Ph.D.
Cheryl Lero Jonson, Ph.D.

ABSTRACT

There can be little doubt that Scott Henggeler's intervention program—multisystemic therapy (MST)—counts as a major criminological invention. Rooted in the science of risk factors and theory based in community psychology, MST seeks to target for reform antisocial youths at high-risk of out-of-home placement. Due to the quality of the intervention and to Henggeler's work over three decades to implement and market the program, MST has been adopted across many contexts.

Several independent studies and large-scale reviews of the program have been conducted over MST's more than 30-year lifespan. These studies, however, have found mixed results and one has even called into question the efficacy of MST altogether. Even further, a recent meta-analysis of MST called attention to the importance of moderating factors or characteristics that may help to explain why mixed results are being seen. In this context, the overall goal of the current dissertation was to advance the literature on the effectiveness of MST and determine more in depth, what factors or circumstances lead to the strongest treatment effects for the program. The objectives of this dissertation were threefold.

First, this dissertation involved the analysis of 44 unpublished and published studies, providing the most comprehensive review of studies on the effectiveness of MST to-date. Second, the effectiveness of MST for youths and their families on nine dependent variables was examined, including: (1) delinquency, (2) problem behavior, (3) psychopathology and mental health, (4) family functioning and relationships, (5) peer relationships, (6) school performance, (7) parent functioning and relationships, (8) substance abuse, and (9) service utilization. Third, to determine the effectiveness of MST on the nine outcome variables, a meta-analysis was conducted. From this quantitative synthesis process, the overall mean effect size, weighted mean

effect size, and corresponding confidence intervals were calculated for the independent variable (i.e., MST) and each dependent variable under review. In addition, the impact of several moderating variables was explored. Moderators were examined across five general categories, including: (1) publication characteristics, (2) study context characteristics, (3) sample characteristics, (4) treatment characteristics, and (5) study characteristics.

The results indicated that, on average, MST treatment significantly reduced delinquency by 11 percent and problem behavior by 15 percent, as well as improved juvenile psychopathology and mental health by 15 percent, family functioning and relationships by 14 percent, parent functioning and relationships by 15 percent, and service utilization by 21 percent. Although treatment effects were in favor of MST for peer relationships, school performance, and substance abuse, the results failed to reach statistical significance. The findings also suggested that MST effectiveness may be moderated by a number of publication, study context, sample, treatment, and study characteristics.

MST is a carefully designed program that is used widely and, according to the current dissertation and other systematic reviews, reduces delinquency and problem behavior, while at the same time, improving a variety of other juvenile behaviors. It is modestly successful overall, and in some instances, more than modest effects have been demonstrated. The challenge for MST is how to proceed in the future. One promising avenue is for MST to embrace the risk-need-responsivity model and create a new version that is a hybrid of the two.

ACKNOWLEDGEMENTS

I thought about this moment so many times over the last few weeks that I started to lose count; the moment when I would write the last word, on the last page, of the last chapter, and the only thing left to do was scroll to the top to acknowledge all of the people who helped me get here. The problem is, that's a lot of people and this dissertation is already quite long! Suffice it to say that if you have been around me in any capacity over the last year or so, you have likely influenced this dissertation in some way. Whether you said something that sparked a moment of creativity for me to write two more pages at 11:00 at night or offered support on the days where I thought I would never see the light at the end of the tunnel, you deserve a big thank you!

There are a handful of people I would specifically like to acknowledge. First, to my dissertation chair, Dr. Cullen. I do not know where I would be or what I would be doing if I had not taken your corrections course seven years ago. You helped me discover a passion for something I did not know I had and showed me how to turn that passion into a career. Thank you for being my cheerleader when others were not always rooting for my team and for your unwavering support as I wrote each section of this dissertation. While it was slow and painful at times (for both of us, I'm sure), I am so grateful to have had your mentorship, words of wisdom, and occasional "nudge" throughout the entire process.

I must also thank the rest of my dissertation committee. To Dr. Latessa. Thank you for always putting things in perspective and for reminding me that once I finished this dissertation, I would never have to do it again. Your encouragement, motivation, and constant reminder to keep things simple helped me to keep pushing through. I must also thank Dr. Smith. This dissertation would not have been possible if it wasn't for your training and guidance on the complexities of meta-analysis. Thank you for your patience and willingness to respond to my

panicked text messages at all hours of the night. Finally, to Dr. Jonson. Thank you for starting this meta-analysis journey with me. The “introductory course” you offered me after completing your dissertation motivated me to start working on my own. I am grateful for your coaching and support throughout, not only this dissertation process, but throughout my graduate school career.

To the UCCI team. You all are so much more to me than co-workers. You are my teammates, comic relief when things get crazy, and most of all, my family away from home. Thank you for “holding down the fort” while I disappeared to finish writing this dissertation. I owe you all big time! I especially want to thank Mindy and Jen Scott for sending me text messages approximately every three hours for 10 days straight to check in on me and send me words of encouragement. Also, to Steph. Thank you for being by my side for every “big” moment in the program. From stats, to the comp, to the last few days of getting this thing done, you have been an absolute blessing throughout this journey.

To my parents. Thank you for showing me early on that hard work and dedication can get your miles ahead in life. I know when I told you I was going to grad school, you were not quite sure what to think, but you two, along with Morg, and Gav, have been my #1 fans ever since. Thank you for never losing faith in me and for believing in me even more than I believed in myself at times. I feel so blessed to be able to call you guys my family.

Finally, to my fiancé TJ. Thank you for hanging on for "just one more year" for the last three years. I am sure when I started this program, you had no idea that you would eat, breathe, and sleep it almost just as much as I would. Your patience and ability to stay calm in the wake of my never ending thunderstorm is something I thank God for every single day. I know I owe you a whole lot of apartment cleaning, dinner making, and most importantly, wedding planning! Here's to us and our new "no dissertation" life.

TABLE OF CONTENTS

LIST OF TABLES AND FIGURES

CHAPTER 1

MULTISYSTEMIC THERAPY AS A CRIMINOLOGICAL INVENTION: ORIGINS AND BEYOND

As Sherman (2011, p. 438) notes, "Criminology is a field of invention just as much as it is a field of research and theory-building." At its best, its inventions can produce "less crime and injustice" and, more broadly, reduce "human suffering" (p. 424). But Sherman adds a word of caution, noting that criminology's inventions not only can do "great good" but also "great harm" (p. 438). Part of the field's responsibility is thus to avoid what Finckenauer (1982) called the "panacea phenomenon"—of jumping onboard a crime-control bandwagon because it seems intuitively plausible and ideologically pleasing. Instead, it is important to embrace the core scientific norm of "organized skepticism" (Merton, 1973), which involves subjecting even our most cherished findings to rigorous empirical scrutiny.

In this context, there can be little doubt that Scott Henggeler's intervention program—multisystemic therapy, know widely by its acronym "MST"—counts as a major criminological invention. Rooted in the science of risk factors and theory based in community psychology, MST seeks to target for reform antisocial youth at high-risk of institutional placement. Due to the quality of the intervention and to Henggeler's work over three decades to implement and market the program, MST has been adopted across many contexts. Currently, MST operates in 34 states and 15 countries, treating more than 23,000 youths a year (MST Services, Inc., 2016).

MST has been hailed as an effective intervention. The program is listed in the Office of Justice Program's website, CrimeSolutions.gov, as an "effective program," and earlier was selected for inclusion in the *Blueprints for Violence Prevention* series that, in Delbert Elliott's (1998, p. xxi) words, identifies "a core set of programs that meet very high scientific standards

for being effective prevention programs." Criminologists have added their support, regularly defining MST as a program that "works" (Cullen & Gendreau, 2000; Farrington & Welsh, 2007; Greenwood, 2006; MacKenzie, 2006).

This seeming consensus on MST's effectiveness, however, was punctured by a systematic review, which included a meta-analysis, conducted by Julia Littell (2005)—an assessment sufficiently compelling to prompt scholars to note that the robustness of MST's impact on antisocial conduct should be viewed with a measure of caution (Farrington & Welsh, 2007; MacKenzie, 2006). Littell's central claim is that methodological weaknesses made it difficult to conclude that MST is a model program that should be widely adopted. MacKenzie (2006, p.178) offers a cogent summary of Littell's critique:

> Littell carefully examined the MST studies over time...[S]he found problems related to randomization procedures, sample sizes, unyoked designs, unstandardized observation periods within studies, and systematic omission of those who refused treatment or did not complete MST. When she controlled for these problems in a meta-analysis she concluded that the effectiveness of the programs was not well established, although she found no evidence that MST had harmful effects or was less effective than other services. Thus, this is a cautionary note about the effectiveness of MST.

In this context, this dissertation was undertaken primarily to provide a more comprehensive assessment of MST evaluation studies. To date, meta-analyses of the MST evaluation literature have been based on a limited number of studies; this also is true of narrative reviews of this body of evidence. The reviews have relied extensively on studies conducted by Henggeler and his associates, often on studies they directed or helped to implement. By contrast, I have searched both for published and unpublished studies. The result is a richer database of

research that includes evaluations undertaken by a diverse group of scholars (although still heavily weighted toward studies by Henggeler and associates).

As neither an advocate nor opponent of MST, I have, as the saying goes, "no dog in this hunt." Rather, the interest in this evaluation was sparked largely by the desire to confirm—or not confirm—the effectiveness of an intervention program that is publicized as "working" and is being used with thousands of troubled youngsters. The extant reviews of MST evaluation studies are mostly qualitative and, again, based on a low number of studies. As suggested by the core scientific norm of organized skepticism (Merton, 1973), it seemed that a more extensive quantitative synthesis of the literature was in order.

To introduce the dissertation, the current chapter considers three issues. The first section "tells the story" of the origins of MST, focusing on how Scott Henggeler came to invent the program. The second section conveys the core elements of MST developed by Henggeler. The third section reviews the existing evidence on MST's effectiveness. This discussion will set the context for justifying the need for a systematic meta-analysis of evaluations of MST—the main purpose of this dissertation.

MST AS A CRIMINOLOGICAL INVENTION: THE STORY OF SCOTT HENGGELER

MST began in 1978 when Scott Henggeler was a professor at Memphis State University in the psychology department. With the help of his psychology students, Henggeler was asked to develop a diversion program for juvenile offenders (Cullen, 2005). Although developing such a program was a difficult task in and of itself, the team was faced with an even greater challenge—creating a program for antisocial youth during a time when support for treatment and rehabilitation was lacking. More specifically, Robert Martinson's controversial "nothing works" article was published in the spring of 1974, sparking an era of "anti-rehabilitation" that reigned

well into the late 1970s and early 1980s when Henggeler and his team were assigned the "diversion project."

Briefly, Martinson's essay, "What Works? Questions and Answers about Prison Reform," reviewed 231 evaluation studies examining the effectiveness of correctional treatment programs between 1945 and 1967. From his review, Martinson ultimately concluded that "with few and isolated exceptions, the rehabilitative efforts that have been reported so far have had no appreciable effect on recidivism" (1974, p. 25). He went on to leave readers with one final thought-provoking question: "Do all of these studies lead irrevocably to the conclusion that nothing works, that we haven't the faintest clue about how to rehabilitate offenders and reduce recidivism?" (p. 48). Martinson stopped short of answering his own question, but it was clear what he was suggesting—that "nothing works" to rehabilitate offenders (Cullen & Gendreau, 2001). Notably, "the response to Martinson's article was pervasive and consequential, outstripping anything modern criminology had seen in recent, or, for that matter, in distant memory" (Cullen, 2005, p. 6). In this way, most policymakers, correctional scholars, and administrators accepted Martinson's findings and embraced the "nothing works" movement, rather than asking for more evidence or research to bolster Martinson's findings.

But why did Martinson's essay have such a powerful and immediate impact? (Palmer, 1992). He was hardly the first person to question the effectiveness of correctional treatment (Cullen & Gendreau, 2000). Donald Cressey (1958, p. 77) argued, for example, "most of the 'techniques' used in 'correcting' criminals have not been shown to be either effective or ineffective and are only vaguely related to any reputable theory of behavior or criminality." Similarly, Berleman and Steinburn (1969, p. 471) noted "uniformly disappointing results" in their review of five major youth programs. And, in January 1974—just a few months before

Martinson's essay was published—Gold (1974, p. 22) advised that it was "time for skepticism" because "the best data at hand demonstrate that we have not yet solved the problem of the effective treatment of delinquency." As Cullen (2005, p. 6) explains, "One critical factor was timing." Martinson's article came at a time when the field of criminology was experiencing an ideological shift—a shift that could be attributed to the social and political turmoil the United States (U.S.) had experienced in the decade prior (Cullen & Gendreau, 2001).

Specifically, from the mid-1960s to the mid-1970s, the country had seen civil rights marches, urban unrest, protests over the Vietnam War, and the Watergate scandal. Not surprising, crime rates were also rising during this time, and in 1971, the Attica Prison riot occurred, forcing law enforcement officials to storm the institution and shoot down both guards and inmates (Cullen & Jonson, 2012). The confluence of these events led to distrust in the government and "caused people to rethink many issues, including the nature of the correctional system" (Cullen & Jonson, 2012, p. 35).

Rehabilitation became the likely scapegoat for both liberals and conservatives to blame for the lack of control and order that had taken hold across the criminal justice system. While the two political parties had differing views regarding what the new correctional ideology should be (Cullen & Gendreau, 2001), they both supported offender punishment and argued that correctional officials had too much discretionary power. As part of the campaign for less discretion, for example, "states began to question indeterminate sentencing and to call for sentencing in which judicial and parole board discretion was eliminated or, at the least, curtailed" (Cullen & Jonson, 2012, p. 33).

Liberals and conservatives' political platforms parted ways, however, on a number of issues. First, liberals argued that the state had become too harsh on offenders. Where once it was

believed that judges used their discretionary power to individualize treatment for the wayward, this power was now seen as one that allowed them to hand out harsher sentences to poor and minority defendants (Cullen, 2005; Cullen & Jonson, 2012). In a similar vein, correctional officials were using indeterminate incarceration as an opportunity to coerce and manipulate offenders into complying with institutional rules, rather than using such sentences to achieve the goal of reformation (Cullen & Jonson, 2012). In short, leftist critics proposed a "justice model" where the guiding philosophy of corrections would be "just deserts" and rehabilitation—if used at all—would be voluntary (Cullen & Gilbert, 2013).

On the other hand, conservatives argued that the criminal justice system had become *too* lenient on offenders, allowing dangerous criminals to remain in the community on nothing more than probation and giving parole boards the discretionary power to release inmates early from prison (Cullen & Jonson, 2012). In this way, conservatives wanted to "get tough" on crime, proposing a "law and order" model that would increase the amount of time offenders spent behind bars (i.e., implement determinant sentencing) and decrease the discretionary powers given to judges and other corrections officials (Cullen & Gilbert, 2013).

As discussed, criminologists and other correctional critics were also influenced by the disorder that had taken place in the larger society, and much like the government, began to doubt the rehabilitative ideal that had been the dominant correctional theory for 150 years. Even in the aftermath of Martinson's publication when his conclusions were challenged by a handful of scholars, the majority of criminologists turned their backs to any evidence that supported rehabilitation and its success in reducing recidivism. Coupled with the social and political context of the times, Martinson's "message confirmed what critics 'already knew' and gave them

a weapon—scientific data—to back up their attack on correctional treatment" (Cullen, 2005, p. 7).

While criminologists continued to "declare rehabilitation dead" (Cullen & Jonson, 2012, p. 35), scholars outside the field of criminology challenged Martinson's "nothing works" doctrine (Cullen & Gendreau, 2001). For example, psychologist, Ted Palmer who worked for years as a Senior Researcher for the California Youth Authority (Cullen & Jonson, 2012), had seen evidence of rehabilitation working with delinquent youths. He was one of the first individuals to not only question Martinson's findings, but went one step further when he carefully reanalyzed Martinson's data. As such, three important considerations came from his (1975) review. First, "only 80 of the studies Martinson reviewed—not 231 as commonly believed—actually examined the impact of treatment interventions on recidivism" (Cullen & Jonson, 2012, p. 34). In other words, included in the original 231 studies identified by Martinson as being studies of treatment effectiveness, were studies that did not actually measure "treatment" (e.g., they measured simply being on probation, parole, or being incarcerated) (Cullen & Jonson, 2012).

Second, about half of the studies reviewed by Palmer "showed that the intervention actually reduced recidivism" (Cullen & Jonson, 2012, p. 35). More specifically, 39 out of the 80 treatment studies (48%) yielded "positive or partly positive results" (Palmer, 1975, p. 142). Further, Martinson categorized the studies into eleven different treatment methods (e.g., probation, individual psychotherapy, milieu therapy) and discovered that within these categories, some treatment strategies worked, while others did not (Cullen, 2005). In this way, Martinson (1974, p. 49) argued that there was "little reason to hope that we have in fact found a sure way of

reducing recidivism through rehabilitation" because there was not one treatment typology or category that worked as a whole (Cullen, 2005).

Third, Martinson only examined studies conducted between 1945 and 1967. "But subsequent reviews of more recent literature—including one by Martinson (in 1979) himself—suggest that many programs do, in fact, 'work' to lower the risk of offenders returning to crime" (Cullen & Jonson, 2012, p. 35).

At the end of Palmer's (1975) rebuttal, he urged scholars to take a closer look at the analytical framework used in Martinson's study and argued that future research examine, not only the type of treatment modality being used with offender populations, but also examine other variables, such as offender and correctional staff characteristics and the setting in which treatment takes place (Cullen, 2005). In this way, the research question becomes not "what works for offenders as a whole" but "which methods work best for which types of offenders, and under what conditions or in what type of setting" (Palmer, 1975, p. 150).

Similar to Palmer, a group of Canadian psychologists continued to advocate for offender rehabilitation by further examining Martinson's findings. Specifically, Paul Gendreau and Robert Ross published two important narrative reviews of the extant empirical literature, both of which revealed the effectiveness of treatment (Gendreau & Ross, 1979; Gendreau & Ross, 1987).

In addition to providing "bibliotherapy for cynics," (Gendreau & Ross, 1979) and "reviving rehabilitation" (Gendreau & Ross, 1987), Gendreau and Ross's articles revealed four other important findings. First, they found that behaviorally oriented programs showed the most promising effects for reducing recidivism—a treatment typology Martinson did not include in his original study (Cullen & Jonson, 2012). Second, programs that "worked" were markedly

different from programs that did not work in that they targeted criminogenic needs. In other words, programs that worked focused on the characteristics about a person that research has shown are related to offender recidivism. Further, these needs are dynamic in nature in that they can be changed (e.g., antisocial personality, substance abusing behaviors) (Cullen & Jonson, 2012). Third, the Canadians discovered that like all other human behavior, "criminal behavior is learned" (Gendreau & Ross, 1979, p. 463). This was a simple, yet important implication, as inherent in Martinson's "nothing works" message was that criminal behavior could not be altered or unlearned. Similarly, it was implied that prosocial behavior could not be re-learned or acquired.

This discussion helps to set the context for understanding Henggeler's entry into the treatment field. Henggeler was not a criminologist and thus was not in a discipline where the "nothing works" view was hegemonic. Rather, his professional ideology followed that of Palmer and the Canadians, for he too was a psychologist whose training "provided an intellectual foundation for a belief in the science of behavioral change" (Cullen, 2005, p. 16). In this way, Henggeler did not jump onboard Martinson's "nothing works" bandwagon and claim that juvenile delinquents were beyond redemption (Cullen & Jonson, 2012). Rather, he argued that interventions failed with at-risk youths because they were based on the wrong model or theoretical underpinnings (Cullen, 2005).

Specifically, the prevailing treatment model at the time was psychotherapeutic in nature and followed principles of psychodynamic theory. Within this framework, clients were prompted to gain "insight" into their problems by examining past events and actions that might have caused or were contributing to the current problems (Burns, Schoenwald, Burchard, Faw, and Santos, 2000). Youths (and sometimes their parents) would come in to an office for weekly

one-hour visits (Cullen & Jonson, 2012). "The length of treatment was open ended and could last months, if not years, on end. This approach often proved costly, was ineffective in reducing antisocial conduct, and often failed to reach inner-city youths" (Cullen & Jonson, 2012, p. 195).

Henggeler's training as a community psychologist helped him understand why a psychotherapeutic approach was ineffective in changing antisocial behavior. Specifically, the community psychology movement, and eventually the tenets of MST, was built upon Bronfenbrenner's (1979) theory of social ecology. Three central features of the theory set the context for Henggeler's invention of MST.

First, a central feature of social ecology theory pertains to the multidetermined nature of human behavior. Similar to a set of Russian dolls (Bronfenbrenner, 1979, p. 3), an individual's ecological environment is comprised of a set of nested structures, each inside the next. The developing person is at the innermost core, with each concentric layer representing an important system in which the person is enmeshed (Henggeler, Schoenwald, Borduin, Rowland, & Cunningham, 2009).

Regarding adolescent functioning, the theory suggests that behavior, including both prosocial and antisocial conduct, is influenced by the "interplay among important aspects of the youth's life, such as family, friends, school, and neighborhood" (Henggeler et al., 2009, p. 15). In this way, clinical assessment is important, because specific risk factors can vary from individual to individual. Further, the assessment must take into account a wide variety of possible contributors to behavior problems both within systems (e.g., lack of parental supervision and drug use) and between systems (e.g., lack of parental knowledge about the youth's drug use) (Henggeler et al., 2009).

Second, according to social ecology theory, to fully understand someone's behavior, it must be ecologically valid or viewed within its naturally occurring context. As such, a clinician must view the youth's functioning in a variety of real world settings (e.g., home, school, neighborhood) and from firsthand sources (e.g., mom, teacher, neighbor) in order to obtain an "ecologically valid assessment" (Henggeler et al., 2009, p. 16). It is also important that therapeutic interventions be conducted with ecological validity, which is why MST is conducted where problems occur—in homes, schools, and in other community based settings (Henggeler et al., 2009).

Finally, the theory of social ecology argues that human interaction is influenced by the reciprocal influence different systems have on one another. Understanding reciprocity is important for both the assessment and intervention components of MST. More specifically, at the assessment level, it is instructive for therapists to determine why, for example, parents might have given up on their son or daughter or why a youth was expelled from a particular classroom. From here, therapists can use this information to design a treatment strategy for the youth and ultimately implement an effective intervention.

Departing from traditional psychotherapeutic approaches, the MST theory of change is consistent with Brofenbrenner's (1979) theory of social ecology in that it targets youths' observable behavioral problems, rather than focusing on unconscious forces from the past (Van Voorhis & Braswell, 2007). Specifically, the MST theory of change rests on two important assumptions.

First, it is assumed that "adolescent antisocial behavior (i.e., criminal activity, substance abuse, conduct problems) is driven by the interplay of risk factors associated with the multiple systems in which youth are embedded (i.e., family, peer, school, and neighborhood)" (Henggeler

et al., 2009, p. 16). In this way, antisocial behavior can only be changed when interventions consider the broader social context that youths are enmeshed in, as well as the individual risk factors that are driving the particular youth to engage in delinquent behavior (Van Voorhis & Braswell, 2007).

Second, it is assumed that the youth's parents/caregivers are the main conduits for change. As such, MST therapists work with parents/caregivers to develop the resources and skills needed to be more effective with their children (Henggeler et al., 2009). Consequently, parents/caregivers are seen "as full collaborators in treatment" (Henggeler et al., 2009; p. 25) and the family network becomes a central focus for achieving treatment success. Figure 1.1 presents the MST theory of change (Henggeler et al., 2009, p. 31).

MST began to proliferate in the early 1990s due in large part to initiatives that advocated for children's mental health reform. Such movements "helped to generate service system and community environments in which MST [could] thrive" (Henggeler, Schoenwald, Borduin, Rowland, & Cunningham, 1998, p. 17). Specifically, research stemming from the Child and Adolescent Service System Program (1984), for example, emphasized that family and community-based services needed to address a wide array of child and family needs and be individualized to meet those dynamic factors (Henggeler et al., 1998). In this way, MST was a program that quickly gained traction because it not only treated youths' presenting problems, but also emphasized the importance of family and individualized treatment strategies.

In 1992, Henggeler left Memphis State University and moved to the Department of Psychiatry and Behavioral Sciences at the Medical University of South Carolina (MUSC). Here, he founded the Family Services Research Center (FSRC), which was established to develop, validate, and study the transport of MST across agencies and departments (Sheidow, Henggeler,

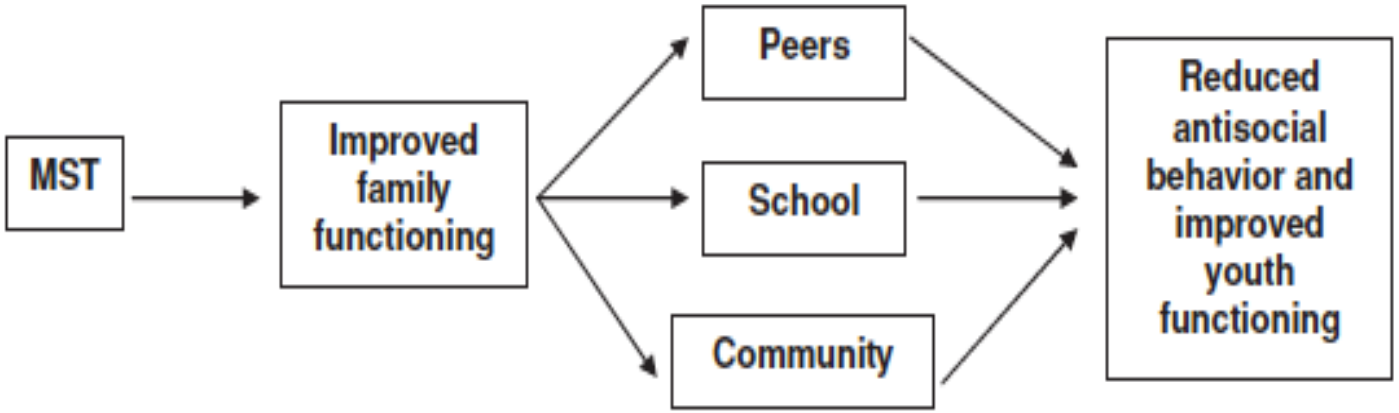
& Schoenwald, 2003). As empirical evidence supporting MST became widely established through clinical and randomized trials, outside agencies began requesting the development of MST programs. "Initially, FSRC faculty, following training and quality assurance protocols used in MST clinical trials, provided program development and ongoing support to new and distal MST sites" (Sheidow et al., 2003, p. 305). Research productivity began to decrease, however, and so did the fidelity of the training and consultation services being provided to off-site agencies (Sheidow et al., 2003). As such, MST Services, Inc. was established in 1996 to support the effective transport and dissemination of MST programs (Sheidow et al., 2003).

As noted, MST currently operates in over 30 states and in 15 countries, treating well over 20,000 youth each year. These data clearly establish that the program—one that started well over three decades ago with a "modest amount of extramural funding" (Sheidow et al., 2003, p. 303)—has grown into an important criminological invention.

THE STRUCTURE OF MST

MST addresses the multiple determinants of serious antisocial behavior in juvenile offenders by providing intensive family- and community-based treatment (Swenson & Duncan, 1998). Beyond the youth and his or her family, MST takes into account the possibility that other contextual systems might be sources that affect a youth's antisocial behavior. Oftentimes, these multiple systems "interact in a reciprocal and dynamic manner" (Bernfeld, Jennings, & Corriveau, 2001, p. 51). The name "multisystemic" comes from the core assumption that criminogenic risks can occur in any of the multiple systems in which a youth is enmeshed. This reality means that the therapy that is

Figure 1.1
The Theory of Multisystemic Therapy (MST)



Source: Henggeler, Schoenwald, Borduin, Rowland, & Cunningham, 2009, p. 31

employed must incorporate "serious flexibility in the design and delivery of interventions" (MST Services, Inc., 2010, p. 9).

The purpose of this section is to describe the structure and features of MST, focusing on several key aspects of the program, including: (1) target population, (2) goals of the intervention, (3) target risk and protective factors, (4) integration of evidence-based practices, (5) intervention location, (6) dosage, (7) core principles, (8) clinical procedures and interventions, (9) MST clinical supervision, (10) structure and responsibilities of the MST treatment team, (11) quality assurance and treatment fidelity.

Target Population

Traditionally, MST programs have targeted persistent, violent, or substance-abusing boys and girls between the ages of 12 and 17 who are at risk of out-of-home placement and their families (Henggeler, 1999). More recently, however, MST programs have expanded in their scope and have been implemented with a variety of juvenile populations in the mental health and medical fields (see, e.g., Borduin et al., 1995; Huey et al., 2004; Borduin, Schaeffer, & Heiblum, 2009). Specifically, MST has been utilized with juveniles suffering from severe emotional disturbances, mental health disorders, and psychiatric illnesses (Henggeler et al., 2003; Henggeler, Pickrel et al., 1999), youths experiencing suicidal, homicidal, and/or psychotic crises (Henggeler et al., 2003; Huey et al., 2004; Huey, Henggeler, Rowland, Halliday-Boykins, Cunningham, & Pickrel, 2005), and abused and neglected youth (Brunk, Henggeler, and Whelan, 1987). Henggeler and his associates argued that the "problem behaviors of youth in these sectors tended to mirror those of youth referred to the juvenile justice system, where MST had shown positive results" (Burns et al., 2000, p. 285). Consequently, MST has become a program that has wide appeal and has been implemented with a variety of adolescent populations.

Goals of the Intervention

As briefly discussed in the previous section, one of the primary assumptions of the MST theory of change is that families are the central focus of the intervention and caregivers are seen as "crucial change agents for their children" (Henggeler et al., 2009, p. 25). In this way, MST clinicians aim to achieve two primary goals when working with youth and their families. First, clinicians aim to provide the youth's caregivers with the skills and resources necessary to independently address antisocial behavior and to identify and teach their child prosocial alternatives (MST Services, Inc., 2010). The second goal is to teach youths the prosocial skills necessary to deal with difficulties that might arise within their family, with their friends, in their school, and/or in their neighborhood (Swenson & Duncan, 1998).

Target Risk and Protective Factors

To achieve the goals described above, MST clinicians seek to reduce the risk factors that are "driving" youth's antisocial behavior (Mitchell-Herzfeld et al., 2008), while simultaneously identifying areas of strength (i.e., protective factors) that can be used as leverage for achieving long-term behavioral change. "In general, these risk and protective factors are relatively constant, whether the examined antisocial behavior is conduct disorder, delinquency, or substance abuse" (MST Services, Inc., 2010, p. 4). Table 1.1 presents a list of identified risk and protective factors commonly seen across the various ecological contexts in which youths and their families are enmeshed (Henggeler, Mihalic, Rone, Thomas, & Timmons-Mitchell 1998). It is instructive that the factors that make up Table 1.1 are criminogenic in nature and are in line with the research on "what works" to change offender behavior (see e.g., Andrews & Bonta, 2010; Andrews, Bonta, & Hoge, 1990; Andrews, Zinger, Hoge, Bonta, Gendreau, & Cullen, 1990).

During initial meetings with the family, the MST therapist works to find the "fit" between the problem behavior and the different social contexts in which the youth is embedded (Henggeler et al., 2009; Swenson & Duncan, 1998). "Consequently, the identification of the key variables in a particular case is the major task of assessment in MST" (MST Services, Inc., 2010, p. 6). This assessment process is important because not all risk and protective factors will pertain to all youth and families; thus, treatment plans must be individualized for each family (Henggeler et al., 2009).

Regarding families, "the MST therapist works to enhance the caregivers' parenting skills (e.g., monitoring, supervision, affective relations) and then leverages these improvements in family functioning to facilitate key changes in other extrafamilial networks" (Henggeler, 2011, p. 355). For example, caregivers might learn how to identify antisocial peers that the adolescent is spending time with and teach the youth how to disengage from these negative relationships. Subsequent interventions then might teach caregivers how to motivate the child towards more positive peer networks. Finally, MST clinicians might encourage caregivers to become involved in other aspects of the youth's social context, such as collaborating with teachers and school administrators and/or other community professionals (e.g., probation officers) (Henggeler, 2011).

At the peer level, delinquent peers can contribute to behavior problems in that they "provide youths with an increased opportunity to learn a variety of criminal behaviors, particularly covert antisocial behaviors" (Andrews & Bonta, 2010, p. 230). Antisocial peers are especially risky for adolescents because they are in the midst of developing a system of morals and values that can be easily influenced by peer interactions (Henggeler et al., 2009). In this way, MST aims to diminish youths' associations with deviant peers by determining what factors

about the youth-peer dynamic are leading them to trouble (e.g., rejection from prosocial peer networks) (Henggeler et al., 2009).

School is an additional factor that can lead the youth to trouble. Specifically, low academic achievement, dropping out of school, and low commitment to education are all major predictors of adolescent problems (Sheidow et al., 2003). Youth who struggle in this realm are given assistance with academics and parents/caregivers are supported in monitoring their child's school activities. Further, teachers are involved to aid in the change process (Andrews & Bonta, 2010).

Finally, neighborhood and community factors can also influence the way an adolescent behaves. For example, youth who are exposed to violence, live in crime-ridden neighborhoods, and/or lack community supports, are more likely to engage in crime themselves (Sheidow et al., 2003). From this perspective, MST clinicians attempt to enlist as many positive community supports in the intervention as possible. The approach truly "leaves no stone unturned in identifying the immediate social and community supports that can increase the rewards for prosocial behavior and interfere with the social forces that support antisocial activity" (Andrews & Bonta, 2010, p. 258).

Integration of Evidence-Based Practices

MST clinicians integrate a series of empirically-based treatment approaches "into a broad-based ecological framework" in order to target the risk factors identified during the family's assessment process (Henggeler, 1999, p. 3). More specifically, treatment techniques are integrated from therapies that have at least some empirical support (Henggeler, 1999). These include, for example, strategic family therapy, structural family therapy, behavioral parent training, and cognitive behavioral therapies. Further, MST clinicians identify and address any

Table 1.1
Risk and Protective Factors

	Risk Factors	Protective Factors
Individual	• Low verbal skills • Favorable attitudes toward antisocial behavior • Psychiatric symptomatology • Cognitive bias to attribute hostile intentions to others	• Intelligence • Being firstborn • Easy temperament • Conventional attitudes • Problem-solving skills
Family	• Lack of monitoring • Ineffective discipline • Harsh and inconsistent discipline • Low warmth • High conflict • Parental difficulties (e.g., drug abuse, psychiatric conditions, criminality)	• Attachment to parents • Supportive family environment • Marital harmony
Peer	• Association with deviant peers • Poor relationship skills • Low association with prosocial peers	• Bonding with prosocial peers
School	• Low achievement • Dropout • Low commitment to education • Aspects of the schools (e.g., weak structure and chaotic environment)	• Commitment to schooling
Neighborhood and Community	• High mobility • Low community support (e.g., neighbors, church) • High disorganization • Criminal subculture	• Ongoing involvement in church activities • Strong indigenous support network

Source: Henggeler, Mihalic, Rone, Thomas, & Timmons-Mitchell, 1998

biological factors that may be contributing to the youth's behavior and integrate psychopharmacological treatment, as well as psychosocial treatment when appropriate (Henggeler, 1999).

Intervention Location

MST clinicians are committed to identifying and addressing any barriers to services that the family might have (MST Services, Inc., 2010). In this way, the MST approach follows a community-based model of service delivery in that interventions occur where problems for the adolescent actually arise. This type of service delivery allows families to receive treatment almost anywhere, which increases the likelihood that they will actually stay and participate in treatment (Burns et al., 2000). Providing the intervention on the family's own "turf" eliminates many barriers to service access, such as lack of transportation or restricted hours of operation (Henggeler et al., 2009). Because MST allows youths and their families to be treated in a familiar place, the stigma of participating in an office-based treatment program is reduced and the client and his or her family members feel more at ease and comfortable discussing the problems and issues they are going through (Sheidow et al., 2003).

Dosage

Each clinician within the MST team carries a caseload between four and six families at any given time; allowing the clinician to provide intensive services to each family in their respective environments (Bernfeld et al., 2001). On average, families receive between 40 and 60 hours of direct clinical contact over the course of a three-month to five-month time span (Burns et al., 2000; Henggeler et al., 2009). When clinicians meet with the family, sessions typically last anywhere between 20 and 75 minutes and are conducted during a time that is convenient for

the family. Finally, sessions can be held as often as daily when the treatment initially begins and, then, generally take place less often over the course of treatment (Burns et al., 2000).

Core Principles

As Henggeler explains, "MST is operationalized through adherence to nine core treatment principles" (MST Services, Inc., 2010, p. 9) that "provide clear guidelines to measure treatment success and the parameters to which a therapist must adhere" (Henggeler, 1999, p. 3). Although these principles are important, "they are not so rigid as to limit the flexibility that is necessary to deliver an ecologically valid treatment on an individual basis" (Burns et al., 2000, p. 286). The nine treatment principles are as follows (Henggeler et al., 2009, pp. 25-26):

1. *Finding the fit*: The primary purpose of assessment is to understand the "fit" between the identified problems and their broader systemic context and how identified problems "make sense" in the context of the youth's social ecology.

2. *Positive and strength focused*: Therapeutic contacts emphasize the positive and use systemic strengths as levers for positive change. Focusing on family strengths has numerous advantages, such as decreasing negative affect, building feelings of hope, identifying protective factors, decreasing frustration by emphasizing problem-solving, and enhancing caregivers' confidence.

3. *Increasing responsibility:* Interventions are designed to promote responsible behavior and decrease irresponsible behavior among family members. The emphasis on enhancing responsible behavior is contrasted with the usual pathology focus of mental health providers and kindles hope for change.

4. *Present focused, action oriented, and well defined:* Interventions are present-focused and action-oriented, targeting specific and well-defined problems. Such interventions enable treatment participants to track the progress of treatment and provide clear criteria to measure success. Family members are expected to work actively towards goals by focusing on present-oriented solutions (versus gaining insight or focusing on the past). Clear goals also delineate criteria for treatment termination.

5. *Targeting sequences:* Interventions target sequences of behavior within and between multiple systems that maintain identified problems. Treatment is aimed at changing family interactions in ways that promote responsible behavior, and broaden family links with indigenous prosocial support systems.

6. *Developmentally appropriate:* Interventions are developmentally appropriate and fit the developmental needs of the youth. A developmental emphasis stresses building youth competencies in peer relations and acquiring academic and vocational skills that will promote a successful transition to adulthood.

7. *Continuous effort:* Interventions are designed to require daily or weekly effort by family members, presenting youth and family frequent opportunities to demonstrate their commitment. Advantages of intensive and multifaceted efforts to change include more rapid problem resolution, earlier identification of treatment non-adherence, continuous evaluation of outcomes, more frequent corrective interventions, more opportunities for family members to experience success, and family empowerment as members orchestrate their own changes.

8. *Evaluation and Accountability:* Intervention effectiveness is evaluated continuously from multiple perspectives, with providers assuming accountability for overcoming barriers to successful outcomes. MST does not label families as "resistant, not ready for change or unmotivated." This approach avoids blaming the family and places the responsibility for positive treatment outcomes on the MST team.

9. *Generalization:* Interventions are designed to promote treatment generalization and long-term maintenance of therapeutic change by empowering caregivers to address family members' needs across multiple systemic contexts. The caregiver is viewed as the key to long-term success. Family members make most of the changes, with MST therapists acting as consultants, advisors, and advocates.

Research has shown that MST programs that adhere to these nine core treatment principle have more favorable outcomes, compared to programs that do not follow such guidelines (see e.g., Henggeler, Melton, Brondino, Scherer, & Hanley, 1997; Henggeler, Pickrel, et al., 1999; Huey, Henggeler, Brondino, & Pickrel, 2000; Schoenwald, Henggeler, Brondino, & Rowland, 2000). In this way, "considerable training, supervisory, and consultative resources are devoted to maximizing therapist adherence to the [above mentioned] MST treatment principles" (MST Services, Inc., 2010, pp. 9-10).

Clinical Procedures and Interventions

By using a standardized diagnostic and decision-making process, it is possible for MST clinicians to adhere to the core principles described above. More specifically, the MST

analytical process, or “do-loop,” (Henggeler et al., 2009, p. 29), guides MST clinicians in that it “structures the treatment plan, its implementation, and the evaluation of its effectiveness” (Henggeler & Schaeffer, 2010, p. 153). As the name “do-loop” suggests, the MST analytical process is a “sequential and recursive process that MST teams use to conceptualize and intervene with families” (Henggeler at al., 2009, p. 29). Figure 1.2 displays this decision-making process.

The do-loop is an important starting point in the MST intervention because it serves as a guide for completing other activities and tasks with the family. Therapists follow the following five-step process (Henggeler et al., 2009, pp. 31-32):

1. The therapist obtains the family’s “buy-in” and works with them and other community supports to establish a concrete list of overarching goals.

2. The therapist examines within and between all of the systems in which the youth and his or her family are enmeshed to understand how the presenting behaviors have manifested and developed.

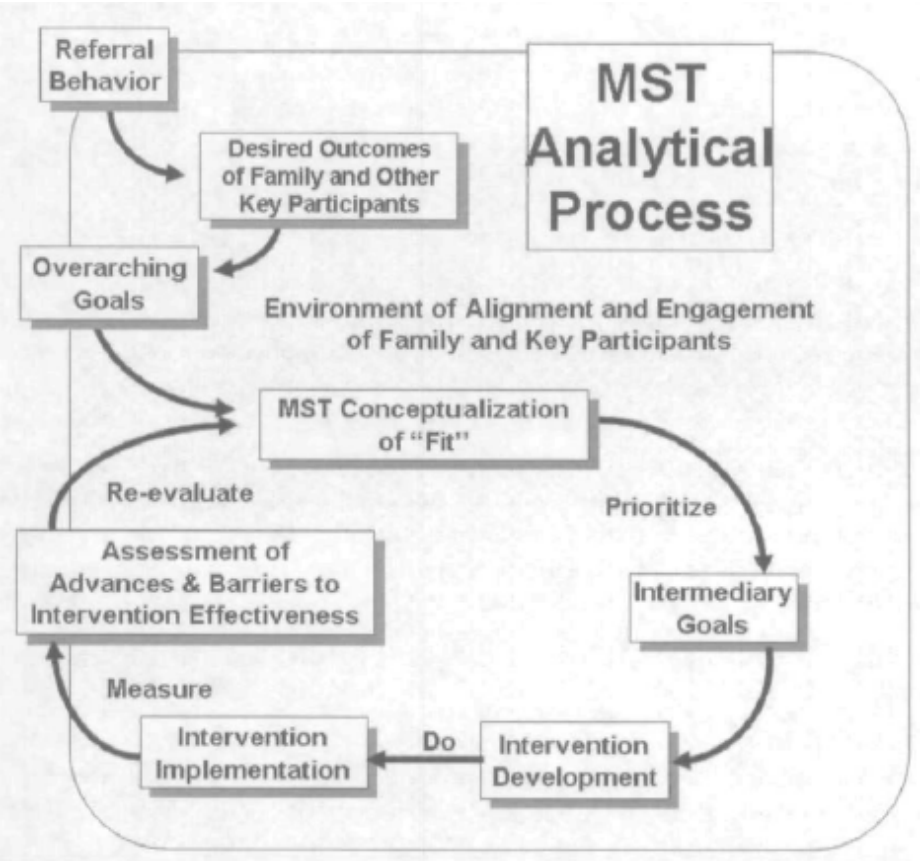
3. The team and family members prioritize hypothesized drivers of the identified problems and develop interventions to address the drivers.

4. The interventions are implemented, and any barriers to effective implementation are identified.

5. Following the do-loop back up toward the top, therapists assess the outcomes of their interventions from multiple perspectives to determine if they are having the intended effects. If not, the information gained during the process is fed back into the loop, and therapists work to develop new hypotheses and modified interventions based on these revised hypotheses.

According to Henggeler and Schaeffer, (2010, p. 154), “The recursive nature of the analytic process reinforces two important features of the MST model.” First, MST clinicians are trained to never give up and to do "whatever it takes” to accomplish the goals and objectives the family and treatment team set forth during the first step in the do-loop process (Henggeler 2011; Henggeler et al., 2009; Henggeler & Schaeffer, 2010). Second, it is the responsibility of the MST clinician to engage the family and other key participants in the treatment process (Swenson

Figure 1.2
Multisystemic Therapy Recursive Clinical Decision-Making Process



Source: Henggeler & Schaeffer, 2010, p. 153

& Duncan, 1998). In this way, youths and their family members help the MST clinician in designing the treatment plan, "which ensures that it will be family-driven rather than therapist-driven" (Swenson & Duncan, 1998, p. 1). If initial strategies are unsuccessful or interventions fail, however, it is the team's failure, rather than the family's failure (Henggeler & Schaeffer, 2010).

As briefly mentioned, the information gathered during the analytic process is used as a guide for completing other activities during treatment. Specifically, MST clinicians use four different "conceptual aids" (background information form, strengths and needs assessment, case summary for supervision and consultation, and fit circles) during the intervention process that "help therapists maintain a clear focus on treatment goals and progress while helping supervisors and consultants maximize the usefulness of their feedback" (Henggeler et al., 2009, p. 31). These aids are generally completed at various points in time and may include participation from the family and/or others participating in the treatment process (Henggeler et al., 2009).

First, the background information form is completed in the initial meeting with the family and updated periodically, when needed. The form begins with a genogram to help the MST clinician and treatment team understand the family structure, dynamic, and relationships across and between family members (Henggeler & Schaeffer, 1998). The form also helps the treatment team identify potential individuals for involvement in the treatment process. Perhaps most importantly, the form identifies reasons why the youth and his or her family are involved with MST services and helps the treatment team understand presenting problem behaviors. "These behaviors are described using brief, specific behavioral terminology, and care is taken to provide information concerning the frequency, intensity, and duration of the behaviors as well as the systems that have been impacted" (Henggeler et al., 2009, p. 31). Finally, at this stage in the

treatment process, the MST clinician meets with the youth's family members (e.g., caregivers, siblings, and other important extended family members) to get an understanding of how they see the situation and to determine "the outcomes they want the youth to achieve during treatment" (Henggeler et al., 2009, p. 31).

Along with the background information form, therapists and the family fill out a strength and needs worksheet at intake and then periodically update as needed (e.g., when treatment is completed or when a traumatic event occurs) (see Appendix A for an example of the background information form and strengths and needs assessment). At this stage, risk and protective factors are identified for each of the identified contextual systems the youth and his or her family are engaged in (family, school, peers, individual, and neighborhood/community) (Henggeler & Schoenwald, 1998).

Third, the case summary for supervision and consultation form is completed or updated each week. The form consists of six sections: (1) overarching/primary MST goals, (2) previous intermediary goals, (3) barriers to intermediary goals, (4) advances in treatment, (5) assessment of "fit" between identified problems and their broader systemic context, and (6) new intermediary goals for next week and is used to guide MST supervision and consultation meetings. (see Appendix B for a brief description and example for each section covered in the MST case summary and consultation form) (Henggeler et al., 2009).

Finally, following from the first principle of the nine MST core treatment principles, fit circles are used to "understand the fit between the identified problems and their broader systemic context" (Henggeler et al., 2009, p. 35) (See Appendix C for an example of a fit circle). These relationships are important, as treatment teams use them to develop hypotheses concerning possible causes of the youth's presenting problems. The MST clinician completes this exercise

with family members at intake (and throughout treatment when needed) and subsequently uses the completed fit circle to make decisions about which intervention strategies and treatment processes should be put in place (Henggeler et al., 2009; Henggeler & Schoenwald, 1998).

In short, the MST treatment process entails several steps that include ongoing assessment and implementation of treatment and intervention strategies that seek to reduce youths' presenting problems. Clinicians are encouraged to test hypotheses when they have "hunches, beliefs, or theories about (a) the causes and correlates of particular problems in a family, (b) the reasons that improvements have occurred, and (c) barriers to change" (Schoenwald, Brown, & Henggeler, 2000, p. 115). To accomplish weekly goals and objectives, clinicians rely on both the MST do-loop and several other conceptual aids to track interactions and developments with the family. In addition, these resources are used to keep treatment staff, supervisors, and MST consultants updated on each family's progress and setbacks (Schoenwald et al., 2000).

MST Clinical Supervision

According to Henggeler et al. (2009, p. 26), *"*As the nine principles provide a foundation for the design of interventions used in the model, so the MST clinical process provides the underlying structure and framework on which therapists build their intervention." In this way, the framework for any MST intervention is built upon three important components. First, programs must consistently use the conceptual aids discussed in the previous section. Second, a strong treatment team must be in place. And third, the program must establish a process for ongoing quality assurance and improvement for the overall program (Henggeler et al., 2009). Agencies that have these processes in place are more likely to achieve long-term success as an organization and in implementation of MST (Henggeler, et al., 1998).

As noted, the first component that lays the framework for any MST intervention—the use of conceptual aids—is important because the aids "serve as a system of communication…to help therapists convey important clinical information to supervisors, which, in turn, the MST team passes along to the team consultant." (Henggeler et al., 2009, p. 31). With so many individuals involved in the youth's rehabilitative process, it is clear why the second component—establishment of a strong treatment team—is also important for achieving positive outcomes.

A strong treatment team starts with clinical supervision where emphasis is placed on understanding human behavior and the multidetermined ways in which these behaviors progress (Schoenwald, Henggeler, & Brown, 2000). In clinical sessions, supervisors are responsible for ensuring that therapists continue to understand the theoretical framework from which MST was developed, as well as ensure that they have developed and maintained the behavioral skills necessary to implement the program effectively. In this context, supervision typically focus on developing and improving the following skills (Henggeler et al., 1998, p. 46):

1. Develop and refine a multisystemic conceptualization of the causes of identified problems presented by each client family.

2. Design and effectively implement intervention strategies that embody the nine treatment principles.

3. Identify barriers to the successful engagement of key participants (family members, school personnel, sources of parental social support) and implement strategies to overcome these barriers.

4. Logically and clearly connect intermediary goals to ultimate goals and intervention strategies to both intermediary and ultimate goals.

5. Identify barriers to the successful implementation of interventions and implement strategies to overcome them.

With this skill development, supervisors are trained to continuously remind clinical staff about the importance of critical thinking and the use of hypothesis testing whenever problems arise (Schoenwald et al., 2000).

Finally, it is instructive to note that within the MST framework, clinical supervision operates under the following five assumptions (Henggeler & Schoenwald, 1998, pp. 1-2):

1. The purpose of clinical supervision is to enable clinicians to adhere to the nine principles of MST in all aspects of treatment-engagement of families, case conceptualization, intervention design and implementation, and evaluation of outcomes.

2. Each clinician implementing MST is a hard-working, competent professional who brings unique personal strengths and professional experiences to the treatment process.

3. Ongoing clinical supervision is necessary to monitor adherence to MST and to achieve positive, sustainable outcomes with youth presenting serious clinical problems and their families.

4. The process of clinical supervision should mirror the process of MST. That is, supervision is present-focused, action-oriented, and targets specific problems that the clinician appears to be having in (a) engaging families in the treatment process, (b) conceptualizing the "fit" of referral problems with the family's ecological context, (c) identifying and using strengths as levers for change, (d) designing interventions, (e) implementing interventions adequately, and (f) overcoming barriers to intervention implementation or success. Supervision also should enable clinicians to sustain MST-like conceptualization and intervention skills across families (generalization).

5. Clinicians, supervisors, and the provider organization that houses the MST program are accountable for outcomes.

Structure and Responsibilities of the MST Treatment Team

MST is typically provided by treatment teams consisting of two to four therapists and one supervisor. These teams are traditionally part of a private company who are contracted by the juvenile court, welfare services, and/or mental health agencies to provide MST services (Sheidow, Schoenwald, Wagner, Allred, & Burns, 2006). The vast majority of both supervisors and clinical staff hold master's-level degrees in a helping profession, such as psychology, social

work, or family counseling (Henggeler et al., 2009). In addition, an off-site MST consultant works with each company and is responsible for initial implementation of MST, helping therapists and supervisors achieve clinical competence in the program model, and ensuring treatment fidelity is being achieved (Henggeler et al., 2009; Schoenwald et al., 2000).

The MST therapist's primary focus is on the interactions within the family and between family members and school personnel, peers, neighbors, and other individuals who might be seen as influencing the youth's behavior (Henggeler et al., 1998). As would be expected, the complexity of MST and the difficulties that arise when helping families with at-risk youths can be a challenge, even for the most experienced MST clinicians. In this way, the group supervision format described above "helps team members to become familiar with developments in all cases and with supervisory feedback regarding those developments. This familiarity, in turn, enables clinicians to reinforce the multisystemic nature of therapeutic efforts throughout the week, generalize as appropriate from supervisory discussion of colleagues' cases, and sustain continuity of care while covering for one another during vacation, weekend, and personal time" (Henggeler et al., 1998, p. 48).

According to Henggeler and his colleagues (2009, p. 51), "The MST supervisor and expert consultant are responsible for developing the clinical competence of the team and identifying and addressing any practitioner-, organizational-, and system-level barriers to the operational and clinical success of the MST program." Specifically, supervisors act as on-site expert consultants for their team of clinicians and play a crucial role in the overall functioning and development of each staff member (Henggeler et al., 2009).

The responsibilities of the MST supervisor can be seen as falling into five general categories (Henggeler et al., 2009). First, supervisors are responsible for consulting the research

related to juvenile justice and corrections (e.g., evidence-based practices, effective treatment and intervention strategies). In addition, they must be able to demonstrate a clear understanding of the MST conceptual and behavioral framework. Second, supervisors must know enough about each therapist's current caseload because they must assist clinicians in all aspects of treatment (e.g., assessment, implementation of interventions, overcoming barriers to implementation). Third, supervisors are responsible for conducting weekly supervision meetings with their team and ensuring that all MST procedures and protocols are followed. Fourth, supervisors are responsible for their own personal development and growth so as to influence the therapists on their team. Personal development can be achieved by working alongside the MST expert consultant and engaging in ongoing training. In a similar vein, the fifth area of responsibility involves monitoring the team's needs and identifying areas for improvement and resources that might be needed (Henggeler et al., 2009).

Finally, Henggeler and his colleagues (2009, p. 300) note that "whereas the provider organization's MST supervisor and therapists are responsible for day-to-day decision making regarding each case, the MST consultant is responsible for contributing to the rapid development of the clinicians' ability to bring MST-like thinking and interventions to the cases and to sustain that ability over time." Similar to the captain of a football team, the consultant is the team's cheerleader when positive outcomes are achieved and their troubleshooter when barriers to treatment and supervision arise.

Quality Assurance and Treatment Fidelity

As the demand for MST increased at the community level, Henggeler and his colleagues began to develop strategies to meet these demands and established policies and procedures to deal with the inevitable challenges that came with the large-scale transport of a traditionally

clinically based program. The goal of their transportability research was to achieve similar results (i.e., reductions in recidivism and other antisocial behaviors) that they had seen in previous clinical trials of MST. "As in CQI [continuous quality improvement] approaches attempted in health care, the idea was to design an implementation system that engaged all individuals involved in the delivery effort by establishing a feedback loop that integrated data-based and qualitative feedback about MST implementation at the level of the client family, therapist, supervisor, expert consultant, and organization with the MST program" (Schoenwald, 2008, p. 74). Specifically, MST uses a comprehensive quality assurance and improvement (QA/QI) system that is comprised of three broad interrelated components: (1) training and monitoring; (2) organizational support; and (3) implementation measurement and reporting (Henggeler, 2011). Figure 1.3 depicts how each of these components are integrated into the aforementioned feedback loop (Henggeler, Schoenwald, Rowland, & Cunningham, 2002).

The process of hypothesis testing is also used during the QA/QI stages to identify what is and is not working in the implementation of the program. Failure to implement the program with fidelity can subsequently affect youth outcomes. For example, if therapists target non-criminogenic factors (i.e., lack fidelity to the MST model), youths are less likely to change their delinquent behavior, and subsequent reductions in recidivism will likely not be observed (i.e., poor youth outcomes).

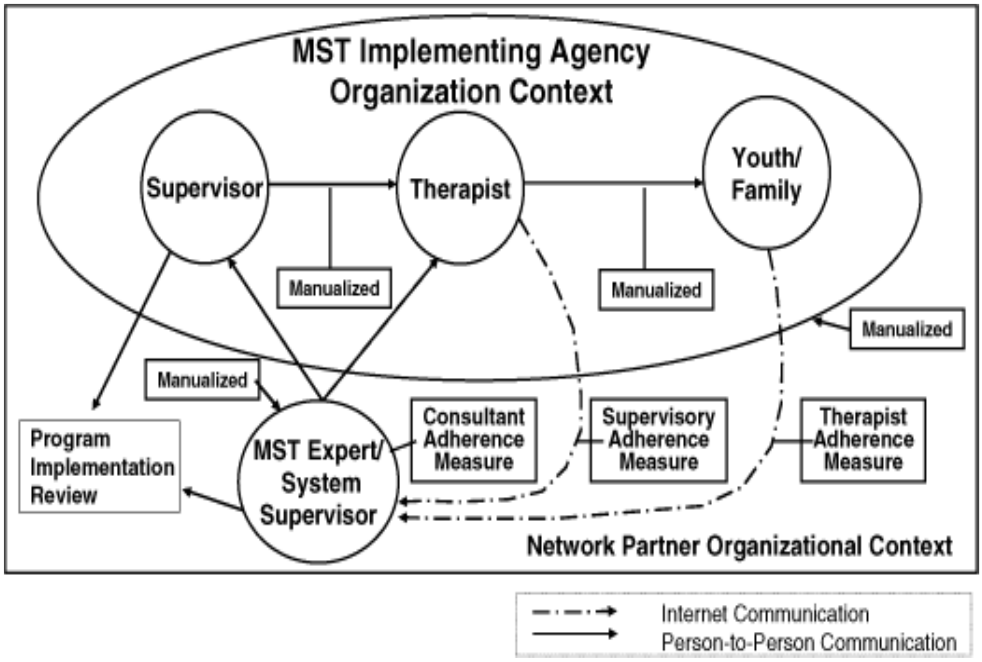
Training and Monitoring. The intensive nature of the MST QA/QI process differentiates it from most other mental health approaches likely because of the highly structured training and monitoring components that are required of all provider organizations (MST Services, Inc., 2010). Specifically, the QA/QI process provides mechanisms at each level (e.g., therapist, supervisor, consultant, and program) for training and support on the elements of the

MST treatment model (MST Institute, 2014). MST Services, Inc. provide a "core training package" (Henggeler et al., 2002, p. 231) to agencies interested in formally implementing the program. Henggeler and his colleagues (2002, p. 229) advise that "in light of the great diversity of children and families served by MST programs, as well as of the practitioners and staff working in MST programs, structures and processes are needed to focus stakeholders on the ultimate goals of MST programs…" Training and monitoring are accomplished in the following four ways: (1) five-day orientation training, (2) 1.5-day booster training sessions on-site, (3) weekly clinical supervision on site, and (4) weekly expert consultation.

First, clinicians, supervisors, and other staff in the provider organization who might be involved in the youth's treatment plan, are trained in the MST model during a five-day orientation training. During this time, the MST program is officially "launched" at the provider organization and brings together stakeholders, key management and leadership personnel who likely played a significant role in bringing the treatment approach to the site (Henggeler et al., 2002). The rest of the five-day training is focused on the collaboration of the clinicians, supervisor, and expert consultant. At this time, the consultant is responsible for teaching the rest of the team about the theory behind MST and the rationale for its approach. The consultant is also responsible for engaging the group in various modeling and role-playing activities. These strategies and techniques allow clinicians and supervisors to conceptualize cases and practice interventions from "an MST perspective" (Henggeler et al., 2009, p. 298).

A second piece of the MST training and monitoring protocol involves quarterly booster sessions. These sessions consist of 1.5-days of on-site training and are "designed to enhance the knowledge and skills of the team so they can more effectively address clinical challenges they are facing (e.g. marital interventions, treatment of caregiver depression)" (Henggeler et al., 2009,

Figure 1.3
The MST Quality Assurance and Improvement System



Source: Henggeler, Schoenwald, Borduin, Rowland, & Cunningham, 2009, p. 296

p. 298). Modeling and role-playing are also used during these trainings to help therapists and supervisors problem-solve particularly challenging cases and come to a consensus on how the situation should be handled. When the session is over, therapists evaluate the session and the feedback is used by the expert consultant to enhance future trainings. Finally, supervisors and consultants are responsible for observing therapist implementation of the skills between booster sessions to determine whether the new plan is working for the clinician and the family (Henggeler et al., 2009).

Part of the MST training process also involves weekly clinical supervision meetings. The purpose of these meetings "is to help therapists learn and use the clinical skills—conceptual and behavioral—needed to effectively implement MST with each and every youth and family served" (Henggeler et al., 2009, p. 298). In this way, MST Services, Inc. provides supervisors with a plethora of resources (e.g., a supervision manual and continuous support from the on-site expert consultant) to ensure they are successful in meeting these goals and objectives with their team.

Finally, weekly expert consultation meetings occur with the MST supervisor and clinicians which aim to identify and address clinical-, team-, organizational-, and systemic-level barriers to achieving positive outcomes for adolescents and families (Henggeler et al., 2009, p. 300). Due to their knowledge and expertise in the field, consultants have the ability to augment and change existing clinical practices to achieve successful outcomes with their team. In this way, it is imperative that the consultants not only develop a strong relationship with the supervisor, but also with the clinicians on their team to ensure their recommendations and adaptations to the procedures are effective (Henggeler et al., 2002).

It is also instructive to note that MST Services, Inc. has created a series of standardized manuals that are provided to all staff involved in the MST continuum of care prior to their five-day orientation. Standardized manuals include, for example, a published treatment manual for clinicians (Henggeler et al., 1998), a supervision manual for on-site MST supervisors (Henggeler & Schoenwald, 1998), a consultation manual for off-site expert consultants (Schoenwald, 1998), and an organizational manual for MST program administrators (Strother, Swanson, & Schoenwald, 1998).

Organizational Support. As Henggeler and his colleagues explain (2009, p. 303), "research reveals that, across numerous industries, individuals and organizations often decide to adopt a new program and frequently fail to implement it successfully." Research also reveals that the organization itself is instrumental to the success and failure of new programs (Andrews & Bonta, 2010). As such, it is important that staff members are not only trained initially on the treatment approach, but continue to be involved in training and booster sessions to keep with the fidelity of the program. Several organizational-level structures and procedures have been adopted to address common barriers to successful program implementation.

First, as briefly discussed in the previous sub-section, the MST organizational manual (Strother et al., 1998) describes the program practices and policies of the MST program and is intended to help key leaders and administers from the provider organization. The policies and practices described in this manual are based on procedures used in randomized trials of MST, as well as factors that emerged during the transport and implementation processes in other community-based settings (Henggeler et al., 2009). The manual provides an overview of the theory and practices behind MST, as well as highlights areas where programs have struggled with implementation and fidelity to the model (e.g., evaluation, program financing, recruitment

and retention, youth referral and discharge criteria). Finally, programmatic features (e.g., on-call system, referral agency communication and relationships), as well as practical needs of the staff (e.g., agency vehicles, insurance, cell phones) are also described in the manual (Henggeler et al., 2002; Henggeler et al., 2009).

Also important to the organizational support of MST is initial and ongoing program development. Initial program development begins when a funding agency (e.g., juvenile court, welfare service provider, mental health provider) contacts MST Services, Inc. and expresses interest in starting a program. During this early conversation, staff from MST Services, Inc. typically ask the interested agency a series of questions to better understand their mission, policies, and practices, as well as identify the specific population the provider organization intends to serve (Henggeler et al., 2002). If it is thought that MST could potentially be a viable program in the agency's community, a needs assessment is conducted. The goal of the needs assessment is four-fold: (1) to identify the community's need for MST, as well as establish the target population, geographic area, and number of potential referrals in the area; (2) to develop a financial plan and assess the adequacy of funding to sustain the program; (3) to identify service provider organizations and referral agencies that would be committed to implementing MST with fidelity and would follow quality assurance and quality improvement protocols; and (4) to identify stakeholders that would promote the program and also be committed to implementing it with fidelity (Henggeler et al., 2009, p. 304).

If MST continues to appear to be a viable program for the community and funding agency, several critical issue meetings are conducted with staff from MST Services, Inc. the funding organization, and key program agencies to specify how components of the program will operate (e.g., inclusion/exclusion criteria, discharge criteria, outcomes to be collected). During

these sessions, an individualized "goals and guidelines" document is created for the program that lays out how these components will work in the specific community (Henggeler et al., 2009).

Staff recruitment and orientation training, as well as a site readiness review meeting are held once all of the funding sources and service provider organizations are identified. The site readiness meeting incorporates those individuals who will be responsible for the day-to-day operations of the MST program, as well as referral service organizations, and other entities that are important to the successful implementation of the program. The purpose of this meeting is to review the implementation plan and to discuss any adaptations or changes that might need to be made before the program is launched (Henggeler et al., 2009).

Finally, program implementation support is provided through MST Services, Inc. on an ongoing basis to also foster organizational support for the program. For example, during the QA/QI process (which occurs the week before the five-day initial orientation training), a schedule for booster trainings, on-site supervision, and expert consultation is established. Procedures for measuring fidelity are also established during this time (Henggeler et al., 2009).

Implementation Measurement and Reporting. To ensure that each member of the MST treatment team keeps with the fidelity of the model, a system of checks and balances has been established that includes measures of therapist adherence as reported by youths' parents/caregivers, supervisor adherence as reported by therapists, and consultant adherence as reported by therapists and supervisors (Schoenwald, 2008). Recently, the MST Institute has created a web-based system (www.mstinstitute.org) "to support the reporting, scoring, and interpretation of therapist adherence, supervisor adherence, [consultant adherence], and youth outcomes (Henggeler et al., 2009, p. 307). These are described below.

The Therapist Adherence Measure-Revised (TAM-R) includes 28 items based on a Likert scale (Henggeler & Borduin, 1992; Henggeler, Borduin, Schoenwald, Huey, & Chapman, 2006) (see Appendix D). Specifically, the TAM-R is completed by youths' caregivers and assesses therapist adherence to the nine principles of MST (Schoenwald, 2008). The assessment is first administered during the second week of MST treatment and is further administered once every four weeks thereafter (Schoenwald, 2008).

Similarly, the Supervisor Adherence Measure (SAM) consists of 43 items in a Likert format and is comprised of four subscales: (1) structure and process of supervision; (2) supervisor promotes adherence to the MST treatment principles; (3) supervisor promotes use of the MST analytic process; and (4) supervisor promotes clinician development of the competencies needed to implement MST (Schoenwald, 1998; Schoenwald, Henggeler, & Edwards, 1998; Schoenwald, Chapman, & Sheidow 2006)) (See Appendix E). Therapists initially complete the SAM one month after the MST program begins and then every two months thereafter (Schoenwald, 2008).

Finally, the Consultant Adherence Measure (CAM) consists of 23 Likert scale items and is completed by MST therapists and supervisors (Schoenwald, 2001) (see Appendix F). The CAM is first completed after two months of MST treatment and completed every two months thereafter (Schoenwald, 2008). When conducting the assessment, therapists and supervisors rate the consultant on three scales. First, the perceived consultant competence (PCC) subscale examines the therapists' perceptions that the consultant is knowledgeable, skilled in, and able to teach MST. Second, the MST procedures (MSTP) subscale examines the consultant's use of MST-specific assessments and intervention strategies, as well as their use of the conceptual aids when attempting to help therapists problem solve and troubleshoot. And finally, the alliance (A)

subscale examines the therapists' perceptions that the consultant is supportive of their efforts and attentive to their needs (Schoenwald, 2008).

As Schoenwald (2008, p. 82) explains, "Empirical evaluation of linkages among the components of the MST quality assurance and improvement system, their impact on youth outcomes, and other factors that could affect the implementation of MST in usual care settings…" have been conducted. Briefly, these studies confirm that the MST QA/QI process is not only a viable program, but also a valid program in that factors derived from respondent ratings on the TAM-R, SAM, and CAM assessments are associated with reductions in youth delinquency and recidivism (e.g., arrests, incarceration, self-reported criminal activity, and out-of-home placement).

MST AS AN EVIDENCE-BASED INTERVENTION

Since the initial development of MST, Henggeler and his colleagues have been devoted not only to providing communities with affordable and effective interventions for high-risk youths and their families, but also to examining the effectiveness of MST in rigorous research studies (Henggeler, 1997; Henggeler et al. 1998). In this way, the purpose of this section is to provide an overview of outcomes achieved in MST clinical trials and other evaluations (e.g., efficacy, effectiveness, and transportability studies) (Henggeler et al., 2009). More specifically, this section provides a broad overview of both early MST research studies, as well as later evaluations of the program conducted by Henggeler and his team. Finally, the section will end with a brief review of the findings from studies conducted independently of Henggeler and the developers of MST.

Henggeler's Early Evaluations

Henggeler and his team conducted several trials of MST that helped to establish initial support for the program. Specifically, four clinical trials conducted between the late 1970s and 1990s set the stage for later effectiveness trials of the program, as well as research on the transportability and dissemination of the MST approach. It should be noted that during each of these studies, close clinical supervision and oversight was provided to graduate students who served as therapists. Also important, the studies were conducted in "research friendly" environments (i.e., academic university departments) where factors that typically pose challenges for researchers were not of issue (e.g., finances, staff turnover) (Henggeler et al., 2009).

The first two evaluations of MST were conducted when Henggeler was in his initial academic position at Memphis State University (now the University of Memphis) in the Department of Psychology in the late 1970s (Henggeler et al., 1998). The first study was conducted between 1978 and 1983 and used a quasi-experimental design to examine the efficacy of the multisystemic approach for inner-city juvenile delinquents. With a small budget, Henggeler served as the principal investigator, while his graduate students served as therapists and data managers (Henggeler et al., 1998). Youths and their families were assigned to the MST condition through a federally funded community delinquency diversion program, while comparison participants were matched on common demographic characteristics of the families in the MST condition. Comparison group participants were also part of the delinquency diversion program; however, they were not assigned to MST. Although the study's result were not published until almost ten years later, the findings demonstrated that MST was a "promising model for improving the relations of dysfunctional families and decreasing the behavior problems of relatively serious delinquents" (Henggeler, et al., 1986, p. 139).

The second clinical trial of MST was published in 1987 and examined the efficacy of MST with maltreating families referred by the Department of Social Services (Brunk et al., 1987). It is instructive that this study was the first randomized trial of MST and the first evaluation to use a well-specified treatment for the comparison group (Henggeler et al., 1998). More specifically, youths and their families were randomly assigned to a home-based intervention program (i.e., MST) or a clinic-based behavioral parent-training program. Similar to the first clinical trial, MST was provided by doctoral students in clinical psychology. Parent training, on the other hand, was provided by mental health professionals employed at a community mental health center (Brunk et al., 1987).

Results from the second clinical trial showed similar improvements across the two treatment conditions. For example, parents in both the MST and parent-training group reported decreased psychiatric symptomatology, reduced overall stress, and a reduction in the severity of problems. Through observations, however, it appeared that parent-child interactions improved considerably for the MST group, compared to their parent-training counterparts. More specifically, "parent-child dyads in the MST condition showed the types of changes in interaction that suggest decreased risk for maltreatment: Maltreating parents controlled their child's behavior more effectively, maltreated children displayed less passive noncompliance, and neglecting parents became more responsive to their child's behavior [compared to the comparison group]" (Henggeler et al., 1998, p. 239).

After graduating from Memphis State University with a Ph.D. in clinical psychology, Charles Borduin took his first academic position in the Department of Psychology at the University of Missouri—Columbia in 1982 where he continued his MST research agenda (Henggeler et al., 1998). With the continued help of Henggeler, Borduin conducted two pivotal

randomized studies that would once again provide support for the MST approach with serious juvenile delinquents (Borduin et al., 1995), as well as support for the program with juvenile sex offenders (Borduin, Henggeler, Blaske, & Stein, 1990).

As Henggeler and his colleagues (2009, p. 282) explain, "Together, the favorable outcomes (e.g., reduced rearrest and incarceration, improved family functioning) achieved in this work set the stage for sponsored research, funded by the National Institutes of Health, other federal entities, foundations, and state and county governments, that has expanded the domain of MST well beyond graduate schools in clinical psychology." In short, Henggeler and his teams' early studies provided clear evidence that MST could be effective in improving family functioning and decreasing antisocial behavior of delinquents. It was now time to transition into the next phase of research to determine whether MST could be implemented successfully in real world clinical settings.

Later Studies by Henggeler's Team

Two Initial Studies. Later studies conducted by Henggeler and his team are also important for understanding how MST became an evidence-based intervention. More specifically, the results from two randomized effectiveness studies in South Carolina were the first trials to closely replicate real-world clinical conditions. Both studies measured the impact that MST had on chronic juvenile offenders at imminent risk of incarceration and were conducted in mental health centers with clinical professionals and master's level practitioners serving as therapists (Henggeler et al., 2009).

During the first effectiveness trial in Simpsonville, South Carolina, families were randomly assigned to MST or usual Department of Juvenile Justice (DJJ) services (e.g., incarceration, probation, and/or referral to community-based services) (Henggeler et al., 1998).

In contrast to the efficacy trials described above, Henggeler served as an off-site consultant during the study where he was responsible for the initial training and for providing clinical assistance and support to the MST clinicians and supervisors (i.e., the role that the expert consultant plays in the present-day MST quality assurance and improvement system described previously) (Henggeler, 2011; Henggeler, Melton, & Smith; 1992; Henggeler, Melton, Smith, Schoenwald, & Hanley, 1993; Henggeler et al., 2009).

The results from the study demonstrated significant effects favoring MST. At a 59-week follow-up, youth in the MST condition evidenced a 43 percent decrease in recidivism and a 64 percent reduction in out-of-home placement (Henggeler et al., 1992). Further, at a 2.4-year follow up, MST recidivism effects remained significant (Henggeler et al., 1993). The study also found that compared to youths in the usual services group, youths in the MST group reported increased family cohesion, as well as decreased adolescent aggression with peers (Henggeler et al., 1998). Although treatment effects in this study were not "quite as powerful as those observed in the efficacy studies, this work made a major step in bridging the science-to-service gap by demonstrating the successful transport of MST programs to community-based providers" (Henggeler, 2011, p. 364).

Following the positive results found in the Simpsonville study, a second project was funded to determine whether MST could be effective when experts in MST did not provide significant off-site clinical supervision (Henggeler, 2011). Although such oversight had taken place in previous successful trials, there was concern that such extensive involvement of an MST expert might create "financial and administrative barriers to the broader dissemination of [the program]…" (Henggeler et al., 1997, p. 822). This was important because if the treatment

effects found in previous studies could not be replicated, the failure could be linked to difficulties in therapist adherence to the treatment protocol (Henggeler et al., 1997; Henggeler et al., 1998).

The study was conducted at two public community mental health center sites in Orangeburg and Spartanburg, South Carolina where youths were randomly assigned to MST or usual DJJ services. Ten master's-level therapists in the field of mental health participated in the project. With the exception of employing an expert consultant, other components of the MST protocol remained in place (i.e., five-day orientation training, booster trainings, and on-site clinical supervision). Because of the emphasis placed on MST treatment adherence in the study, "parent, adolescent, and therapist ratings of MST treatment adherence were assessed, and associations between adherence and clinical outcomes [were] reported" (Henggeler et al., 1997, p. 322).

Although MST improved adolescent psychiatric symptomology at post-treatment and achieved a 47 percent reduction in days incarcerated at 20-months follow-up, significant effects on instrumental outcomes (e.g., peer relations, family relations) and criminal activity were not found (Henggeler et al., 1997). Significant effects were found between treatment outcomes and therapist adherence to the MST treatment principles, however. More specifically, "Parent and adolescent ratings of treatment adherence predicted low rates of rearrest; and therapist ratings of treatment adherence and treatment engagement predicted decreased self-reported index offenses and low probability of incarceration, respectively" (Henggeler et al., 1997, p. 829). In short, higher treatment fidelity was linked with lower recidivism (Henggeler et al., 2011).

The findings from the Orangeburg/Spartanburg study "demonstrated the importance of including fidelity measures in clinical trials as well as the significance and value of the expert consultant role in MST programs" (Henggeler, 2011, p. 364). In other words, because an MST

expert did not provide ongoing consultation to support practitioner fidelity to the MST model, there was variability in following the MST treatment protocol and thus, an overall reduction in the effectiveness of MST (Henggeler et al., 1997).

Studies of Other Clinical Outcomes. Up to this point, MST targeted chronic and violent juvenile offenders and their families. However, two medical university-based research groups began to examine the MST approach for treating other serious clinical problems (Henggeler et al., 2009). In addition, Henggeler (or Borduin who was also directly involved in the initial MST clinical trials) provided ongoing clinical oversight. This oversight was the responsibility of a second generation of MST expert consultants who were trained by the developers of MST (Henggeler, 2011). Specifically, researchers trained in MST from the department of Pediatrics at Wayne State University and researchers from the Department of Psychiatry at MUSC served as the expert consultants for these projects (Henggeler et al., 2009).

Notably, the results from these second generation-led studies supported the generalizability of MST to a range of serious clinical problems presented by adolescents and their families (Henggeler, 2011). Specifically, promising results were found in studies that examined the effectiveness of MST for youths with serious emotional disturbance (Henggeler, Pickrel et al., 1999; Huey et al., 2004; Rowland et al., 2005), juvenile offenders with substance use disorders (Henggeler, Rowland, et al., 1999; Henggeler, Clingempeel, Brondino, & Pickrel, 2002), physically abused adolescents (Swenson, Penman, Henggeler, & Rowland, 2010; Swenson, Schaeffer, Henggeler, Faldowski, & Mayhew, 2010), and adolescents with chronic health care problems (Ellis et al., 2004; Ellis, et al., 2005; Naar-King et al., 2009).

The studies also demonstrated that second-generation expert consultants could potentially be the key to successfully transporting the program on a larger scale (Henggeler, 2011). In this

way, MST Services, Inc. was established to actively work with agencies to facilitate program adoption and implementation with fidelity (Henggeler, 2011). In addition, MST Services, Inc. also provides training and ongoing quality assurance to other network partners who are responsible for carrying out all of the aspects of the program, from development, to implementation, to monitoring for program fidelity and outcomes. These network partners "might be the treatment developers themselves, individuals trained by the developers (e.g., second-generation MST experts), or individuals trained by second-generation experts (e.g., third-generation MST experts), and so forth" (Henggeler, 2011, p. 366).

Independent Studies of MST

Beyond Henggeler and his colleagues, MST has gained credibility from a number of diverse sources. These evaluations were undertaken by researchers having no ties to Henggeler and his research team. Because theses studies were independent, they added credence to the conclusion that MST is an effective, evidence-based intervention.

The first independent replication of the MST model was conducted in Norway in 2004 with youths presenting serious antisocial behavior and their families (Ogden & Halliday-Boykins, 2004). Youths were randomly assigned to the MST condition or usual child welfare services (e.g., foster care, institutional placement). Findings from this four-site Norwegian study revealed that, compared to their child welfare services counterparts, youths in the MST condition had decreased externalizing symptoms (e.g., disruptive, hyperactive, and aggressive behaviors) and internalizing symptoms (e.g., withdrawn, anxious, and depressed behaviors) of behavior at six-months post-recruitment. In addition, the MST group had a 78 percent reduction in out-of-home placements, as well as significantly increased social competence skills, compared to the child welfare services group. A 24-month follow-up revealed that MST effects on youths'

internalizing symptoms and out-of-home placements were maintained (Henggeler, 2011; Ogden & Hagen, 2006). Importantly, these first independent studies demonstrated that MST could be implemented with fidelity outside of the United States and without the developers' direct involvement (Henggeler, 2011).

Timmons-Mitchell and her colleagues conducted the first real-world randomized clinical trial of MST in the United States without direct oversight by the developers of the program (Timmons-Mitchell, Bender, Kishna, & Mitchell, 2006). Study participants were juvenile felons at imminent risk of incarceration who were randomly assigned to MST or treatment as usual (TAU). At approximately 12-months post recruitment, youths assigned to the MST condition showed improved mood and school/work functioning, and decreased substance use. Further, at approximately 24-months post recruitment, results indicated that the MST group had significantly decreased rearrest rates, compared to their TAU counterparts (37% decrease in rearrests).

In addition to the independent clinical trials discussed above, independent researchers have conducted quasi-experimental and benchmarking studies that have also found favorable support for MST (Henggeler et al., 2009). For example, using a quasi-experimental design, Stambaugh and her colleagues compared the effectiveness of MST versus wraparound for youths with serious emotional disturbance who were at risk of out-of-home placement (Stambaugh et al., 2007). Findings at an 18-month follow-up revealed that MST was more effective at decreasing youth symptoms, improving youth functioning, and decreasing out-of-home placements (a 54% decrease) (Henggeler 2011; Henggeler et al., 2009; Stambaugh et al., 2007).

Benchmarking studies "compare the strength of treatment effects in a community-based implementation of an evidence-based treatment with the strength of the effects achieved in

previous clinical trials of that treatment" (Henggeler et al., 2009, p. 288). For example, in a benchmarking study conducted in Norway, Ogden, Hagen, and Anderson (2007) examined the sustainability of MST effectiveness. Specifically, the researchers investigated whether the effects of MST on a second group of adolescents could match those reported in the first independent study discussed above (Ogden & Hagen, 2006; Ogden & Halliday-Boykins, 2004). Briefly, it was found that MST clinical outcomes (e.g., reducing antisocial behavior and out-of-home placement) in the second year of program operation matched or surpassed those achieved during the program's first year of operation. These results are important because they demonstrated that MST could sustain its effectiveness over time (Henggeler et al., 2009).

In another benchmarking study, Curtis and her colleagues compared pre-post findings of MST programs in New Zealand with results from clinical trials conducted in the United States (Curtis, Ronan, Heiblum, & Crellin, 2009). The findings from this New Zealand study found that rates of successful completion were considerably high (98%) and that clinical outcomes (e.g., days in out-of-home placement, frequency and severity of offending behavior, school and vocational attendance) were consistent with those achieved in previous MST studies. In short, both the Norwegian and the New Zealand studies "provide further support for the viability of MST in both real world and international contexts" (Henggeler et al., 2009, p. 289).

Independent Large-Scale Reviews in Support of MST. In light of the increasing evidence favoring MST, several large-scale studies and program reviews have been conducted that further support the program's dissemination. These areas of research fall into three general categories: (1) meta-analyses; (2) federal initiatives; and (3) cost-benefit analyses.

First, three meta-analyses conducted between 2003 and 2014 yielded results in favor of MST. For example, in their review and meta-analysis of the effectiveness of family-based crime

prevention programs (e.g., home visiting programs, day care/preschool programs, parent training programs, school based programs, home/community programs with older children, and multisystemic therapy programs), Farrington and Welsh (2003) found that the "most effective family-based approach was multisystemic therapy" (Farrington & Welsh, 2007, p. 135; see also, Farrington & Welsh, 2003).

Similarly, Curtis, Ronan, and Borduin found positive results for MST in their meta-analysis of the program in 2004. More specifically, Curtis and her colleagues found, that across the 11 studies examined, the rate, frequency, and seriousness of adolescent criminal activity was better for MST participants, compared to participants who received usual services or another treatment program. In addition, MST participants, on average, spent fewer days incarcerated, had better school attendance, and reduced substance use behaviors, compared to comparison group participants (Curtis et al., 2004).

Finally, a recent meta-analysis conducted by van der Stouwe and her colleagues examined the effectiveness of MST and updated previous meta-analyses by Curtis et al. (2004) and Littell et al. (2005) by including a larger number of studies (k = 22 studies; 332 effect sizes, N = 4,066 juveniles) (van der Stouwe, Asscher, Stams, Dekovic, & van der Laan, 2014). Small but significant treatment effects were found for the primary outcome of interest—delinquency (d = 0.20)—and also on secondary outcome measures including, psychopathology (d = 0.27), substance use (d = 0.29), family factors (d = 0.14), out-of-home placement (d = 0.27), and peer factors (d = 0.21). There were no significant treatment effects found for the skills and cognitions outcome, however.

Moderator analyses revealed several important findings in van der Stouwe et al.'s study. Specifically, study characteristics (e.g., country where the research was conducted, efficacy

versus effectiveness studies, and quality of study), treatment characteristics (e.g., single versus multiple control treatments, duration of MST treatment), sample characteristics (e.g., target population, age, gender, and ethnicity), and outcome characteristics (e.g., general delinquency versus violent/non-violent offending) moderated the effectiveness of MST. In general, the scholars concluded that "MST seems most effective with juveniles under the age of 15, with severe starting conditions" (2014, p. 468). The authors argue, however, that the effectiveness of the program may be improved for older youthful offenders if treatment focuses more on peer relationships and risks and protective factors related to school outcomes (van der Stouwe et al., 2014).

Second, several federal supporters and initiatives have also contributed to the success of MST and its dissemination across the United States and abroad. The Office of Juvenile Justice and Delinquency Prevention (OJJDP), for example, funded several projects that "focused on identifying and supporting the broader use of effective juvenile justice prevention and intervention strategies" (Schoenwald, 2010, p. 566). One of the initial projects stemming from OJJDP's efforts was conducted by the University of Colorado, Center for the Study and Prevention of Violence (CSPV) (Elliott, 1998). Specifically, CSPV reviewed the literature on the prevention and treatment of youth violence and delinquency and was responsible for developing blueprints or "summaries of clinical procedures and outcomes from evaluations and implementation measures for their transport" (Henggeler, 2011, p. 374). To date, "more than 1,300 programs have been reviewed [by CSPV and a distinguished Advisory Board], but less than 5 [percent] of them have been designated as model and promising programs" (Blueprints Programs, 2014, p. 1).

During the development of the Blueprints for Violence Prevention Series, MST was designated as a model program by CSPV which aided in its widespread dissemination during the late 1990s. Subsequent projects funded by OJJDP, provided funding for the development of various program manuals (e.g., consultation, supervision, and organizational manuals) and an automated outcomes tracking system that are all central to the MST quality improvement processes discussed previously (Henggeler, 2011, Schoenwald, 2010). Finally, OJJDP continues to provide grants to states for "activities to reduce juvenile offending and criteria for measuring the effects of funded activities" (Schoenwald, 2010, p. 567).

It is important to note that MST's recognition and support stems beyond OJJDP and CSPV. For example, the Office of Justice Programs' CrimeSolutions.gov was created for practitioners and policymakers to inform them about evidence-based programs and practices in criminal justice, juvenile justice, and crime victim services. From their rigorous rating process, MST was deemed "effective" on CrimeSolutions.gov's three-point scale (ranging from no effects, to promising, to effective). Several other federal and state supporters have recognized MST as an effective program (e.g., Washington State Institute for Public Policy [WISPP], National Institute of Health [NIH], Substance Abuse and Mental Health Services Administration [SAMHSA], Institute of Public Policy Research [IPPR]), which subsequently have helped to transport MST across the United States and overseas.

Finally, cost benefit analyses have shown that MST is not only an effective early intervention strategy for reducing delinquency and improving a variety of other outcomes, but it is also cost effective (i.e., it lowers the total costs to taxpayers and crime victims) (Cullen & Jonson, 2011). Steve Aos and his colleagues at WSIPP, for example, have developed a cost-benefit analysis model to estimate a range of costs and benefits for both taxpayers and crime

victims (Aos, Lieb, Mayfield, Miller, & Pannucci, 2004; Aos, Phipps, Barnoski, & Lieb, 2001; Aos et al., 2011; Drake, Aos, & Miller, 2009). The WSIPP model is important because it is able to obtain the "bottom line" (i.e., the return on investment) of each program included in the CBA (Dopp, Borduin, Wagner, & Sawyer, 2014). Notably, WSIPP reported net benefits ranging from $9,316 to $131,918 for each MST participant. This results in returns of $2.64 to $28.81 for every dollar spent on MST (Aos et al., 2001, 2004, 2011; Drake et al., 2009). In this way, WSIPP has identified MST as a treatment program that is likely to reduce taxpayer and crime victims' costs (Dopp et al., 2014).

Independent Large-Scale Studies Not in Support of MST. Not all independent studies of MST have been able to replicate favorable outcomes. In a large-scale randomized trial of MST in Ontario, Canada, for example, Leschied and Cunningham (2002) (see also, Cunningham, 2002) found that MST programs averaged a 10 percent reduction in convictions across four-sites; however, these results were "considerably below the findings observed in published trials of MST" (Henggeler et al., 2009, p. 289). Further, although short-term family- and youth-level outcomes appeared to have promising effects, these results failed to reach statistical significance (Leschied & Cunningham, 2002). It is instructive to note that across the four MST sites included in Leschied and Cunningham's (2002) study, overall adherence to the MST model (i.e., treatment fidelity) was the lowest in the site with the poorest outcomes.

Sundell and his colleagues also failed to find favorable outcomes for treatment participants in their multisite study of MST in Sweden (Sundell, Hansson, Lofholm, Olsson, Gustle, & Kadesjo, 2008). Specifically, the researchers found that MST participants did no better or worse than youths who participated in usual treatment services. In most cases, youths in *both* the MST and TAU groups decreased their problem behaviors, showed improved relations

with family members, and improved their social skills. Importantly, none of these findings were statistically significant between the groups.

Similar to Leschied and Cunningham (2002), treatment fidelity was also a concern in Sundell et al.'s study. Although treatment adherence was remarkably low across all of the MST sites, results indicated that sites with higher overall adherence scores were associated with more favorable youth outcomes. "Interestingly, with regard to reduced symptomatology, youth in the MST condition in Sweden fared at least as well as counterparts in the successful Norwegian and U.S. trials [discussed previously], suggesting that the failure to attain MST effects might have been due to the relative strength of usual services (i.e., the comparison condition) in Sweden" (Henggeler et al., 2009, p. 289).

Finally, Littell, Popa, and Forsythe (2005) stirred up a fair amount of controversy in their meta-analytic review of MST that concluded "MST is not consistently more effective than other alternatives for youth with social, emotional, or behavioral problems" (p. 25) (see also, Littell, 2005). Due to rigorous inclusion criteria, only eight of 35 distinct MST outcome studies conducted between 1985 and 2004 met the inclusion criteria for Littell and her colleagues' study. "Although seven of the eight studies included in [their] review found significant differences favoring MST on one or more outcome measures, these effects were not consistent across studies, and the heterogeneity among studies [made] it impossible to know which factors or combinations of factors account[ed] for the differences between them" (Fonagy, Cottrell, Phillips, Bevington, Glaser, & Allison, 2015, pp. 167-168). Because of the variation in the studies' quality and context, the scholars argued that it was premature to draw conclusions about the effectiveness of MST compared with other services (Littell et al., 2005, p. 4).

THE NEED FOR A SYSTEMATIC REVIEW: RESEARCH STRATEGY

It has been more than three decades since MST was born at Memphis State University when Henggeler and his graduate students were assigned the "diversion project." As this chapter has described, MST has grown exponentially since its birth, and as a result, research exploring its effectiveness has also expanded at a rapid rate.

A quasi-experimental study of MST with juvenile delinquents (Henggeler et al., 1986) and three small-scale randomized efficacy trials of MST for other populations (e.g., maltreated youths and their families, juvenile offenders, and juvenile sex offenders) were published in academic journals between 1986 and 1990 that help set the stage for the program's success early on (Borduin et al., 1985; Borduin et al., 1990; Brunk et al., 1987). These initial clinical trials were important particularly because they set the stage for two large-scale community-based trials of the program in South Carolina (Henggeler et al., 1992; Henggeler et al., 1993; Henggeler et al., 1997). The results of these studies further caught the attention of agencies at the national level and "the demand to transport MST to usual-care settings began…" (Schoenwald, 2010, p. 564). With the transport of MST across the United States and abroad, MST Services, Inc. was established to work closely with network partners and subsequently, build a structure that would continue to keep up with the program's growth and expansion (Henggeler, 2011).

With the attention MST has received over its more than 30-year lifespan, it is not surprising that independent studies and large-scale reviews of the program have been conducted. As can be gleaned from the previous discussion, however, these studies have found mixed results and have even called into question the efficacy of MST altogether (i.e., Littell et al., 2005). Even further, a more recent study conducted by van der Stouwe and her colleagues (2014) found small, but significant treatment effects for MST. Importantly, their findings demonstrate that

particular moderating variables may be able to explain why MST might work better when conducted in particular settings and/or with particular types of youths.

In line with the above discussion, the overall goal of the current dissertation is to advance the extant literature on the effectiveness of MST and determine more in depth, what factors or circumstances lead to the strongest treatment effects for the program. Specifically, the objectives of the study are threefold. First, this dissertation involves the analysis of 44 unpublished and published studies, providing the most comprehensive review of studies on the effectiveness of MST to-date.

Second, this dissertation examines the effectiveness of MST for youths and their families on nine dependent variables, including: (1) delinquency, (2) problem behavior, (3) psychopathology and mental health, (4) family functioning and relationships, (5) peer relationships, (6) school performance, (7) parent functioning and relationships, (8) substance abuse, and (9) service utilization.

Third, to determine the effectiveness of MST on the abovementioned outcome variables, a meta-analysis is conducted. From this quantitative synthesis process, the overall mean effect size, weighted mean effect size, and corresponding confidence intervals are calculated for the independent variable (i.e., MST) and each dependent variable under review. In addition, the impact of several moderating variables is also explored. Moderators are examined across five general categories, including: (1) publication characteristics, (2) study context characteristics, (3) sample characteristics, (4) treatment characteristics, and (5) study characteristics.

CONCLUSION

Given that MST is increasingly being disseminated across the U.S. and abroad, it is important to understand the program's overall treatment effects. Although many individual

studies and a small number of comprehensive reviews have been conducted on the effects of MST on subsequent delinquency and other outcome measures, results have been mixed and have recently called attention to the importance of moderating factors or characteristics that may help to explain why we are seeing these mixed results. As such, this dissertation serves as an independent assessment of the existent studies and comprehensive reviews.

In the chapters to come, the methods used to conduct the meta-analysis will be described. In addition, the results of the study will be presented and the theoretical and policy implications of these results will be highlighted. Specifically, Chapter 2 will focus on the methods used in the study, including a review of the strengths and weaknesses of meta-analysis, as well as an overview of the sample of the included independent studies, a description of the dependent and independent variables, and finally, a review of the analyses employed to conduct the meta-analytic review. Chapter 3 will present the results of the meta-analysis for each of the previously identified outcome variables, as well as present the results of moderating analyses. Finally, Chapter 4 will present the theoretical and policy implications of the current study, as well as highlight potential areas of research that should be examined in the future.

CHAPTER 2

METHODS

The purpose of the current chapter is to describe the methods used to examine the effectiveness of MST. First, a brief description of meta-analysis is presented, along with a discussion of the strengths and weaknesses of this statistical technique. Second, the sample of studies selected for the current project is described. This discussion provides an overview of the eligibility criteria used to select studies into the project, as well as describes how the studies were located. Third, the nine dependent variables examined in the study are described. Fourth, the independent variable, MST, is presented. Fifth, the moderating variables included in the study are described. Finally, the analyses employed as part of the meta-analysis are explained.

QUANTITATIVE SYNTHESIS OF RESEARCH STUDIES: CONDUCTING A META-ANALYSIS

Reviewing Research: Two Traditional Methods

In a utopian world, results of a study would extend, or at the very least, support, findings from previous scientists' work so that research on a particular topic was unanimous and consistent. Trends in research, however, tend to do the opposite. Specifically, as research on a topic increases, so too does the diversity of results (Hunt, 1997). Scientists are subsequently left with an accumulating body of evidence that they must disentangle and make sense of—an exercise that is likely to conclude with nothing more than a list of "what might work" and "what might not work" (Hunt, 1997).

Traditionally, researchers have used two methods to synthesize or make sense of large bodies of research: (1) narrative reviews, and (2) vote counting or ballot box reviews (Hunter & Schmidt, 1990). Both approaches, however, are plagued with methodological flaws that

consequently impact the findings of the topic under review (Hunt, 1997; Wolf, 1986). These approaches are discussed below as a prelude to further examining the technique of meta-analysis.

The Narrative Review. First, in a narrative review, the author reads the existing literature on a specific topic and draws general conclusions based on what this research has found (Cullen & Gendreau, 2000). In this way, a narrative review is a qualitative study whereby the author reviews the research—reading the available literature study by study—and then draws conclusions on whether a relationship exists between the measures of interest. As discussed in the introduction to Chapter 1, for example, Gendreau and Ross (1979, 1987) attempted to "revive rehabilitation" by publishing two large-scale narrative reviews during the post-Martinson era (Cullen & Gendreau, 2001). Their first review assessed 95 studies conducted between 1973 and 1978 and their second review assessed 130 studies conducted between 1981 and 1987. Both of these studies ultimately concluded that correctional treatment *can* be effective.

Although the narrative review allows the reviewer to "focus on the richness of individual studies" and also allows them to "interpret what the research, taken together, 'really means'" (Cullen & Gendreau, 2000, p. 133), this approach can be problematic for several reasons. First, there are no set criteria or standardized procedures for how to conduct a narrative review, how to find studies for review, or how to organize the studies that were found (Hunt, 1997). Even further, because the narrative review is a qualitative technique, it is more likely to bring in the reviewer's judgment or biases (i.e., it is subjective). In this way, two reviewers may read the same literature on a topic and come to very different conclusions (Glass, 1976).

Second, a narrative review is difficult to conduct when the research on a topic has considerably expanded (Hunter & Schmidt, 1990). That is, it is easy to get lost in the details and variations of each study and then have no way to quantitatively summarize the results. Some

scholars have chosen a subset of studies to review to make the existing evidence more manageable; however, this causes another set of problems. For example, because the reviewer has omitted some studies, he or she is not providing the audience with a full review of the results to-date on the topic of interest. Even further, when authors conduct a narrative review, they tend to decide what the truth is, either before starting the review, or after their initial exploration of the topic (Chalmers & Lau, 1994). This in turn leads them to "defend their conclusions by citing all the evidence they can find" (cited in Hunt, 1997, p. 7). In other words, they only present the evidence supporting their "truth," leaving out studies that might go against it.

The Ballot Box Review. A second method traditionally used to synthesize research includes the vote counting or "ballot-box" review. Attempting to overcome the subjective nature of narrative reviews, a ballot box review divides studies under examination into three categories: (1) studies that show positive effects on the outcome of interest; (2) studies that show negative effects on the outcome of interest; and (3) studies that do not show any effect on the outcome of interest (Hunter & Schmidt, 1990). Once studies are placed into the appropriate categories, the studies showing statistically significant results are tallied or counted. The category with the most counts or the most statistically significant outcomes is considered to be the overall effect (for an example, see Palmer, 1975).

Similar to narrative reviews, however, ballot box studies can be problematic. Most importantly, ballot box reviews give equal weight to all studies included in the evaluation and do not take the size of the sample into consideration (Hunt, 1997). For example, studies with a sample size of 50 participants are given the same weight in a ballot box review as studies with a sample size of 1,000 participants. In addition, "ballot box reviews reduce each individual study to a dichotomy: each individual study is rated as either a success or a failure" (Whitehead & Lab,

1989, p. 278). In this way, it simplifies the results of a study, which creates room for subjectivity.

Related to the above discussion, the magnitude of the effect is not examined in a ballot box review. That is, a ballot box review will demonstrate whether a relationship exists between the independent variable and the dependent variable, but it will not tell the reviewer how strong that relationship is. For example, two studies in a review might yield statistically significant results—one study demonstrating positive effects and one study demonstrating negative effects on the outcome variable. Although the study demonstrates whether a relationship in both directions exists, there is no way to tell which relationship is stronger between the two studies (Hunt, 1997).

Finally, ballot box reviews are prone to type II errors (i.e., researchers run the risk of failing to detect a relationship when in fact, there is a relationship present). Because ballot box reviews only count studies that are statistically significant, studies with smaller sample sizes oftentimes do not receive a vote. It other words, because it is more difficult for studies with smaller sample sizes to reach statistical significance, they do not count in a ballot box review, even when the study's findings show modest effects.

Taken together, both the narrative review and the ballot box review allow for a considerable amount of subjectivity from the reviewer. In addition, "even when the results are agreed upon, it is often difficult to determine whether the glass is 'half full" or "half empty'" (Cullen & Gendreau, 2000, p. 125). Ultimately, different reviewers may interpret the benchmark for success differently; some concluding that the program or treatment is successful, while others concluding that the program or treatment is unsuccessful.

What is a Meta-Analysis?

Meta-analysis is a more sophisticated way of assessing the evidence on a particular topic; because it is quantitative in nature, it allows the reviewer to take a more objective approach to synthesizing a large body of literature. Glass first coined the term meta-analysis in 1976, defining it simply as the "analysis of analyses" (1976, p. 3). Hunt (1997, pp. 1-2) further explains that meta-analysis:

> is a means of combining the numerical results of studies with disparate, even conflicting, research methods and findings; it enables researchers to discover the consistencies in a set of seemingly inconsistent findings and to arrive at conclusions more accurate and credible than those presented in any one of the primary studies. More than that, meta-analysis makes it possible to pinpoint how and why studies come up with different results, and to determine which treatments—circumstances or interventions—are more effective and why they succeed.

Whereas previous approaches "tallied the votes" to determine how many studies worked and how many studies did not work, meta-analysis computes an effect size for each study under review. This effect size tells a reviewer not only what direction the relationship is headed in, but also tells them the magnitude or strength of the relationship between the independent variable (e.g., juvenile correctional treatment) and the dependent variable (e.g., recidivism). The effect size, for example, may tell the reviewer that the treatment has no effect on recidivism (i.e., the effect size is zero), that the treatment is actually increasing recidivism (i.e., the effect size is a negative number), or that the treatment is working to decrease recidivism (i.e., the effect size is a positive number). Once these individual effect sizes are computed, the reviewer calculates the

average effect size, yielding a "precise point estimate of the relationship of the treatment on the outcome measure across all studies" (Cullen & Gendreau, 2000, p. 134).

Since Glass first introduced meta-analysis in the mid-1970s, meta-analytic techniques have been used increasingly in both the physical and social sciences (Cullen & Gendreau, 2000). For example, meta-analysis has been used in medicine (see e.g., He, et al., 1999; Strazzullo, D'Elia, Kandala, & Cappuccio, 2009; Rhodes, et al., 2004; Wilson & Giguere, 2008), agriculture (Alston, Chan-Kang, Marra, Pardey, & Wyatt, 2000; Armstrong, 1994; Bengtsson, Ahnstrom, & Weibull, 2005; Thiam, Bravo-Ureta, & Rivas, 2005), and other behavioral sciences (Albarracin, Fishbein, Johnson, & Muellerleile, 2001; LeFrance, Hecht, & Paluck, 2003; McClure, 2000; Stanne, Johnson, & Johnson, 1999; Wolf, 1986), as well as more recently in the field of criminal justice (see e.g., Andrews, Zinger, et al., 1990; Dowden & Andrews, 1999, 2000; Gendreau et al., 1996; Hanson & Bussiere, 1998; Lipsey, 1992). Despite the numerous studies that have been published across disciplines, meta-analysis has not been without its critics. The following subsection will present the strengths and weaknesses of this methodological technique.

Strengths of Meta-Analysis

As discussed in the previous section, traditional methods for combining data from multiple studies had been completed primarily through narrative or ballot box reviews prior to the 1990s (Borenstein, Hedges, Higgins, & Rothstein, 2009). For a number of reasons, however, both methods suffer from several important limitations, which has led researchers in many fields to adopt and favor meta-analytic techniques (Borenstein et al., 2009). Described by some scholars as the new "gold standard" in corrections (see e.g., Gendreau, Smith, & French, 2006, p. 429), meta-analysis is superior to other methods for several reasons.

Objectivity and Replication. First, because meta-analysis is a quantitative technique,

reviewers can systematically investigate the relationship between two variables using statistics, rather than drawing general conclusions on their own. Ultimately, results are driven by data, making meta-analysis a more objective research strategy compared to more qualitative research methods such as narrative and ballot box reviews.

Meta-analysis is an objective statistical technique for several other reasons. For example, authors of a meta-analysis must establish a clear set of rules and standards for conducting their review. In this way, "the rules and strategies that are followed in a meta-analysis for including, abstracting, weighting, and integrating studies are generally more objective" (Mullen, 1989, p. 8), compared to the methods followed by those conducting a narrative or ballot box review. Since any given author must set these rules and standards when conducting a meta-analytic review, there is inevitably an element of subjectivity; however, because all decisions are clearly specified and made "public," the "mechanisms are transparent" (Borenstein et al., 2009, p. 736; Mullen, 1989; Pratt, 2001).

In addition, because rules and strategies are made public and data-driven results lend themselves to more objectivity, meta-analyses can be replicated, especially if the same studies are included for analysis and the same rules are followed (Mullen, 1989). Thus, scholars who are skeptical of the original author's methods and "decision-rules," may conduct their own meta-analysis to determine whether the results are valid. Replicating a narrative or ballot box review may not be as easy to achieve because these methods do not require the authors to make the strategies and standards employed explicit or public.

Number of Studies. Second, as Cullen and Jonson (2012, p. 159) candidly state, "a key strength of meta-analysis is that it can take several hundred studies and tell you with one itty-bitty number what the effect of treatment is on recidivism." In this way, even when the research

on a topic has considerably expanded, meta-analysis makes it more feasible for reviewers to "wade through the data" because it treats each study as a unique case in a dataset (Hunt, 1997). From here, the reviewer can compute the effect size for each study and then calculate the mean effect size overall across all studies. In short, even when there are hundreds, or even thousands of studies on a topic, meta-analysis allows for a true synthesis of the existing empirical research on that topic (Lipsey & Wilson, 2001).

Magnitude, Direction, and Precision of the Effect Size. Third, unlike narrative reviews where the reviewers simply describe the relationship they think they see, or a ballot box review where the reviewers "tally" up the significant effects that are reported, authors conducting a meta-analysis can present an effect size estimate that tells them the magnitude of the effect and the direction that effect is headed (i.e., positive, negative, or no direction). Even further, they may compute the confidence interval around the effect size to determine how precise their estimates may be (Gendreau & Smith, 2007). Importantly, both effect size estimates, as well as confidence intervals around those effect sizes, can be calculated for all studies, regardless of whether the individual study found statistically significant results. It is also important to note that meta-analysis allows the reviewer to take into account studies' sample sizes, which in turn, decreases their probability of committing a type II error—something, recall, that a ballot box review is prone to do. Specifically, each study included for review is weighted so that studies with larger sample sizes are given more weight, while studies with smaller sample sizes are given less weight (Hunt, 1997; Hunter and Schmidt, 1990).

Moderating Variables. Finally, a meta-analysis allows reviewers to examine the relationship between the independent variable and dependent variable, while also examining how other moderators or key characteristics of a study are associated with the relationship. Although

moderating variables differ across meta-analyses, the literature categorizes moderators into four general areas: (1) demographic characteristics (e.g., age, sex, race); (2) study characteristics (e.g., year published, publication type); (3) methodological characteristics, (e.g., design of the study); and (4) measurement characteristics (e.g., how the independent and dependent variables are measured) (Durlak and Lipsey, 1991; Lipsey and Wilson, 2001). By examining some of these key characteristics, reviewers may be able to determine, for example, whether their outcome varies by a certain publication year or by a specific set of authors, and whether these differences are statistically significant (Lipsey & Wilson, 2001).

Criticisms of Meta-Analysis

Although meta-analysis has been embraced by a wide range of scholars across various fields, not every researcher has jumped on its bandwagon (Borenstein et al., 2009). In this way, the following section will briefly describe the criticisms associated with this statistical technique.

The File Drawer Problem. First, a meta-analysis will yield a mathematically accurate synthesis of the studies included in the analysis; however, if these studies represent a biased sample of all studies related to a particular topic, the meta-analysis itself (i.e., the overall effect size) is also likely to be biased (Borenstein et al., 2009). Not surprising, evidence shows that studies are more likely to be published if they report relatively high treatment effects, compared to those studies that find lower treatment effects. In this way, studies with unfavorable results, or less than exciting conclusions, are likely to "lie dormant in researchers' filing cabinets" (Borenstein et al., 2009, p. 8129). Since meta-analyses tend to rely exclusively on published research, any bias in the literature is likely to be reflected in the meta-analysis as well (Borenstein et al., 2009; Pratt, 2001). Scholars often refer to this problem as publication bias or the file drawer problem (Lipsey & Wilson, 2001; Rosenthal, 1979; 1984).

It is important to note that publication bias is not a limitation unique to meta-analyses. It can be problematic for scholars conducting both narrative and ballot box reviews as well. Borenstein and his colleagues (2009, p. 6249) argue that the issue "tends to receive more attention with regard to systematic reviews and meta-analyses, possibly because these are promoted as being more accurate than other approaches to synthesizing research."

To resolve the file drawer problem, two solutions have been presented. First, a truly comprehensive review of the existing literature should include both published and unpublished studies. In this way, reviewers can confidently state that they conducted a thorough investigation of the evidence related to their topic of interest. Evidence suggests that the increased resources and extra effort invested in to locating unpublished works—such as dissertations, theses, conference papers, and technical reports—is worthwhile, especially if the scholars wish to publish their findings (Borenstein et al., 2009).

Reviewers may also wish to statistically test for biases in their meta-analytic review by conducting a fail-safe N test—a method that cannot be used when conducting a narrative or ballot box review (Rosenthal, 1979). Put simply, a fail-safe N test helps a meta-analyst determine the magnitude of the file drawer problem. Once a meta-analysis has been conducted, the issue to be addressed is whether omitted studies—those presumably never published and hidden away in a file draw—would have changed the results if they had been included in the study. The assumption is that these omitted studies would have reported more non-significant findings. However, at some point, a meta-analysis might include a sufficiently large number of studies that the omitted file-draw studies would not have altered the findings even if they had been included. In this context, "a fail-safe number indicates the number of nonsignificant, unpublished (or missing) studies that would need to be added to a meta-analysis to reduce an

overall statistically significant observed result to nonsignificance." (Rosenburg, 2005, p. 464). Ultimately, the larger the fail-safe number relative to the number of studies included for analysis, the more confident reviewers can be about the representativeness of their dataset and the robustness of any significant findings (Harrison, 2010; Rosenburg, 2005).

Apples to Oranges Problem. Another common criticism of meta-analysis is that researchers combine different kinds of studies, or attempt to compare "apples to oranges," in the same study (Borenstein et al., 2009; Lipsey & Wilson, 2001). Opponents argue, for example, that the statistical technique combines studies with different independent and dependent variables and subsequently analyzes these studies as if the variables were the same. In this way, a meta-analysis is simply reporting a meaningless effect size because the studies included have different samples and different measures (Hunter & Schmidt, 1990; Logan & Gaes, 1993).

As discussed previously, one of the more subjective areas of meta-analysis is related to the inclusion/exclusion criteria the reviewer sets for his or her study. As such, different scholars may have desperate opinions surrounding the appropriateness of the studies included for review. What may look like all apples to one reviewer may look like a fruit basket to another.

It has been contended that scholars conducting a meta-analytic review should include only those studies with high methodological rigor, such as experimental and/or quasi-experimental studies. This approach, however, creates another set of problems because the vast majority of studies in any given field are typically not quasi- or experimental studies. In turn, relatively few studies would meet the eligibility criteria set forth by the author (Lipsey & Wilson, 2001).

It is also possible for the reviewer to statistically control for the variations found among studies. More specifically, the author may control for different study characteristics and/or

different methods used across studies to determine how variation is related to outcome (Lipsey & Wilson, 2001; Rosenthal, 1984). This technique would be the same as exploring different moderating variables related to outcome, as discussed in the previous section.

Garbage in, Garbage out. Finally, the "garbage in, garbage out" metaphor has been used to describe the quality of studies that are included in a meta-analysis. Put simply, if a meta-analysis includes mainly low quality studies, the meta-analysis is likely to produce a low quality effect size. Similarly, if there are errors in the primary study, these errors will not only be carried over into the meta-analysis, but be even harder to identify and disentangle (Borenstein et al., 2009).

Of course, the garbage in garbage out debate would be diminished if reviewers simply included only high quality studies in their meta-analysis; however, as discussed throughout the previous sections, this method would decrease the number of studies eligible for inclusion in the review and could possibly spark even more debate amongst researchers. For these reasons, Borenstein and his colleagues (2009, p. 9380) contend that meta-analysis should be a process of *waste management*, rather than a process of *garbage in, garbage out*. From their point of view, authors should include both high quality and low quality studies in their review, but they should "manage" it in such a way that high quality studies are given more weight, compared to low quality studies. In short, weighting the lower-quality studies allows the reviewer to include the studies while statistically correcting for their less accurate results (Harris, 1997).

SAMPLE OF STUDIES

In order to conduct a meta-analysis, the researcher must first gather all of the relevant literature on his or her topic of interest (Hunt, 1997; Lipsey & Wilson, 2001). In the current study, various techniques were used in an attempt to collect all of the studies examining the

effectiveness of MST. The following section briefly describes how both published and unpublished studies were found for the current meta-analysis.

First, to find published studies, a keyword search was conducted in various online databases, including Criminal Justice Abstracts, Criminal Justice Periodical Index, Dissertation Abstracts Online, ERIC, National Criminal Justice Reference Service, PsycINFO, Social Sciences Index, Sociological Abstracts, and SocINDEX. Second, the reference sections of the four MST meta-analyses discussed in Chapter 1 (Curtis et al., 2004; Farrington & Welsh, 2007; Littell et al., 2005; van der Stouwe et al., 2014) were examined to ensure all of the studies included in their meta-analytic reviews were also included in the current study. Third, the MST website (http://mstservices.com) was also referenced as MST Services, Inc. keeps an up-to-date list of relevant articles and publications on their site. Fourth, Scott Henggeler was contacted during the initial stages of the project. Although most of the studies Henggeler provided had already been found in online scholarly journals, requests for some outcome studies were made through MUSC.

Several techniques were also utilized to uncover unpublished studies related to MST. First, for example, annual conference programs for the American Society of Criminology and the Academy of Criminal Justice Sciences were examined to discover unpublished research. Second, a Google search was conducted that revealed several technical reports, as well as state and federal reports that were not uncovered during the initial keyword search. Finally, each state's Department of Youth Services (or their equivalent—e.g., Department of Children and Families, Division of Youth Services, Department of Youth Corrections) was contacted via e-mail and/or telephone. Agencies were specifically asked whether they had conducted any MST evaluations that might not have been published or posted online. E-mail addresses and telephone

numbers were obtained from states' juvenile justice department or corrections websites. Staff who were in charge of the state's MST program, head of the research/evaluation department, and/or executive directors were contacted for information. If the correct person was not contacted, most staff were willing to forward the e-mail to the appropriate person who could help with the study. If a response was not received within a week, a phone call was made to follow-up. Again, if the wrong person was contacted, the majority of staff were willing to forward the message on to the appropriate person who could help answer questions.

Although all 50 states, and the District of Columbia, were contacted (either by e-mail, telephone, or both), not all states responded to the investigation. Of the states that did respond (n=25), four states did not use MST, one state stopped using MST services due to the limited population it served, three states had recently implemented MST programs, (and thus their programs were too new to conduct outcome evaluations), and 13 states used MST in their facilities but had not conducted any research evaluating their program. Three states said they had conducted evaluations on MST but would not release these evaluations to outside parties. In the end, Connecticut was the only state that had an unpublished study on the effectiveness of MST programs with their juvenile and adolescent offenders; however, it could not be included for analysis because it did not include a comparison group.

Once studies were collected, several inclusion criteria were established (i.e., criteria that the studies had to meet in order to be included in the current meta-analysis). To be included, more specifically, the study must have: (1) compared youths who received MST to youths who received another type of treatment (e.g., individual therapy, family counseling) and/or service (e.g., probation); (2) examined the effectiveness of MST for youths presenting at least one of the nine dependent variables of interest in the current study; (3) presented an effect size estimate for

one or more of the outcomes of interest and/or provided data that could be used to calculate an effect size for one or more of the outcomes of interest; and (4) been gathered and coded by December 2015.

DEPENDENT VARIABLES

As highlighted in the conclusion of Chapter 1, the current dissertation seeks to examine the effectiveness of MST for youths and their families on a variety of outcomes. Specifically, nine dependent variables were examined: (1) juvenile delinquency, (2) problem behavior, (3) psychopathology and mental health, (4) family functioning and relationships (5) peer relationships, (6) school performance, (7) parent functioning and relationships, (8) substance abuse, and (9) service utilization. Each study was coded using a detailed coding guide for recording outcomes and moderating factors (see Appendix G for a copy of the coding guide). Inter-rater agreement was based on a random selection of 10 studies that were coded by the author of the current dissertation and a member of her dissertation committee who is familiar with meta-analytic techniques. Rates of agreement were perfect (i.e., 100% agreement) for publication characteristics, study context characteristics, and study characteristics and relatively high (between 85% and 100% agreement) for sample characteristics, treatment characteristics, and effect size calculations.

Before each outcome measure is described, two issues should be discussed. First, several studies focused on multiple outcome measures of interest (e.g., delinquency, problem behavior, school performance, and family functioning and relationships), as well as employed various methods to measure those outcomes (e.g., parent/caregiver, teacher, and youth self-report assessments and official records to measure youths' performance in school). While the "multifocus issue" (i.e., a single study examining treatment effects across multiple dependent

variables) was overcome by separating dependent variables into nine distinct categories upon the onset of the study, a solution was needed to handle the "multimethod issue" (i.e., to address the studies that used a battery of assessments and official records to measure the same dependent variable). The goal was to find a methodological technique designed to handle statistical dependence so that each study contributed one effect size per dependent variable (Scammacca, Roberts, & Stuebing, 2014).

One approach used by scholars to resolve statistical dependence is to select a single measure to include based on a series of "decision rules" (e.g., for delinquency studies, choose the measure of recidivism at the most initial stage of the juvenile justice system, choose violent offenses over non-violent offenses, etc.) (Scammacca et al. 2014). These scholars caution, however, that this approach is appropriate only when the meta-analyst can make a strong case for including one measure over others (Card, 2012; Scammacca et al., 2014). Other scholars (see e.g., Smith & Glass, 1977), on the other hand, advocate treating the outcome measure as the unit of analysis by using all measures from assessments and records in a study regardless of how many are reported. While this approach seems intuitively pleasing, it permits studies that measured an outcome various ways "to contribute differentially more data to the analysis" (Durlak & Lipsey, 1991, p. 311). For example, study #1 could simply measure delinquency as new adjudications for felony offenses, while study #2 could measure delinquency as new arrests for felony offenses, new arrests for misdemeanor offenses, new adjudications, and new commitments to a residential facility. Ultimately, study #1 contributes one effect size whereas study #2 contributes four effect sizes. Thus, study #2 receives four times more weight in the meta-analysis than study #1 "even though on all other grounds they may be of equivalent importance" (Durlak & Lipsey, 1991, p. 311).

To preserve as much data as possible without violating independence, the current study used a third approach that is also commonly implemented in social science meta-analyses (see e.g., Bonta et al., 1998; Curtis et al., 2004; Morgan et al., 2016; Scammacca, Roberts, Vaughn, & Stuebing, 2015). Specifically, each assessment method was treated as the unit of analysis and effect sizes from multiple assessments were averaged so that each study yielded one mean overall effect size per dependent variable (see e.g., Durlak & Lipsey, 1991; Borenstein et al., 2009; Card, 2012). Although this method runs the risk of "punish[ing] studies for attempting to measure the impact of their treatment across a broad array of measures" (Scammaca et al., 2014), it does not force the reviewer to choose an estimate at random when they may not be able to make a case for including one measure or assessment over the other, nor does it allow particular studies to carry more weight than others.

The second factor that should be noted upfront is in regards to studies' follow-up periods. In particular, several scholars measured treatment effects across different points in time, either in the same article or report or in subsequent articles and reports. For example, in their randomized controlled trial of MST and a court mandated therapeutic intervention for young offenders, Butler, Baruch, Hickey, and Fonagy (2011) measured youth delinquency at three different points in time: (1) 6-months post treatment, (2) 12-months post-treatment, and (3) 18-months post-treatment. In another randomized trial, Henggeler and his research team published a series of articles between 1999 and 2004 examining the effectiveness of MST over psychiatric hospitalization in changing various behaviors for youths presenting psychiatric emergencies (Henggeler, Rowland, et al., 1999; Schoenwald, Ward, Henggeler, & Rowland, 2000; Henggeler et al., 2003; Huey et al., 2004; Sheidow et al., 2004). Follow-up periods across the five articles ranged from post-treatment to 12-months post-treatment. When multiple follow-up periods were

presented either in one study, or across a series of studies, one time period was chosen in order to keep study participants independent of one another. In this way, follow-up periods at or closest to 12-months were selected, as 12-months was the follow-up period presented in the majority of included studies.

In summary, the ultimate goal was to include as many studies as possible across as many outcomes as possible without violating independence. This was accomplished by taking three steps: (1) dependent variables were separated into nine overall categories and nine meta-analyses were conducted accordingly; (2) individual effect size estimates were computed and the mean of those estimates was calculated for studies that measured treatment effects on an outcome using multiple assessments and official records measures; and (3) follow-up periods at or closest to 12-months were selected for studies that presented multiple follow-up periods across dependent variables. With this information in mind, the next subsections briefly describe the dependent variables included in the current study.

Delinquency

The primary outcome in the current study was delinquency, ultimately defined as any delinquent act or illegal behavior occurring post-treatment (i.e., after MST or comparison group treatment/services concluded). Both self-reported delinquency and official reports of delinquency were included. In addition, all types of delinquency were included (i.e., violent, non-violent, and overall offenses). Individual effect sizes were calculated and then the average of these estimates was taken to represent the effectiveness of MST in reducing delinquent behavior for a particular study when the study measured delinquency in a variety of ways.[1]

[1] It should be noted that when studies presented data for each separate offense (e.g., violent and non-violent offenses or sexual and non-sexual offenses), as well as presented summary data related to any offense, only the summary data was included for analyses.

Problem Behavior

Several other dependent variables were examined in the current study, including juvenile problem behavior. Consistent with the individual studies included for meta-analysis, the problem behavior measure was kept separate from the delinquency outcome discussed above. Studies that examined problem behaviors specifically assessed the effect of MST on youths' externalizing behaviors, such as aggression, hyperactivity, and conduct problems. Although these behaviors are considered antisocial, they are not necessarily considered delinquent or criminal behaviors; thus, two separate categories were created.

The majority of studies used the Child Behavior Checklist (CBCL), which measures behavioral functioning of children (Achenbach, 1991). Oftentimes, studies included parent/caregiver, teacher, and youth ratings on the CBCL for two broadband behavior problem scales: (1) externalizing (e.g., acting out, aggression, oppositional defiant or conduct disorder) and (2) internalizing (e.g., depression, withdrawal, anxiety) (Swenson et al., 2010). For the purpose of the problem behavior measure, only the CBCL-externalizing behavior scale was included (the internalizing scale is included in the following section). If the study included parent/caregiver, teacher, and youth reports on the externalizing scale, individual effect size estimates were computed for each report at the follow-up period and a mean of the effect sizes for the study was calculated. Other studies that did not use the CBCL to measure problem behavior, used parent/caregiver reports on the Revised Behavior Problem Checklist (RBPC) (Quay & Peterson, 1987). Again, effect size estimates from both reports were computed and the mean effect size was calculated for the study.

Psychopathology and Mental Health

Treatment effects related to youths' psychopathology and mental health symptoms were also examined. A variety of assessment instruments were employed as pre-test/post-test measures to examine whether changes in youths' behaviors occurred. Assessment tools included, for example, the internalizing subscale of the CBCL, the Global Severity Index-Brief Symptom Inventory (GSI-BSI) (Derogatis, 1993), the mood and emotion sub-scale on the Child and Adolescent Functional Assessment Scale (CAFAS) (Hodges & Wong, 1995), and the problem severity sub-scale on the Child and Adolescent-Texas Recommended Assessment Guidelines (CA-TRAG) (Texas Department of MHMR, 2004). Consistent with the discussion above, effect size estimates were calculated individually and then the average effect size for the estimates was computed for studies that presented parent/caregiver, teacher, and youth ratings on an assessment and/or when studies measured psychopathology and mental health symptoms from multiple assessment scales (e.g., CBCL-internalizing subscale and the GSI-BSI).

Family Functioning and Family Relationships

Some studies also examined the impact MST had on family functioning and relationships. Similar to the dependent variable describe above, a variety of pre-test/post-test assessments were used to measure whether any changes in this area occurred between the youths and their family members. Assessment instruments included, for example, the Family Adaptability and Cohesion Evaluation Scale (FACES-III) (Olson, Portner, & Lavee, 1985), the Family Assessment Measure (FAM-III) (Skinner, Steinhauer, & Santa-Barbara, 1983), and the Unrevealed Differences Questionnaire-Revised (URDR) (Blaske, Mann, & Henggeler, 1989). In instances where both the parent/caregiver and youth ratings were reported and/or more than one

assessment instrument was employed (e.g., FACES and the FAM-III), effect size estimates were computed for each report and the mean of the effect sizes was calculated.

Peer Relationships

The effect that MST had on youths' peer relationships was also examined in several studies. A variety of pre-test/post-test assessments were used prior to the treatment period and then again after the treatment period ended to examine this relationship. For example, studies utilized the Missouri Peer Relations Inventory (MPRI) (Borduin, Blaske, Cone, Mann, & Hazelrigg, 1989), the social sub-scale from the CBCL (Achenbach, 1991), and the Social Competence with Peers Questionnaire (SCPQ) (Spence, 1995). Consistent with the outcome measures described above, individual effect sizes were computed for each person reporting on the behavior (i.e., parent/caregiver and youth), as well as for each type of assessment used to measure the behavior and the mean effect estimate was used to represent the study's treatment effect on peer relationships

School Performance

School performance was measured in a variety of ways across independent studies, including parent/teacher reports of grades, as well as official records related to the youth's school attendance, grades, and placement (e.g., general education setting versus alternative school placement). When a study included parent/caregiver and teacher reports and/or a combination of self-report assessments and official records information, individual effect size estimates were computed for each type of report and the mean of the effect sizes for the study was computed.

Parent Functioning and Parent Relationships

Outcomes related to parent functioning and relationships were also examined in a number of studies. Pre-test/post-test assessments used to measure this relationship included, for example, the Symptom Checklist (SCL-90) (Derogatis, Lipman, & Covi, 1973), the GSI-BSI (Derogatis, 1993), and the Family Inventory of Life Events and Changes (FILE) (McCubbin, Patterson, & Wilson, 1985). In some instances, a composite score of the GSI-BSI and SCL-90 were reported. Notably, the mean of individual effect sizes was estimated if the study measured parent functioning and relationships in a multitude of ways.

Substance Abuse

Substance abuse was also examined in a variety of studies to determine whether MST reduced or eliminated youths' illicit drug use and/or alcohol consumption. The majority of studies measured substance abuse by presenting biological measures (i.e., drug test results). Other studies used pre-test/post-test assessment instruments, such as the Personal Experiences Inventory (PEI) (Winters & Henly, 1989), the drugs sub-scale of the CAFAS (Hodges &Wong, 1996), the Alcohol Use Disorder Identification Test (AUDIT) (Babor, de la Fluente, Saunders, & Grant, 1992), and the Drug Use Disorder Identification Test (DUDIT) (Berman, Bergman, Palmstierna, & Schlyter, 2005). When both biological and self-report assessment measures were reported, individual effect size estimates were calculated and the average effect size representing substance abuse for the study was computed.

Service Utilization

Finally, several studies examined the impact that MST had on youth's service utilization (e.g., hospitalization; outpatient or inpatient mental health services). Both dichotomous (e.g., "is

the youth currently living at home—yes or no?") and continuous (e.g., number of days in hospitalization) measures, as well as self-report assessments and official documentation were used to assess this relationship across included studies. Once again, a mean effect size was calculated for studies that employed various types of assessments and records information, which consequently represented the average service utilization effect for each corresponding study.

INDEPENDENT VARIABLE

As discussed in Chapter 1, the goal of the current dissertation is to advance the extant literature on the effectiveness of MST and determine more in depth, what factors or circumstances lead to the strongest treatment effects for the program. Consistent with the inclusion criteria outlined above, studies were collected for analysis that compared outcomes for MST participants and their families to participants and their families who received some type of comparison group treatment and/or service. In this way, the independent variable for the current study is whether youths and their family members participated in MST treatment services.

MODERATING VARIABLES

Although MST was the main predictor examined in this study, other moderating variables were also examined to determine their impact on the outcomes discussed above. Moderating characteristics were broken down into five general categories: (1) publication characteristics, (2) study context characteristics, (3) sample characteristics, (4) treatment characteristics, and (5) study characteristics. The following subsections explore these moderating variables more in depth.

Publication Characteristics

Various publication characteristics were coded in the current meta-analysis. For these measures, the publication itself was the unit of analysis. For example, publication type (dissertation, journal, state or federal report), publication decade, and the first author's affiliation (university, state/local agency, federal agency or other) and discipline (criminal justice/criminology, psychology, sociology, social work, or other) were coded.

Study Context Characteristics

Several study context characteristics were also coded that were related to the MST project itself. For example, the agency from which the evaluators received funds to conduct the project was coded (unfunded, agency funded, state funded, federally funded, or other), as well as the location in which the project took place (United States or outside of the United States), and the decade data collection for the project began. Whether the evaluators on the project and whether the MST developers were involved in the study in any way were also coded.

Sample Characteristics

In addition to the characteristics highlighted above, several demographic characteristics related to MST and comparison group participants were coded. First, the target population, or the type of youths examined in each study, were coded and grouped into three distinct categories (1) conduct problems (e.g., serious juvenile offenders, youths with serious conduct problems, and youths at-risk for delinquency), (2) psychological and mental health (e.g., youths presenting serious emotional disturbances or mental health issues, substance-abusing or –dependent, maltreating families, and dual-diagnoses), and (3) sexual offenders. Second, the mean age for the sample was coded as a continuous variable. Third, frequency and percent of males and

females was coded for the overall sample. Fourth, race was coded for the overall sample– frequency and percent Black, White, Hispanic, and other, as well as the frequency and percent White and non-White, Fifth, the frequency and percent of youths who had at least one prior arrest, as well as at least one prior incarceration was coded.

Measures related to youths' risk levels were also coded. The sixth measure in this category coded, for example, youths' overall risk level (low, moderate, high, or not reported), while the seventh measure noted how those risk levels were measured (using a valid psychometric instrument, using a recidivism percentage, or not reported) and if there was a psychometric instrument used, the name of the tool was indicated.

Treatment Characteristics

Characteristics related to treatment from each study were also coded. First, the source from which participants were recruited for MST and comparison group treatment/services was coded (non-criminal justice agency, such as a mental health provider, criminal justice agency, or mixed). Second, treatment completion rates for both the MST and comparison groups were coded, as well as attrition rates for both groups (i.e., percent of MST and comparison group participants who dropped out of the treatment/services prior to completion). Third, the type of treatment and/or services comparison group participants received was coded (no treatment/services, individual counseling/therapy, non-criminal justice services, traditional probation, traditional probation plus individual or group therapy, family or group counseling, mixed, or other). Fourth, whether fidelity to the MST treatment model, as well as the average adherence score/level was coded. Specifically, this study coded for whether adherence to the MST model was measured that went beyond the normal QA/QI process discussed in the previous chapter that is required of all provider organizations (e.g., audiotapes with feedback, use of the

TAM instrument). Fifth, the number of study sites (single or multisite) and the study setting (efficacy trial versus effectiveness study) was coded. Sixth, the average number of treatment/service days and hours for both MST and comparison groups was coded. Seventh, how youths were selected into treatment (through a randomized controlled study or through a quasi-experimental study) was coded. Finally, the total sample size, along with the number of participants in each group (MST and comparison) was coded.

Study Characteristics

The last set of variables examined specific characteristics related to each study's methodology, including the follow-up period employed, source of data, direction of treatment effects, and the type of analysis used in the study to calculate treatment results and subsequently, the type of analysis used to compute the effect size for the outcome of interest. First, as discussed previously, follow-up periods at or closest to 12-months were selected for all studies if the study utilized multiple follow-up times. In addition, time at-risk for many studies examining juvenile delinquency was different from the follow-up time period employed for other outcomes of interest. For example, a study might have examined recidivism for juveniles 12-months post-treatment, but examined their performance in school immediately post-treatment. In this way, the codebook included a line for a delinquency follow-up period and a line for an other outcome measures follow-up period. Second, whether the measures employed were examined via self-report, official records, observations, or a variety of sources (i.e., mixed) was coded. Third, the direction of the treatment effect was coded. More specifically, whether MST youths did better than comparison group youths or vice versa, or if the treatment favored neither group (i.e., they were equal) was coded. Finally, whether the authors calculated frequencies/percentages, means/standard deviations, used a variety of methods, or used a different type of analysis to

examine treatment effects was coded. The results of these analyses were subsequently used to calculate each study's effect size on the outcome of interest.

ANALYSES

In order to determine the impact that MST had on each dependent variable, frequencies and percentages were calculated for each moderating variable described in the previous section to better understand included studies' characteristics at the aggregate level. From here a series of more complex analyses were conducted with assistance from Borenstein and colleagues' Comprehensive Meta-Analysis (CMA) software (Borenstein, Hedges, Higgins, & Rothstein, 2005). First, analyses to calculate effect size estimates were conducted, along with several other corresponding analyses, including Fisher's *r* to *z* transformation, inverse variance, mean effect size, confidence interval around the mean effect size, and the *Q*-statistic. These analyses were performed for the independent variable (MST) and each dependent variable of interest (juvenile delinquency, problem behavior, psychopathology and mental health, family functioning and relationships, peer relationships, school performance, parent functioning and relationships, substance abuse, and service utilization) for each study. Second, the binomial effect size display (BESD)–a measure of the magnitude of the effect size–was also computed. Third, the fail-safe N statistic was calculated to determine how many studies were needed in order to reduce the findings to a trivial level (Smith, Cullen, & Latessa, 2009). Finally, when applicable, mean effect size estimates and corresponding confidence intervals were calculated for each of the moderating variable categories. These analyses are described further in the subsections below.

Effect Size Estimates

As discussed previously, one advantage of conducting a meta-analysis is that it allows the researcher to take the statistics from each included study and transform them into one common statistic, called the effect size. The effect size not only helps the meta-analyst understand the magnitude of the effect, but also the direction of the relationship under review (Bonta, Law, & Hanson, 1998).

Although there are several types of effect size estimates (e.g., mean differences, logged odds ratio), the current dissertation used Pearson's *r* standard correlation coefficient as the effect size estimate. Pearson's *r* was selected for two reasons. First, it can be easily interpreted because the value of *r* is always between +1 and -1 (Rosenthal, 1984). Second, various formulas exist that allow the researcher to easily convert other test statistics (e.g., t, F, and chi-square) into an *r* value (Bonta et al., 1998; Lipsey & Wilson, 2001; Rosenthal, 1984). Once an effect size was calculated for each study included in the current meta-analysis, several related analyses were conducted, including Fisher's *r* to *z* transformation, the inverse variance weight, the mean effect size, the confidence interval around the mean effect size estimate, and the *Q*-statistic.

Fisher's r to z Transformation. First, each effect size (*r*) was converted into a standardized score, as *r* is not a normal distribution. Standardization was done by transforming each *r* value into a z_r score using Fisher's *r* to *z* transformation. With this transformation, the sampling distribution was able to approach normality (Wolf, 1986).

The Inverse Variance Weight. Second, recall that traditional synthesis methods (i.e., narrative and ballot box reviews) give equal weight to all studies included in the evaluation and do not take into consideration the size of the sample (Hunt, 1997). Meta-analytic techniques, on the other hand, allow the researcher to place more emphasis on studies with larger sample sizes

by weighing them more heavily, compared to studies with smaller sample sizes (Hunt, 1997; Hunter and Schmidt, 1990). More weight can be given to larger samples because they tend to have a minimal amount of sampling error, and thus, produce more reliable results. In line with this discussion, the current dissertation weighted each study by sample size using the formula suggested by Rosenthal (1984). The formula is calculated by taking the product of z_r and the inverse of its variance (n-3 [n=sample size]), from each study (Lipsey & Wilson, 2001).

The Mean Effect Size. After each effect size was calculated, converted to z_r, and weighted for the studies included in the meta-analysis, a weighted mean effect size was calculated. This analysis combines the results of all of the studies by taking the average of all of the effect sizes from the individual studies (Hunt, 1997; Lipsey & Wilson, 2001) and is calculated by taking the sum of each individual weighted effect size and dividing it by the sum of the inverse of the variance for each weighted effect size (Lipsey & Wilson, 2001).

Confidence Intervals. Next, a 95 percent confidence interval (CI) around the mean effect size was calculated to assess the precision and magnitude of each point estimate (Smith et al., 2009). The CI, more specifically, provides the reviewer with a range of values around the mean effect size for the population parameter (μ) under review (Cumming & Finch, 2005). According to Smith and her colleagues (2009, p. 191), "the utility of the CI lies in its interpretability: If the interval does not contain 0, then it can be concluded that the mean effect size is significantly different from 0 (i.e., better than chance alone)." Related, if the reviewer is assessing the relationship between two independent mean effect size estimates, the relationship is considered statistically significant if the CIs do not overlap or if there is a gap between estimates' respective intervals (Cumming & Finch, 2005).

Finally, CIs are useful in that the precision of each effect size estimate can be interpreted by a quick review of the CI's width (i.e., the margin of error or *w*) (Cumming & Finch, 2005). In particular, "as the width of the CI increases, the precision of the estimate of μ decreases (or is associated with more uncertainty)" (Smith et al., 2009, p. 191). Scholars, such as Snook et al., argue that CIs with a width greater than 0.10 should be considered imprecise (Snook, Eastwood, Gendreau, Goggin, & Cullen, 2007). Ultimately, the larger the width of the CI, the less confidence the reviewer may have in his or her results and the more likely replication of those results will be warranted.

The Q-statistic. Finally, the *Q*-statistic was computed for each of the dependent variables to identify if outliers were problematic for the studies included under each outcome of interest (Rosenthal, 1991). Specifically, the *Q*-statistic determines the homogeneity of the effect sizes and was computed for each effect size using the following formula:

$$Q_{ES} = \frac{(k-3)}{(z_r - z^+)^2}$$

where k is the number of effect sizes per measure, z_r is the standardized Pearson r correlation coefficient, and z^+ is the weighted mean effect size per measure. Once this was calculated for each effect size, all the Q_{ES} values were summed resulting in Q_{OBT} and compared with a critical value of χ^2 with k–1 degrees of freedom.

If the distribution was considered heterogeneous, outliers were identified using two methods. First, values that were greater than two standard deviations away from the mean were removed. Second, if there were large gaps between subsequent values in the distribution (i.e., estimates were discontinuous across the distribution), these estimates were removed (Durlak & Lipsey, 1991). Once outliers were removed, the mean effect size, weighted mean effect size, and

corresponding confidence intervals, were recalculated. Results are presented in Chapter 3 with both outliers included and outliers removed.

Binomial Effect Size Display (BESD)

The BESD was calculated to aid in interpreting the results of the current meta-analysis (Rosenthal, 1991). Specifically, the BESD represents the difference between the treatment and comparison groups on the outcome of interest. For example, in the current study, the BESD might show the difference in rates of recidivism between the group who received MST treatment and the group who received comparison group treatment/services. Before the analysis is computed, it is necessary to convert Z(*r*) back to *r*. The BESD is then calculated as follows:

$$\text{BESD} = 0.50 \pm \frac{r}{2}$$

For instance, if $r = 0.10$, favoring the treatment group (i.e., MST), the recidivism rate for the MST group would be 45 percent (0.50 - 0.10/2) and the recidivism rate for the comparison group would be 55 percent (0.50 + 0.10/2). Even further, the BESD shows that there are 10 percentage points separating the comparison group from the MST group (Lipsey & Wilson, 2001). Because the BESD is easy to interpret and can be calculated fairly efficiently, it was computed for each effect size in the current study.[2]

Fail-Safe N Statistic

As discussed previously, it can be difficult to gather all of the studies on a particular topic, especially if that topic is highly researched and/or several unpublished studies exist. Undoubtedly, the failure to secure any studies—published or unpublished—can have significant ramifications for researchers conducting a meta-analysis. Not securing unpublished studies,

[2] It was not possible to assess whether the distributions for each dependent variable had extreme base rates (i.e., rates substantially greater than or less than 50%), which is an assumption of the BESD. As such, it was infeasible to determine whether 50 percent was an appropriate base rate and/or if an adjustment statistic should have been employed (e.g., Ley statistic).

however, could potentially be more problematic, as research has shown that published research tends to favor significant results, while unpublished research tends to find nonsignificant or null results and subsequently, "get pushed in a file drawer" (Glass, McGaw, & Smith, 1981).

To assess the impact of publication bias on a meta-analysis, several approaches can be taken (Sutton, 2009). One of the simplest methods—the fail-safe *N* statistic—was first proposed by Roesenthal (1979) and "considers the question of how many new studies averaging a null result are required to bring the overall treatment effect to non-significance" (Sutton, 2009, p. 442). The larger the fail-safe *N*, the more confident the reviewer can be in his or her findings. Conversely, the smaller the fail-safe *N*, the more cautious he or she needs to be when interpreting the results (Hunter & Schmidt, 1990; Lipsey & Wilson, 2001).

Meta-analysts have noted that there are two major problems associated with Rosenthal's fail-safe *N* formula. First, it is based on the question of statistical significance, rather than substantive significance. That is, it takes into account the number of studies that would be needed in order to make the reviewer's findings statistically nonsignificant, rather than taking into account how many studies are needed to reduce the effect to a trivial level or a level of substantive unimportance (Smith et al., 2009; Borenstein et al., 2009). Second, it assumes that the mean effect size in the missing studies is zero, not taking into consideration that it could be negative (and therefore fewer studies would be needed to deem the relationship nonsignificant) or could just be a very small positive effect (Borenstein et al., 2009).

To overcome the limitations described above, Orwin (1983) proposed an alternative fail-safe *N* method. In Orwin's fail-safe model, more specifically, the reviewer can determine how many missing studies would be needed to bring the overall effect size reported to a predetermined level or below that predetermined level (i.e., trivial level) (Smith et al., 2009).

Orwin's fail-safe *N* formula also allows the researcher to specify the effect size estimate in the missing studies as some value other than zero (Borenstein et al., 2009). His formula is as follows (Orwin, 1983):

$$N_{fs} = \frac{N_o(\bar{d}_o - \bar{d}_c)}{d_c - \bar{d}_{fs}}$$

where N_o is the number of studies, d_o is the mean effect size calculated, d_c is the predetermined or trivial effect size, and d_{fs} is the mean effect size of the missing studies. This formula computes N_{fs}, which is the number of additional studies needed to obtain the desired effect size (d_c).

The fail-safe *N* formula presented above is based on effect size estimates calculated as Cohen's *d*. The formula was adapted by Lipsey and Wilson (2001), however, so that it could be used with other effect size estimates, such as Pearson's *r* (i.e., the estimate used in the current meta-analysis. This formula is as follows (Lipsey & Wilson, 2001):

$$k_o = k\left[\frac{\overline{ES_k}}{\overline{ES_c}} - 1\right]$$

where k_o corresponds to the number of studies needed to reduce the mean effect size for the meta-analysis to an alternative or criterion effect size level ($\overline{ES_c}$), k is the number of studies used to calculate the weighted mean effect size, $\overline{ES_k}$ is the weighted mean effect size, and $\overline{ES_c}$ is the criterion effect size level. In line with Hedges and Olkin's (1985) recommendation, the criterion effect size level set forth in the current study was 0.01; a value that would be considered negligible in the criminal justice and corrections fields.

Moderating Variables

Mean effect sizes and their respective confidence intervals were calculated across each moderating variable category described in the previous section (publication characteristics, study context characteristics, sample characteristics, treatment characteristics, and study characteristics). Each confidence interval was examined to determine whether it overlapped with any of the other confidence intervals. If it did overlap, it was assumed there were no significant moderating effects. On the other hand, if confidence intervals did not overlap, it was determined that there was a moderating effect. It is instructive to note that many moderating variables were unable to be calculated because a considerable amount of data were missing for the analyses being performed. This will be addressed further in Chapters 3 and 4.

CONCLUSION

Because meta-analysis is quantitative in nature, it overcomes many of the obstacles associated with traditional research synthesis methods (i.e., narrative and ballot box reviews). Specifically, meta-analysis allows the reviewer to systematically investigate the relationship between two variables using statistical techniques, making it a more objective approach to synthesizing the literature on a particular topic. In addition, because of the technique's objectivity, meta-analyses can be replicated by other researchers. Finally, and perhaps most importantly, meta-analysis can take hundreds of studies and yield one statistic that provides a lot of information about a particular topic, including the direction and magnitude of the effect size.

Meta-analysis is not without its critics; however, the present chapter has attempted to describe how each of these points of criticism were overcome. For example, efforts were made to find both published and unpublished studies. Related, the fail-safe *N* statistic was calculated for each mean effect size related to the various outcome variables examined in the study in order

to address the potential "file drawer" problem. In addition, the "apples to oranges" problem was addressed by examining various moderating variables related to publication characteristics, study context characteristics, sample characteristics, treatment characteristics, and study characteristics.

Ultimately, this dissertation seeks to examine the effectiveness of MST on nine dependent variables, including: (1) delinquency, (2) problem behavior, (3) psychopathology and mental health, (4) family functioning and relationships, (5) peer relationships; (6) school performance; (7) parent functioning and relationships, (8) substance abuse, and (9) service utilization. In addition, the impact of various moderating variables is examined to determine whether the main effects of the independent variable and dependent variables varied by particular characteristics. In short, the current study attempts to contribute to a better understanding of MST's effectiveness by including more studies than previous MST meta-analyses, as well as by examining the impact of various moderating variables.

CHAPTER 3

RESULTS

The previous chapters described the development of MST, took stock of the current empirical evidence surrounding the program, and described the research methods used in the present study. The current chapter presents the results of the study in three sections. First, descriptive statistics are presented for a variety of publication, study context, sample, treatment, and study characteristics. Second, mean effect sizes are presented for the nine outcomes described in the previous chapter (juvenile delinquency, problem behavior, psychopathology and mental health, family functioning and relationships, peer relationships, school performance, parent functioning and relationships, substance abuse, and service utilization). Finally, the impact of the aforementioned publication, study context, sample, treatment, and study characteristics for all outcomes together and then for delinquency and psychopathology and mental health outcomes separately are examined (i.e., moderating effects).

As previously discussed, an extensive literature search was conducted to collect all studies examining the effectiveness of MST. Over 100 studies were uncovered during this search process. For a variety of reasons, however, a number of studies were deemed ineligible during coding. Studies were excluded, for example, because they did not include a control or comparison group, they did not examine the effectiveness of MST on one of the nine dependent variables of interest, they did not provide appropriate data to calculate an effect size, or they utilized a previously included sample to examine outcomes with a longer follow-up period. In all, 44 studies, contributing 127 effect sizes were included in the current meta-analysis. Of the 127 effect sizes:

- 26 (19.0%) examined the effectiveness of MST on juvenile delinquency,

- 19 (14.0%) examined the effectiveness of MST on problem behavior,
- 18 (13.0%) examined the effectiveness of MST on psychopathology and mental health,
- 14 (10.0%) examined the effectiveness of MST on family functioning and relationships,
- 9 (7.0%) examined the effectiveness of MST on peer relationships,
- 9 (7.0%) examined the effectiveness of MST on school performance,
- 14 (10.0%) examined the effectiveness of MST on parent's functioning and relationships,
- 8 (6.0%) examined the effectiveness of MST on substance abuse, and
- 10 (8.0%) examined the effectiveness of MST on out-of-home placement.

PUBLICATION, STUDY CONTEXT, SAMPLE, TREATMENT, AND STUDY CHARACTERISTICS

The following subsections present various characteristics of the studies that were deemed eligible for inclusion in the current meta-analysis. Again, characteristics separated into five general categories: (1) publication characteristics, (2) study context characteristics, (3) sample characteristics, (4) treatment characteristics, and (5) study characteristics. It should be noted that nine of the included studies utilized non-unique samples. These studies were included in the current project however, because they each measured the effectiveness of MST on different outcomes of interest. For example, Henggeler et al. (1999), Brown et al. (1999), and Henggeler et al. (2002) all utilized the same data set for their publications; however, Henggeler et al., (1999) examined the effectiveness of MST in changing delinquent and substance abusing behaviors, as well as lowering youths' risk of out-of-home placement. Brown et al. (1999), on the other hand, examined MST's ability to improve juveniles' school performance, while Henggeler et al. (2002) examined MST's ability to improve youths' problem behavior, as well as their psychiatric and mental health functioning. Thus, in the subsections to follow, it will be specified whether analyses were based on: (1) 44 non-unique studies, (2) 35 unique datasets, *or* (3) 127 effect sizes.

Publication Characteristics

Table 3.1 presents descriptive statistics related to studies' publication characteristics. Since the characteristics presented in this subsection focus on features of the publications themselves, analyses were based on the 44 non-unique studies deemed eligible for inclusion. As can be seen from the table, almost 80 percent of the studies were published in journals, while the remaining studies were non-published dissertations (approximately 11%), or state or federal reports (approximately 9%). Regardless of the type of publication, almost all of the studies were

Table 3.1
Descriptive Statistics: Publication Characteristics (N = 44)

Publication Characteristic	N	%
Publication Type		
Dissertation	5	11.4
Journal	35	79.5
State or Federal Report	4	9.1
Publication Decade		
1980 or 1990	8	18.2
2000 or 2010	36	81.8
Primary Author Affiliation		
University	39	88.6
State or Federal Agency	5	11.4
Primary Author Discipline		
Criminal Justice/Criminology	4	9.1
Psychology	31	70.5
Social Work	2	4.6
Other	4	9.1
Missing	3	6.8

written in more recent decades–approximately 82 percent in 2000 or 2010. The remaining studies were written in the 1980s or 1990s (approximately 18%). Finally, out of the 44 studies, over half were written by a primary author who was affiliated with a university (approximately 89%) and trained in psychology (approximately 71%).

Study Context Characteristics

Table 3.2 presents descriptive statistics related to the study context. The analyses for this subsection were based on the 35 unique datasets deemed eligible for the current study because the variables that were coded were directly related to the circumstances in which the project took place. As can be seen from the table below, funding for each project came from the state in which the study was conducted (approximately 37%), a federal agency (approximately 30%), or the researchers received a mix of both state and federal funding for the project (approximately 5%). The table also demonstrates that slightly more than 77 percent of the studies were conducted in the U.S. while the remaining, almost 23 percent, were conducted outside of the U.S. Regarding data collection, it appears that the start of data collection processes began for most projects in the early 2000s (approximately 49%), followed by the late 1900s (approximately 40%).

Some evaluators, in addition to assessing MST's effectiveness, oversaw MST clinicians during the project's duration and/or provided technical assistance and quality assurance monitoring to the agency responsible for providing MST treatment. In this way, Table 3.2 examines whether the scholar(s) who wrote the article was also involved in the implementation of MST (in addition to being the project evaluator[s]). As can be seen, the majority of evaluators *were* involved in the study in an additional capacity (almost 57.1%).

Table 3.2
Descriptive Statistics: Study Context Characteristics (N = 35)

	N	%
Funding Agent		
State Funded	13	37.1
Federally Funded	10	28.6
Mixed	5	14.3
Missing	7	20.0
Geographic Location		
Outside of the United States	8	22.9
United States	27	77.1
Decade Data Collection Started		
1980 or 1990	14	40.0
2000 or 2010	17	48.6
Missing	4	11.4
Involvement of the Evaluator		
No	15	42.9
Yes	20	57.1
Involvement of MST Developers		
No	21	60
Yes	14	40.0

Related, Table 3.2 demonstrates that more than half of the studies (60.0%) were conducted independently from the MST developers (i.e., Henggeler and his colleagues who originally created MST), while the remaining studies were completed by Henggeler and/or other developers of the program (40.0%).

Sample Characteristics

Table 3.3 presents descriptive statistics related to sample characteristics. Because these analyses focused on characteristics of the youths participating in MST and comparison

treatment/services, the analyses for this subsection were based on the 35 unique datasets deemed eligible for the current study. It should be noted upfront that several variables related to sample characteristics were coded, but due to missing data, could not be included for analysis (please refer to the codebook in Appendix G to see all of the variables that were coded). As can be seen from the table, the effectiveness of MST was examined for various types of youths. Specifically, 60.0 percent of the studies examined MST's effectiveness for conduct problem youth, 31.4 percent of the studies examined the program's effectiveness for youths presenting mental health-related problems, and a small number of studies (8.6%) examined MST's effectiveness for sexual offenders.

Basic demographic characteristics (age, sex, and race,) for youths who participated in MST and comparison group treatment/services across all included studies are also presented in Table 3.3. In line with van der Stouwe et al.'s (2014) meta-analysis of MST, age was collapsed into two meaningful categories–under 15 years of age and 15 years of age or older. As can be seen, the majority of the study samples yielded average ages of 15 or older (approximately 63%), while a smaller number of studies yielded average ages under the age of 15 (approximately 37%). In addition, in over half of the studies, the majority (70% or more) of the participants were male, while participants in a slightly smaller percentage of studies (45.7%) included both males and females. One study (2.9%) examined MST's effectiveness for an all female participant sample. Related to race, almost half (approximately 46%) of the study samples consisted of a mix of White and non-White participants, while another half of the samples consisted of mainly White or mainly non-White participants (approximately 26% and 23%, respectively).

Table 3.3
Descriptive Statistics: Sample Characteristics (N = 35)

	N	%
Type of Youth		
Conduct Problems	21	60,0
Mental Health	11	31.4
Sex Offenders	3	8.6
Average Age ($\bar{x}$ = 14.4)		
Under 15	13	37.1
15 and Older	22	62.9
Sex		
Exclusively Females	1	2.9
Mainly Males (over 70.0%)	18	51.4
Mixed	16	45.7
Race		
White (over 70.0%)	9	25.7
Non-White (over 70.0%)	8	22.9
Mix	16	45.7
Missing	2	5.7
History of Arrest (over 70.0%)		
No/Cannot Tell	15	42.9
Yes	12	34.3
Mixed	8	22.9
Measurement of Risk		
Valid Psychometric Instrument	4	11.4
Author Reported Risk Level/Score	3	8.6
Coder Defined (based on criminal history)	13	37.1
Cannot Tell /Not Reported	15	42.9
Risk Level		
Low	0	0.0
Moderate	1	2.9
High	16	45.7
Mix	2	5.7
Cannot Tell/Not Reported	16	45.7

Finally, Table 3.3 provides information related to participants' history of arrest, level of risk for reoffending, and measurement of those risk levels. As can be seen, the preponderance of studies included samples where the majority of youths had never had a previous arrest or arrest rates could not be determined (42.9%). A smaller number of studies included participants who mainly all had a history of arrest (34.3%), followed by studies where arrest histories were mixed (22.9%). Several studies did not report levels of risk for their samples or levels of risk could not be determined based on the information provided in the study (45.7%). When risk levels were reported or could be determined based on the information provided, 45.7 percent were deemed high-risk juveniles, 2.9 percent were deemed moderate-risk juveniles, and 5.7 percent were a mix of low-, moderate-, and high-risk participants. Of the studies where risk levels could be determined, 11.4 percent utilized a valid psychometric instrument (e.g., Risk Needs Assessment [RNA], Positive Achievement Change Tool [PACT]), 8.6 percent did not indicate whether they used a valid psychometric tool, but reported participants' overall risk level to reoffend, and the remaining 37.1 percent of risk levels were defined by the coder based on youths' criminal history information.

Treatment Characteristics

Table 3.4 presents descriptive statistics related to treatment characteristics. Similar to the sample characteristics described above, analyses for this subsection were based on the 35 unique datasets deemed eligible for the current study. As can be seen from the table below, the majority of youths participating in the included studies were referred to treatment or services (MST or comparison group treatment/services) through a criminal justice agency (62.9%). Slightly more than a quarter of these studies (25.7%) included youths who were referred for treatment/services through a non-criminal justice agency (e.g., mental health provider, child and

family services, local hospital). Finally, 11.4 percent of the studies included participants referred for treatment/services through a mix of both criminal justice and non-criminal justice entities.

Related to treatment completion rates for MST participants, Table 3.4 demonstrates that 60.0 percent of the studies yielded completion rates at or above 80.0 percent. The remaining studies either did not report completion rates (14.3%) or yielded a mix of completers versus non-completers (25.7%). When referring to comparison group completion rates, the table demonstrates that 80.0 percent or more of comparison group youths did not complete treatment/services in 5.7 percent of the studies. Almost half (48.6%) of the studies, on the other hand, yielded completion rates at or above 80.0 percent for comparison group youths. The remaining studies either did not report completion rates (28.6%) or yielded a mix of completers and non-completers (17.1%).

Table 3.4 also presents statistics related to the number of treatment days MST youths and comparison group youths received, respectively. As can be seen, youths in 43.9 percent of the studies received between 100 and 200 days of MST treatment, followed by youths in 8.6 percent of the studies who received between 201 and 300 days of MST treatment, and youths in 5.7 percent of the studies who received less than 100 days of MST treatment. Related to the number of treatment/service days comparison group youths received, Table 3.4 indicates that youths in 11.4 percent of the studies received between 100 and 200 days of treatment/services, followed by youths in 5.7% of the studies who received between 301 and 400 days of treatment/services, and youths in another 5.7 percent of the studies who received less than 100 hours of comparison group treatment/services. The majority of studies (74.2%) did not indicate how many treatment/service hours comparison group youths received, however.

While youths in the treatment group across all included studies received MST, comparison group youths received a variety of treatment and/or services during studies' durations. As can be seen from Table 3.4, 62.9 percent of the studies compared the MST treatment group to youth who received both treatment (e.g., individual counseling, group or family therapy) and traditional supervision services (e.g., probation, community service, restitution). A smaller percentage of comparison group youths received treatment (28.6%) *or* traditional supervision services (8.6%) alone.

The last four variables included in this subsection are related to therapist adherence, study sites, study design, and study settings. As can be seen in Table 3.4, it appears that therapist adherence to the MST model was measured in the majority of studies (60.0%), while a smaller percentage did not indicate that therapist adherence was measured (40.0%). Treatment/services were conducted at a single site or location for the preponderance of studies (65.7%), while a small percentage were conducted across two or more locations (34.3%). A study was coded as an efficacy trial if MST therapists were graduate students, while they were coded as effectiveness trials if master's level licensed clinicians or therapists facilitated treatment. Descriptive statistics related to study setting thus indicate that approximately 71 percent of the studies were considered effectiveness studies, while the other one-third were considered efficacy trials. Finally, a smaller percentage of studies (40.0%) were quasi-experimental in nature, while the majority were randomized controlled trials (60.0%).

Table 3.4
Descriptive Statistics: Treatment Characteristics (N = 35)

	N	%
Source of Clients for Treatment		
Non-Criminal Justice Agency	9	25.7
Criminal Justice Agency	22	62.9
Multiple Sources	4	11.4
Program Completion Status – MST		
Did Not Complete Treatment (over 80.0%)	0	0.0
Completed Treatment (over 80.0%)	21	60.0
Mixed	9	25.7
Missing	5	14.3
Program Completion Status – Comp.		
Did Not Complete Treatment (over 80.0%)	2	5.7
Completed Treatment (over 80.0%)	17	48.6
Mixed	6	17.1
Missing	10	28.6
Average Number of Treatment Days – MST		
< 100	2	5.7
100 – 200	15	42.9
201 – 300	3	8.6
Missing	15	42.9
Average Number of Treatment/Service Days – Comp.		
< 100	2	5.7
100 – 200	4	11.4
201 – 300	1	2.9
301 – 400	2	5.7
Missing	26	74.2
Comparison Group Received		
Treatment	10	28.6
Probation	3	8.6
Treatment & Probation	22	62.9

Table 3.4
Descriptive Statistics: Treatment Characteristics (N = 35)

	N	%
Adherence to MST Model		
Appeared Adherence was Measured	21	60.0
No Indication Adherence was Measured	14	40.0
Number of Study Sites		
Single-Site	23	65.7
Multi-Site	12	34.3
Study Design		
Randomized	21	60.0
Quasi-Experimental	15	40.0
Study Setting		
Efficacy Trial	10	28.6
Effectiveness Study	25	71.4

Study Characteristics

Table 3.5 presents descriptive statistics related to study characteristics. Because this table presents data related to effect size characteristics, analyses were based on the 127 unique effect sizes presented in the beginning of this section. As discussed in the Methods section, several studies employed different follow-up periods for juvenile delinquency and other outcomes of interest. In this way, follow-up time is presented separately for delinquency measures and the other eight outcome variables included for analysis. It is also important to note that all follow-up periods were standardized to reflect post-treatment follow-up periods.

Recall that 26 effect sizes measured the effectiveness of MST in reducing juvenile delinquent behavior. Of these 26 effect sizes, 65.4 percent employed a follow-up period between 0-months (i.e., post-treatment) and 12-months, 15.4 percent employed a follow-up period

between 13- and 24-months, 19.2 percent employed a follow-up period greater than or equal to 25-months. For the remaining 101 effect sizes, 83.2 percent employed a follow-up period between 0- and 12-months, 14.9 percent employed a follow-up period between 13- and 24-months, and 2.0 percent employed a follow-up period between 25- and 48-months.

The next three sets of analyses examined the type of information gathered to calculate effect sizes, the direction of those effect sizes, and the type of data scholar(s) used to measure program effectiveness. First, the type of analyses refers to the statistical procedures the scholar(s) used to present his/her study results (and subsequently, the data that were extracted for the current study to calculate effect size estimates). As can be seen, the preponderance of effect sizes (57.5%) were calculated using the mean and standard deviation. Studies were coded as "mixed" if the scholar(s) used several different types of analyses to examine outcome measures and/or used different assessment/official records measures to assess a particular outcome. As can be seen from the table, almost 30.0 percent fell into the "mixed" category, indicating that effect sizes were calculated based on frequencies and percentages, means and standard deviations, and/or another type of statistical method. A smaller number of effect sizes (7.9%) were computed using frequencies and percent estimates only, while an even smaller percentage (4.7%) used some other type of statistical analysis (e.g., multivariate analysis of variance and F-test).

In the current study, positive effect sizes signify that outcome was associated with treatment effects in favor of the MST group. That is, positive effects indicate that MST was associated with a greater reduction in delinquency, problem behavior, substance abuse and/or greater improvements in psychopathology and mental health, family functioning and relationships school performance, peer relationships, parent functioning and relationships, and/or

Table 3.5
Descriptive Statistics: Study Characteristics (k = 127)

	k	%
Average Follow-Up Period – Delinquency (k = 26)		
0–12 Months	17	65.4
13–24 Months	4	15.4
≥ 25 Months	5	19.2
Average Follow-Up Period – All Others (k = 101)		
0–12 Months	84	83.2
13–24 Months	15	14.9
25–48 Months	2	2.0
Type of Analyses		
Frequency/Percent	10	7.9
Mean/Standard Deviation	73	57.5
Other	6	4.7
Mixed	38	29.9
Direction of Effect Size		
Favored Treatment Group	90	70.9
Favored Comparison	32	25.2
Favored Neither	5	3.9
Source of Data		
Self-Report	42	33.8
Official Records	10	7.4
Mixed	75	58.8

service utilization, compared to comparison group treatment/services. On the other hand, negative effect sizes signify that outcome was associated with treatment effects in favor of the comparison group. Effect size estimates that fell right at zero indicate that MST treatment and comparison group treatment/services did no better or worse than one another (i.e., the effects were neutral). In line with this discussion, Table 3.5 indicates that 90 effect size estimates

(70.9%) favored MST treatment, while 32 effect size estimates (25.2%) favored comparison group treatment. A small number (3.9%) of estimates fell at 0.00, indicating that MST treatment and comparison group treatment and/or services did no better or worse than one another.

Finally, Table 3.5 presents data related to the source of the data scholar(s) utilized to measure treatment outcomes. As can be seen, the majority of studies (58.8%) measured outcomes based on both self-report and official records (i.e., mixed). A smaller percentage of studies (33.8%) measured outcomes based on self-report measures, followed by an even smaller percentage of studies (7.4%) that measured outcomes based on official records.

MEAN EFFECT SIZE ESTIMATES

The results of the meta-analysis by each outcome measure of interest are presented in the current section. For each measure, the number of effect sizes included in the mean calculation (*k*) is presented, along with the total number of participants included in the samples used to calculate the respective mean effect size estimates (N). The mean effect size (*r*), weighted mean effect size (Z^+), and corresponding 95 percent confidence intervals are also presented. In line with Snook and his colleagues (2007), CIs with a width greater than 0.10 were deemed less precise, compared to CIs with a width of 0.10 or less. As such, less precise estimates should be interpreted with caution until further lines of research are able to replicate (or not) the findings. All of the analyses identified herein were calculated for both fixed effects and random effects models.

Briefly, in a fixed effects model, it is assumed that there is only one true effect size for all of the studies included in the analysis, and "that all differences in observed effects are due to sampling error" (Borenstein et al., 2009, p. 83). A random effects model, on the other hand, assumes that the true effect size varies from study to study. For example, the point estimate

might be higher for studies where treatment participants were older, or for studies where the majority of participants were male. Ultimately, the random effects model takes into account that each study is likely to yield a different effect size because they "differ in the mixes of participants and in the implementation of interventions" (Borenstein et al., 2009, p. 83). It is important to note that due to its calculations, CIs in a fixed effects model are narrower, compared to CIs in a random effects model. When a fixed effects model is used in a meta-analysis then, it is often considered inappropriate to generalize the results "to a universe of studies that could not be located and studies that have not been conducted" (Gendreau & Smith, 2007, p. 1551); a generalization that can be claimed when using a random effects model.

Three additional sets of analyses were also computed for each outcome of interest. First, to interpret the findings in a more simplistic manner, the BESD was calculated. Second, the *Q*-statistic was calculated to better understand the amount of variability across studies' effect size estimates. Recall from the Methods chapter that when the *Q*-statistic is significant, it demonstrates that the sample of studies included in the model is heterogeneous and that further analyses should be conducted to examine the influence of moderating factors. Before moderators are explored however, it is important to investigate potential outlying effect sizes on the distribution. As such, fixed effects and random effects models were re-calculated to examine relationships without outliers and discontinuous point estimates. Third, the fail-safe *N* statistic was computed to account for missing studies by calculating the number of studies that would need to be included in the current meta-analysis in order to reduce the effect size to 0.01.

Before the results for each outcome are presented, the effect of MST on outcomes in general is evaluated for all dependent variables combined, including: (1) delinquency, (2) problem behavior, (3) psychopathology and mental health, (4) family functioning and

relationships (5) peer relationships, (6) school performance, (7) parent functioning and relationships, (8) substance abuse, and (9) service utilization. As a caveat, this methodological technique certainly violates the "rule of statistical independence" that, as previously discussed, many scholars discourage against. This approach was taken, however, because due to a small number of total effect sizes and missing data in some subsequent analyses, several moderators did not reach statistical significance (that might have otherwise if cell frequencies were not limited). By comparing findings from the small subset of studies to those yielded from all of the studies, however, general trends or patterns could be gleaned.

All Outcomes

Table 3.6 presents the effects of MST when all of the outcomes of interest are combined (N =19,343 non-unique youths; $k = 127$). For ease of interpretation, the word outcome is used to represent outcome generically. The random effects model below indicates that MST significantly improved outcome. More specifically, the average effect size and weighted mean effect size were both 0.13 (SD = 0.20) and the 95 percent confidence interval around both r and Z^{+} fell between 0.10 and 0.17. Employing the BESD, differences between MST and comparison groups can be better demonstrated. For youths participating in MST treatment, more specifically, outcome improved, on average, by 56.5 percent, compared to 43.5 percent, on average, for youths participating in comparison group treatment/services.

Table 3.6 also reveals that the Q-statistic was significant for the relationship between MST and outcome ($Q = 569.65$, $p = 0.00$). Although the Q value and corresponding p value suggest that particular factors may moderate the relationship between treatment and outcome (and therefore, explain some of the variability in effect sizes), it is important to examine this

Table 3.6
Mean Effect Size Estimates—All Outcomes

Model	*k*	N	*r*	SD	95% CI	Z^+	95% CI	*Q*
Fixed Effects	127	19,343	0.10	0.20	0.08–0.12	0.10	0.08–0.12	569.65**
With outliers removed	120	18,618	0.08	0.17	0.07–0.10	0.07	0.07–0.10	293.08**
Random Effects	127	19,343	0.13	0.20	0.10–0.17	0.13	0.10–0.17	569.65**
With outliers removed	120	18,618	0.11	0.17	0.08–0.14	0.11	0.08–0.14	293.08**

**$p < 0.001$, *$p < 0.05$

relationship without outlying effect sizes or effect sizes that are discontinuous on the distribution. In this vein, outliers that were greater than two standard deviations away from the mean effect estimate and/or were discontinuous on the distribution were identified. This resulted in the removal of seven point estimates.

Table 3.6 also presents the mean effect size estimates with the seven outliers identified above removed from the distribution. As can be seen, the magnitude of the point estimates decreased slightly (r and Z^{+} = 0.11, SD = 0.17) and the confidence interval around r and Z^{+} also narrowed in width (CI_r and CI_{Z+} = 0.08-0.14). Importantly, the relationship between MST and outcome remained significant, as the CI did not include zero.

The fail-safe N statistic was also computed to determine the number of studies needed in order to reduce the mean effect size for MST and outcome to 0.01. The results from this analysis revealed that almost 900 studies (N = 896) would need to be included in order to reduce the estimate to 0.01 when the seven outlying estimates were removed. This number increased to more than 1,000 studies (N = 1,140) when outliers were kept in the distribution. Note that normally we can have faith in these results with this large amount of studies; however, recall that the effect sizes included for analysis were not independent of one another and as such, may be biased.

Delinquency

Table 3.7 examines MST's effectiveness in reducing rates of delinquency for more than 7,000 juveniles who participated in MST treatment and comparison group treatment/services (k = 26). When referring to the random effects model, the table indicates that MST significantly reduced youths' delinquent behavior, with an average effect size of 0.11 (SD = 0.17) and a 95 percent confidence interval between 0.05 and 0.16. Notably, the weighted mean effect size and

corresponding confidence interval were almost identical to the unweighted mean effect size estimate and associated CI ($Z^+ = 0.11$, $CI_{Z+} = 0.05–0.17$). These findings can also be interpreted using the BESD statistic. Assuming a base rate of 50 percent, for example, the *r* value of 0.11 suggests that the success rate of MST represented an 11 percent improvement over the comparison group condition (or 55.5% versus 44.5%) (Gendreau & Smith, 2007). While independent study effect sizes ranged from -0.11 to 0.49, the majority of studies (65%) were in favor of MST treatment over comparison group treatment/services (please see Appendix H for a complete list of independent study results, including author, publication year, effect size, and sample size).

Table 3.7 also reveals that the *Q*-statistic was significant for the relationship between MST and juvenile delinquency ($Q = 74.05$, $p = 0.00$). While this information is important and indicates a need for further exploration of moderating effects, it also indicates a need to examine the results with outliers removed. In this way, outliers that were greater than or less than two standard deviations from the mean effect size estimate and/or were discontinuous on the distribution were identified. This resulted in the removal of one effect size.

Table 3.7 illustrates that the magnitude of the mean effect size for the relationship between MST treatment and juvenile delinquency without the outlier identified above, is slightly smaller ($r = 0.10$) than the estimate with the outlier included; however, the width of the CI for both *r* and Z^+ were larger than the 0.10 threshold for precision suggested by Snook et al. (2007) (w_r and $w_{Z+} = 0.12$). Ultimately, the removal of the outlier did not result in substantial changes to the mean effect size estimate and corresponding 95 percent confidence interval. Similarly, *Q* remained significant even after the removal of the outlier ($Q = 71.34$, $p = 0.00$). This suggests

Table 3.7
Mean Effect Size Estimates—Delinquency

Model	*k*	N	*r*	SD	95% CI	Z^+	95% CI	*Q*
Fixed Effects	26	7,344	0.06	0.17	0.03–0.10	0.06	0.03–0.10	74.05**
With outliers removed	25	7,328	0.06	0.16	0.03–0.09	0.06	0.03–0.09	71.34**
Random Effects	26	7,344	0.11	0.17	0.05–0.16	0.11	0.05–0.17	74.05**
With outliers removed	25	7,328	0.10	0.16	0.04–0.16	0.10	0.04–0.16	71.34**

**$p < 0.001$, *$p < 0.05$

that significant heterogeneity exists among sampling distributions, which may be attributed to systematic differences in studies, rather than a product of sampling error.

Finally, because the majority of included studies in the meta-analysis were published (approximately 80%), the fail-safe *N* statistic was calculated to determine the number of studies needed in order to reduce the effect size estimate for MST and juvenile delinquency to 0.01. Using Lipsey and Wilson's (2001) formula and examining the random effects model with the outlier removed, analyses revealed that 130 additional studies would be needed in order to reduce the effect size to 0.01. When the outlying estimate was added back into the model, the fail-safe *N* value was slightly higher ($N = 139$). Regardless of whether outliers are included or not included, we can have faith in these results given that over 100 studies would be needed in order to reduce the relationship between MST and delinquency to trivial results.

Problem Behavior

Table 3.8 examines MST's effectiveness in changing problem behavior for 2,200 youths who participated in MST and comparison group treatment/services ($k = 19$). The random effects model below indicates that MST treatment significantly reduced juveniles' problem behavior. More specifically, the average effect size and weighted mean effect size were both 0.15 (SD = 0.19), with associated 95 percent confidence intervals between 0.05 and 0.24 and 0.05 and 0.25, respectively. Employing the BESD, differences in these effect sizes can be better illustrated. For youths participating in MST treatment, for example, problem behavior decreased by 57.5 percent, while it decreased by 42.5 percent for youths participating in comparison group treatment and/or services (a 15% difference). While individual study effect sizes ranged between -0.08 and 0.58, the majority of studies (79%) favored the MST group. Independent study results

Table 3.8
Mean Effect Size Estimates–Problem Behavior

Model	*k*	N	*r*	SD	95% CI	Z^{+}	95% CI	*Q*
Fixed Effects	19	2,228	0.14	0.19	0.10–0.18	0.14	0.10–0.18	98.03**
With outliers removed	17	1,959	0.07	0.13	0.03–0.12	0.07	0.03–0.13	22.90
Random Effects	19	2,228	0.15	0.19	0.05–0.24	0.15	0.05–0.25	98.03**
With outliers removed	17	1,959	0.08	0.13	0.03–0.13	0.08	0.03–0.13	22.90

**$p < 0.001$, *$p < 0.05$

(author, publication year, effect size, and sample size) related to problem behavior can be found in Appendix I.

Table 3.8 also reveals that the *Q*-statistic was significant for the relationship between MST and problem behavior ($Q = 98.03$, $p = 0.00$). While it is important that moderating effects be explored, it is also important to examine this relationship without outlying effect sizes or effect sizes that are discontinuous on the distribution. As such, outliers that were greater than or less than two standard deviations from the mean effect estimate and/or were discontinuous on the distribution were identified. This resulted in the removal of two effect sizes.

As can be seen from Table 3.8, there are considerable differences between mean effect size estimates with and without outliers ($r = 0.15$ and $r = 0.08$, respectively); however, the relationship between MST and problem behavior remained significant, as it does not include zero. It is also important to point out that the width of the confidence interval became much narrower when the two outlying estimates were removed from the distribution ($w_r = 0.19$ to $w_r = 0.10$).

The fail-safe *N* statistic was also computed for the problem behavior outcome measure to determine the number of studies needed in order to reduce the effect size estimate for MST and youth problem behavior to 0.01. The results from this analysis revealed that 109 studies would need to be included in order to reduce the estimate to 0.01 when the two outlying estimates were removed. This number more than doubled when the two outlying studies were kept in the model ($N = 243$). Regardless of whether outliers were included or not included, we can have faith in these results given that over 100 studies would be needed in order to reduce the relationship between MST and youth problem behavior to trivial results.

Psychopathology and Mental Health

Table 3.9 presents the relationship between MST treatment and youth psychopathology and mental health for almost 2,200 MST and comparison group youths ($k = 18$). The effect size estimates generated from the random effects model, more specifically, reveal that MST treatment, on average, reduced youths' psychopathological and mental health symptoms (r and $Z^+ = 0.15$, SD = 0.18). This relationship is significant, in that the 95 percent confidence interval about mean r and Z^+ does not include zero (CI_r and $CI_{Z+} = 0.07–0.23$); however, the widths of the corresponding CIs are greater than 0.10 (w_r and $w_{Z+} = 0.16$). In line with the previous findings presented above, independent study results ranged from negative results ($r = -0.08$) to positive results (r = 0.63), but the majority of the estimates favored MST treatment (83%) (please see Appendix J for a complete list of independent study results, including author, publication year, effect size, and sample size). When the BESD was computed, the psychopathology and mental health effect size estimates translated into a 15 percent improvement rate for the MST group over the comparison treatment/services group (or 57.5% versus 42.5%).

Table 3.9 also reveals that the Q-statistic was significant ($Q = 56.68$, $p = 0.00$) for the relationship between MST treatment and psychopathology and mental health, which indicates that further moderating analyses need to be explored. Before moderators are explored, however, it is important to examine the model when outliers are removed. In this case, one outlier was removed that was two standard deviations away from the mean effect size, which resulted in a slightly smaller (but significant) mean estimate ($r = 0.12$, SD = 0.15) and a narrower confidence interval ($CI_{Z+} = 0.05$—0.18) (albeit, the CI is still wider than 0.10). The Q-statistic also remained significant ($Q = 33.88$, $p = 0.01$), which may indicate that heterogeneity in effect sizes stems beyond sampling error.

Table 3.9
Mean Effect Size Estimates–Psychopathology and Mental Health

Model	*k*	N	*r*	SD	95% CI	Z^{+}	95% CI	*Q*
Fixed Effects	18	2,230	0.11	0.18	0.07–0.15	0.11	0.07–0.15	56.68**
With outliers removed	17	2,171	0.10	0.15	0.05–0.14	0.10	0.05–0.18	33.88**
Random Effects	18	2,230	0.15	0.18	0.07–0.23	0.15	0.07–0.23	56.68**
With outliers removed	17	2,171	0.12	0.15	0.05–0.18	0.12	0.06–0.18	33.88**

**$p < 0.001$, *$p < 0.05$

The last analysis related to youth psychopathology and mental health was conducted so that potential publication bias could be explored. For this analysis, the fail-safe *N* statistic was calculated to determine the number of studies needed in order to reduce the mean effect estimate to 0.01. It was found that 147 studies would need to be added to the current study in order to reduce the estimate to 0.01 when the one outlying estimate was removed. The number of studies needed in order to reduce the mean effect size to 0.01 was slightly higher (n = 185) when the outlying estimate was added back into the distribution. Regardless of whether outliers are included or not included, we can have faith in these results given that over 100 studies would be needed in order to reduce the relationship between MST and youth psychopathology and mental health to trivial results.

Family Functioning and Relationships

Table 3.10 demonstrates the relationship between treatment and family functioning and relationships for 1,492 youths who participated in MST or comparison group treatment/services (k = 14). The random effects model below indicates that family functioning and relationships improved significantly for MST youths and their families (r and Z^{+} = 0.14, SD = 0.20). In addition, the confidence intervals about the weighted and unweighted estimates imply that we can be 95 percent confident that the mean effect size is contained within the interval estimate (Snook et al., 2007; Cumming & Finch, 2005) because they do not include zero (CI_r = 0.01-0.26 and CI_{Z+} = 0.01-0.27); however, these estimates cannot be considered precise, as their widths are greater than 0.10 (w_r = 0.25 and w_{z+} = 0.26). Translating the findings into the BESD statistic, results indicate that the success rate of MST represented a 20.0 percent improvement in family dynamics (or 60% versus 40%) over comparison group family dynamics.

Notably, effect size estimates varied considerably across individual studies (r_{min} = -0.10

Table 3.10
Mean Effect Size Estimates–Family Functioning and Relationships

Model	*k*	N	*r*	SD	95% CI	Z^{+}	95% CI	*Q*
Fixed Effects	14	1,517	0.12	0.20	0.07–0.17	0.12	0.07–0.18	75.43**
With outliers removed	13	1,424	0.6	0.13	0.00–0.11	0.06	0.00–0.11	17.13
Random Effects	14	1,517	0.14	0.20	0.01–0.26	0.14	0.01–0.27	75.43**
With outliers removed	13	1,424	0.07	0.13	0.00–0.13	0.07	0.00–0.14	17.13

**$p < 0.001$, *$p < 0.05$

and $r_{max} = 0.66$) for the relationship between MST and family functioning and relationships; however, well over half of studies' estimates favored MST (64%) (please see Appendix K for a complete list of independent study results, including author, publication year, effect size, and sample size). Also related to individual study results, the *Q*-statistic ($Q = 75.43$, $p = 0.00$) presented in the table indicates that significant heterogeneity across independent effect sizes exists. Moderating analyses are explored further in the section to follow; however, the significant value of *Q* warrants that the model be examined without outliers first. The random effects model with outliers removed suggests that one study was removed that yielded an effect size that was more than two standard deviations away from the mean effect size estimate.

When the outlying estimate was removed, the mean effect size decreased by half of its original value ($r = 0.14$ to $r = 0.07$) and the confidence interval became substantially narrower ($CI_r = 0.00$-0.13 and $CI_{Z+} = 0.00$-0.14). Note, however, that the width of the CIs for both weighted and unweighted estimates is greater than 0.10 and that the lower boundary of the CI is 0.00, indicating that the findings may not be any different than what we may expect from chance alone.

Peer Relationships

Table 3.11 presents the relationship between MST and peer relationships for over 1,200 youths who participated in MST and comparison group treatment/services ($k = 9$). As can be seen from the random effects model, weighted and unweighted effect size estimates were both 0.09 (SD = 0.16; BESD of 54.5% versus 45.5%) and associated confidence intervals were between 0.00 and 0.17. Note that the lower boundary of both CIs about mean *r* and Z^+ were 0.00, suggesting that the point estimates were not significantly different from 0 and the widths of their associated CIs were substantially larger than the 0.10 threshold suggested by Snook et al. (2007).

Table 3.11
Mean Effect Size Estimates–Peer Relationships

Model	*k*	N	*r*	SD	95% CI	Z^+	95% CI	*Q*
Fixed Effects	9	1,227	0.08	0.16	0.02–0.14	0.08	0.02–0.14	15.98*
With outliers removed	8	1,179	0.06	0.06	0.00–0.12	0.06	0.00–0.12	3.35
Random Effects	9	1,227	0.09	0.16	0.00–0.17	0.09	0.00–0.17	15.98*
With outliers removed	8	1,179	0.06	0.06	0.00–0.12	0.06	0.00–0.12	3.35

**$p < 0.001$, *$p < 0.05$

(2007) to be deemed precise. These findings, however, are in the right direction (i.e., favor MST) and an attempt should be made to replicate the results with additional effect sizes.

Importantly, seven out of nine effect sizes (78%) examining the effectiveness of MST for improving peer relationships for youths yielded point estimates in favor of MST (please see Appendix L for a complete list of independent study results, including author, publication year, effect size, and sample size). The lack of statistical significance then, is likely due to the small number of studies included for analysis on this dependent variable.

Because the *Q*-statistic was statistically significant ($Q = 15.98$, $p = 0.04$), outliers that were two standard deviations away from the mean effect size and estimates that were discontinuous on the distribution were identified. The random effects model with outliers removed indicates that only one study was removed based on these criteria and that the relationship between treatment and peer dynamics remained nonsignificant ($r = 0.06$, SD = 0.06), with the lower limit of the CI falling at 0.00 once again and the upper limit falling at 0.12.

The fail-safe *N* statistic was calculated in order to determine the number of studies needed in order to reduce the mean effect size for peers to 0.01. It was found that a minimal number of studies ($N = 38$) would need to be added to the current study in order to reduce the estimate to 0.01 when the one outlying estimate was removed. The number of studies needed in order to reduce the mean effect size to 0.01 was slightly higher ($N = 64$) when the outlying estimate was added back into the distribution. Note that these fail-safe values are much smaller than those yielded for the main effects examined in prior subsections.

School Performance

The effect that MST treatment had on youths' performance in school is presented in Table 3.12 for almost 960 MST and comparison group youths ($k = 9$). Examining the results

associated with the random effects model, mean *r* and Z^+ were 0.16 (SD = 0.35) and their associated CIs were -0.13 to 0.42 and -0.13 to 0.45, respectively. Note that the 95 percent confidence intervals about the unweighted and weighted estimates included 0.00 and the width of these limits were five times greater than the suggested width for declaring estimates precise (w_r = 0.55 and w_{z+} = 0.58). Nonetheless, the magnitude and direction of the effect sizes, are worth noting.

Briefly, while the results above may not have reached statistical significance, 7 out of the 9 effect sizes (78%) generated independent effect sizes in favor of MST for improving juveniles' performance in school over comparison group treatment/services (r_{min} = -0.59 and r_{max} = 0.57). When translating these findings into the BESD statistic, the mean effect size of 0.16 suggests that the success rate of MST represented a 16 percent improvement over the comparison group condition (or 58% versus 42%). The lack of statistical significance then, is likely due to the small number of studies included for analysis on this dependent variable. Please see Appendix M for a complete list of individual study results, including author, publication year, effect size, and sample size.

With the rather large difference between the lowest and highest estimates on the scale, it is not surprising that the *Q*-statistic reached statistical significance (Q = 131.56, p = 0.00). As such, further analyses were conducted to examine the relationship between treatment and youths' school performance after one outlying point estimate was removed from the distribution. The estimate that was removed was discontinuous and negatively associated with youths' performance in school. After the study was removed, findings revealed that youths who participated in MST treatment significantly improved their performance in school, compared to comparison group youths (*r* and Z^+ = 0.26, SD = 0.25), yielding a confidence interval about

Table 3.12
Mean Effect Sizes With and Without Outliers Removed–School Performance

Model	*k*	N	*r*	SD	95% CI	Z^{+}	95% CI	*Q*
Fixed Effects	9	979	0.09	0.35	0.02–0.16	0.09	0.02–0.16	131.56**
With outliers removed	8	823	0.23	0.22	0.16-0.30	0.23	0.16–0.31	43.96**
Random Effects	9	979	0.16	0.35	-0.13–0.42	0.16	-0.13–0.45	131.56**
With outliers removed	8	823	0.26	0.22	0.07–0.43	0.26	0.07–0.46	43.96**

**$p < 0.001$, *$p < 0.05$

mean *r* between 0.07 and 0.43 and about Z^+ between 0.07 and 0.46. While margins of error around the unweighted and weighted estimates without the outlier were narrower than those associated with the distribution including the outlier, the width of the CIs were still almost four times greater than the 0.10 precision threshold ($w_r = 0.36$ and $w_{z+} = 0.39$). Importantly, the *Q*-statistic remained significant after the one outlier was removed ($Q = 43.96$, $p = 0.00$), suggesting that there are likely some moderating effects at play that are influencing study results. Finally, the fail-safe *N* test revealed that 179 additional studies would need to be uncovered and added to the current study in order to bring the results without the outlier down to a trivial level. Alternatively, a much smaller number of studies ($N = 70$) would need to be uncovered and added to the current study if the outlier estimate was kept for inclusion.

Parent Functioning and Relationships

Study results examining the effect of treatment on parent functioning and relationships for approximately 1,600 MST and comparison group families is presented in Table 3.13 ($k = 14$). The random effects model indicates that the relationship is significantly different from zero, yielding a mean effect size of 0.15 (SD = 0.22) and a confidence interval between 0.05 and 0.25. When the point estimates were weighted by sample size, the findings were nearly identical ($Z^+ = 0.15$, $CI_{Z+} = 0.05\text{-}0.26$). Note that the width of the CIs associated with both mean *r* and mean Z^+ were greater than 0.10 ($w_r = 0.20$ and $w_{z+} = 0.21$). To illustrate these results further, the BESD was computed, which yielded a 15 percent improvement rate in parent functioning and relationships for MST participants, compared to comparison group participants (or 57.5% versus 42.5%).

Independent study effect sizes for the relationship between MST and parent relationships and functioning ranged between -0.14 and 0.50, with nine out of the 14 estimates (64%) favoring

Table 3.13
Me Mean Effect Size Estimates–Parent Functioning and Relationships

Model	*k*	N	*r*	SD	95% CI	Z^{+}	95% CI	*Q*
Fixed Effects	14	1,659	0.12	0.22	0.07–0.17	0.12	0.07–0.17	51.31**
With outliers removed	13	1,575	0.13	0.20	0.08–0.18	0.13	0.08–0.18	47.47**
Random Effects	14	1,659	0.15	0.22	0.05–0.25	0.15	0.05–0.26	51.31**
With outliers removed	13	1,575	0.17	0.20	0.07–0.27	0.17	0.07–0.28	47.47**

**$p < 0.001$, *$p < 0.05$

MST. A list of individual study results (author, publication year, effect size, and sample size) can be found in Appendix N. Because the *Q*-statistic was statistically significant ($Q = 51.31$, $p = 0.00$), the distribution was examined further to determine whether there were any outliers that were two standard deviations away from the mean effect size and/or estimates that were discontinuous. From this investigation, one discontinuous effect size was removed. The resulting model is presented in the table above under the random effects model without outliers.

As can be seen, the relationship between MST treatment and parent functioning and relationships remained significant, yielding marginally larger unweighted and weighted average effect size estimates (*r* and $Z^+ = 0.17$, SD = 0.20) and CIs about these estimates between 0.07 and 0.27 and 0.07 and 0.28, respectively. Note that the widths of the CIs were still greater than the 0.10 threshold considered for precise estimates ($w_r = 0.20$ and $w_{z+} = 0.21$). Similar to the above, the *Q*-statistic is significant ($Q = 47.47$, $p = 0.00$), suggesting that moderating effects may be influencing the effect between MST treatment and parent functioning and relationships. Finally, the fail-safe *N* test indicated that 152 studies would need to be added to the current meta-analysis in order to bring the mean effect size to 0.01 when the discontinuous estimate was removed from the model. A slightly smaller number of studies ($N = 150$) would need to be added to the model if the discontinuous estimate was not removed.

Substance Abuse

The effectiveness of treatment in reducing substance-abusing behaviors for almost 970 MST and comparison group youths ($k = 8$) is presented in Table 3.14 below. From the random effects model below, it can be seen that the relationship between MST participation and drug use did not significantly differ from zero (*r* and $Z^+ = 0.04$, SD = 0.23) (BESD of 52% versus 48%). The width of the confidence intervals about mean *r* and Z^+ were considerably wide (w_r and $w_{z+} =$

Table 3.14
Mean Effect Size Estimates–Substance Abuse

Model	*k*	N	*r*	SD	95% CI	Z^{+}	95% CI	*Q*
Fixed Effects	8	991	0.02	0.23	-0.05–0.09	0.02	-0.05–0.09	26.25**
With outliers removed	7	936	0.04	0.36	-0.03–0.11	0.04	-0.03–0.11	21.39**
Random Effects	8	991	0.04	0.23	-0.10–0.17	0.04	-0.10–0.17	26.25**
With outliers removed	7	936	0.07	0.36	-0.06–0.20	0.07	-0.06–0.21	21.39**

**$p < 0.001$, *$p < 0.05$

0.27) and also included zero (CI_r and CI_{Z+} = -0.10-0.17), which validates that MST youths did not do significantly better than comparison group youths in changing their drug and alcohol consumption.

Unlike the results discussed thus far, the majority of individual study effect sizes favored comparison group treatment/services (63%) (please see Appendix O for a complete list of study results, including author, publication year, mean effect size, and sample size). While the *Q*-statistic was small ($Q = 26.25$, $p = 0.00$), it was significant, which indicates that moderating effects still need to be explored for this relationship. The distribution was examined first, however, to identify whether any outlying studies existed that yielded effect size estimates that were two standard deviations away from the mean estimate and/or discontinuous on the distribution. This resulted in the removal of one discontinuous effect size.

When the random effects model with outliers removed is examined, it can be seen that the average unweighted and weighted effect sizes increased slightly (r and $Z^+ = 0.07$, SD = 0.36), but the relationship remained nonsignificant, as evidenced by the confidence interval constructed around the mean estimates including 0.00 (CI_r and CI_{Z+} = -0.06–0.20). Precision of these estimates was also not considered, as the width of the CIs was more than two times greater than the 0.10 threshold suggested by Snook et al. (2007). Interestingly, the *Q*-statistic remained significant after the discontinuous estimate was removed ($Q = 21.39$, $p = 0.00$).

Service Utilization

Table 3.15 examines the impact that MST treatment had on youths' service utilization (e.g., hospitalization, outpatient, or inpatient mental health services) (N = 1,094; $k = 10$). The random effects model below indicates that MST youths' rates of service utilization improved (r

Table 3.15
Mean Effect Size Estimates–Service Utilization

Model	*k*	N	*r*	SD	95% CI	Z^+	95% CI	*Q*
Fixed Effects	10	1,168	0.19	0.19	0.11–0.26	0.19	0.11–0.26	20.47*
With outliers removed	9	916	0.23	0.17	0.15–0.31	0.23	0.15–0.32	11.25
Random Effects	10	1,168	0.21	0.19	0.09–0.33	0.21	0.09–0.32	20.47*
With outliers removed	9	916	0.24	0.17	0.14–0.34	0.25	0.14–0.35	11.25

**$p < 0.001$, *$p < 0.05$

and $Z^+ = 0.21$, SD = 0.19). Employing the BESD, the findings translate to a 21 percent improvement rate for MST youths over comparison group youths for the service utilization outcome (or 60.5% versus 39.5%). Importantly, the widths of the CIs associated with both mean *r* and mean Z^+ were quite large ($w_r = 0.24$ and $w_{Z+} = 0.25$, respectively), but did not include 0.00, suggesting that the relationship between MST treatment and service utilization is significant (albeit, imprecise).

Individual effect sizes ranged between -0.14 and 0.52; however, 90 percent of the studies yielded effect size estimates that were in favor of MST (please see Appendix P for a complete list of study results, including author, publication year, effect size, and sample size). A significant *Q* value ($Q = 20.47$, $p = 0.02$) also indicated that outliers and discontinuous estimates needed to be examined further, which uncovered only one discontinuous estimate. While the mean effect size remained significant and increased to 0.24 (SD = 0.17) when the study was removed from the analysis, the width of the confidence interval increased slightly falling between 0.14 and 0.34. Similarly, the weighted average effect size estimate and corresponding CI increased ($Z^+ = 0.25$, $CI_{Z+} = 0.14$-0.35). Finally, the fail-safe *N* test indicated that more than 175 studies would need to be added in order to deem the relationship between MST treatment and service utilization trivial for both the model with outliers and the model without outliers ($N = 178$ and $N = 201$ studies, respectively).

MODERATING EFFECTS

As can be gathered from the previous section, several main effect relationships yielded results that warrant a further examination of moderating factors. In this way, the current section explores the impact of moderating effects on MST treatment for all outcomes together and then explores moderating effects for MST and delinquency and MST and psychopathology and

mental health outcomes separately.[3] Consistent with the previous section presenting studies' descriptive statistics, moderating effects in this section are presented for five categories, including: (1) publication characteristics, (2) study context characteristics, (3) sample characteristics, (4) treatment characteristics, and (5) study characteristics. It is important to note that once effect sizes were separated by their respective outcome measures and further separated by moderator subgroups, much missing data were yielded. When possible, subgroups were collapsed to allow more cases to fall into each subcategory; however, even after subcategories were collapsed, some variables could not be included for analysis.

For each moderator variable, the number of effect sizes included in the mean calculation for each subgroup (*k*) is presented, along with the total number of participants included in those subgroups used to calculate the respective mean effect size estimates (N). The mean effect size (*r*), weighted mean effect size (Z^+), and corresponding 95 percent confidence intervals are also presented. The CIs, were computed for each subcategory of a variable, not only to determine if the point estimates were significantly different from zero, but also to determine whether subgroups were significantly different from each other. Overlapping CIs indicate that subcategories were not significantly different from one another, and thus no real moderating effects existed among the point estimates (Durlak & Lipsey, 1991). Non-overlapping CIs, on the other hand, indicate that subcategories were significantly different from one another and subsequently, that a significant moderating effect existed between point estimates. In addition, the BESD statistic was calculated in order to understand the findings in a more parsimonious manner. All of the analyses were calculated for the random effects model.

[3] Cell frequencies were too small to examine moderating effects for the other dependent variables of interest.

Moderating Effects—All Outcomes

Publication and Study Context Characteristics. Heterogeneity in effect sizes may be attributed to characteristics related to the publications themselves and/or the context in which the studies were conducted. Table 3.16 reports the mean effect size estimates for publication and study context characteristics for MST and the outcomes of interest in the current meta-analysis (k = 127). More specifically, enough data were available to compute moderating effects for two publication characteristics: (1) publication type and (2) publication decade and for four study context characteristics: (1) geographic location, (2) decade data collection started, (3) involvement of the evaluator, and (4) involvement of MST developers.

First, it can be seen that published studies were significantly associated with outcome and MST treatment effectiveness. More specifically, published studies yielded a mean effect size of 0.15 (SD = 0.21) and a confidence interval around the estimate of 0.11-0.18. Notably, the weighted mean effect size estimate was identical to the unweighted mean effect size estimate (Z^+ = 0.15) and the width of the corresponding confidence interval minimally increased by 0.01 (CI_{Z+} = 0.11-0.19). It should be further noted that the width of both CIs for the unweighted and weighted mean effect sizes were less than the 0.10 threshold for precision suggested by Snook et al. (2007). As such, we can be 95 percent confident that the mean effect size of 0.15 is contained within the confidence interval estimate (Cumming & Finch, 2005; Snook et al., 2007).

To translate the findings above more simply, the BESD was computed. Assuming a base rate of 50 percent, for example, the r value of 0.15 means that for youths participating in MST treatment, outcome improved, on average, by 57.5 percent, compared to 42.5 percent for youths participating in comparison group treatment/services when examining published studies collectively. While the mean effect size yielded for non-published studies appears to be

minimally associated with outcome and treatment effectiveness (r = 0.04, SD = 0.11), the CI around this estimate contained 0.00, indicating that the relationship was not significantly different from what might be expected from chance alone (CI_r = -0.02-0.09). Importantly, the confidence intervals around the estimates for published versus non-published studies did not overlap with one another, indicating that the larger effect size estimate generated from published studies was significantly different from the effect size estimate generated from non-published studies.

Second, when examining the findings related to publication decade, Table 3.16 indicates that studies written in the late 1900s and early 2000s were significantly associated with outcome and MST treatment effectiveness. For studies written in 1980 or 1990, in particular, both the unweighted mean effect size and weighted mean effect size were 0.15 (SD = 0.20). This relationship was statistically significant, as the CIs around the mean effect and weighted mean effect sizes did not contain 0.00 (CI_r = 0.08-0.22 and CI_{Z+} = 0.08-0.23, respectively). Similar to the discussion above, the r value of 0.15 indicates that outcome improved, on average for the MST group by 57.5 percent, compared to the rate of 42.5 percent for comparison group youths when examining studies published in 1980 or 1990 collectively. The width of the CIs for the unweighted mean effect size and weighted mean effect size, however, were larger than the 0.10 threshold for precision (w_r = 0.14 and w_{z+} = 0.15). As such, results should be considered tentative until future studies are able to replicate (or not) these findings (Snook et al., 2007).

For studies written in the early portion of the 2000s, the mean effect and weighted mean effect sizes were just slightly lower than the estimates yielded for studies written in the later portion of the 1900s (r and Z^+ = 0.13, SD = 0.20). The estimates were both statistically significant from 0.00, however, generating identical confidence intervals between 0.09 and 0.17.

With this relatively narrow CI width (w_r and w_{z+} = 0.08), we can conclude that the precision of the estimate was satisfactory (Gendreau & Smith, 2007). When interpreting the findings using the BESD, the results indicate that outcome, on average, improved by 56.5 percent for the MST group and 43.5 percent for the comparison group for studies written in 2000 or 2010 collectively.

While both publication decade subgroups were statistically significant from zero, they were not significantly different from each other, as evidenced by overlapping CIs around associated unweighted mean effect size and weighted mean effect size estimates. The findings suggest, however, that MST was slightly more effective than comparison group treatment/services in studies where decade of publication was 1980 or 1990.

Third, the table also presents information related to where the study took place. As can be seen, studies that were conducted in the United States were significantly associated with outcome and MST treatment effectiveness. Specifically, the mean effect size was 0.15 (SD = 0.20) and the 95 percent confidence limit was between 0.11 and 0.19; an interval that did not include 0.00. The mean effect size remained at 0.15 after studies were weighted by sample size and a similar confidence interval was produced, increasing minimally in width by 0.01 (CI_{Z+} = 0.11-0.20). Notably, the width of the CI for the mean effect size and weighted mean effect size estimates were under the 0.10 threshold for precise estimates (w_r = 0.08 and w_{Z+} = 0.09, respectively). In this way, we can be 95 percent confident that the point estimate is contained within the population parameter (Cumming & Finch, 2005).

Table 3.16 also reveals that studies conducted outside of the United States were significantly associated with outcome overall and MST treatment effectiveness; however, the point estimates were almost two times smaller than the estimates computed for studies conducted

Table 3.16
Mean Effect Sizes: Publication and Study Context Characteristics–All Outcomes

	k	N	*r*	SD	95% CI	Z^+	95% CI
Publication Type							
Unpublished	14	2,916	0.04	0.11	-0.02–0.09	0.04	-0.02–0.09
Published	113	16,427	0.15	0.21	0.11–0.18	0.15	0.11–0.19
Publication Decade							
1980 or 1990	32	3,702	0.15	0.20	0.08–0.22	0.15	0.08–0.23
2000 or 2010	95	15,641	0.13	0.20	0.09–0.17	0.13	0.09–0.17
Geographic Location							
U.S.	97	14,880	0.15	0.20	0.11–0.19	0.15	0.11–0.20
Outside the U.S.	30	4,463	0.08	0.19	0.01–0.14	0.08	0.01–0.14
Decade Data Collection Started							
1980 or 1990	56	6,208	0.22	0.21	0.16–0.28	0.22	0.16–0.29
2000 or 2010	58	11,506	0.06	0.17	0.02–0.11	0.06	0.02–0.11
Involvement of Evaluator							
No	42	9,731	0.04	0.18	-0.01–0.10	0.04	-0.01–0.10
Yes	85	9,612	0.18	0.20	0.13–0.22	0.18	0.13–0.23
Involvement of MST Developers							
No	59	12,289	0.10	0.20	0.05–0.15	0.10	0.05–0.15
Yes	68	7,054	0.16	0.20	0.12–0.21	0.17	0.12–0.21

in the U.S. (r = 0.08, SD = 0.19, CI_r = 0.01-0.14). Note that the weighted mean effect size and corresponding CI were both identical to the unweighted estimates (Z^+ = 0.08, CI_{Z+} = 0.01-0.14). We cannot be 95 percent confident about the results, however, because the width of the CI is more than the 0.10 threshold identified by Snook and his colleagues (2007) for being precise.

When examining the results between subcategories, the findings suggest that the effectiveness of MST was not significantly moderated by geographic location. It appears, however, that this relationship was approaching significance because the CIs overlapped minimally from the lower bound of one interval (U.S.) to the upper bound of the other (outside the U.S.). Recall, in addition, that the point estimate generated from studies conducted in the U.S. was almost two times larger than the point estimate generated from studies conducted outside of the U.S. In this way, MST may have been more effective when conducted in the United States.

Fourth, Table 3.16 indicates that studies where data collection began in 1980 or 1990 and 2000 or 2010 were significantly associated with outcome and MST treatment effectiveness. For studies where data collection began in the later portion of the 1900s, in particular, both the unweighted and weighted mean effect size estimates were 0.22 (SD = 0.21), with confidence intervals between 0.16-0.28 and 0.16-0.29, respectively. When employing the BESD, these findings suggest that outcome improved, on average by 61 percent for MST youths, compared to 39 percent for comparison group youths when examining studies where data collection began in 1980 or 1990. Notably, both CIs for the mean effect and weighted mean effect sizes were slightly larger than the 0.10 threshold for precision (w_r = 0.12 and w_{z+} = 0.13), suggesting that the results should be regarded as tentative (Snook et al., 2007).

For studies where data collection began in the early portion of the 2000s, the mean effect size was almost four times smaller than the estimate generated from studies where data collection began in the later portion of the 1900s (r = 0.06, SD = 0.17). The CI, however, did not contain 0.00, suggesting that studies where data collection began in 2000 or 2010 were significantly associated with outcome and MST treatment effectiveness (CI_r = 0.02-0.11). It should be noted that the weighted mean effect size estimate and corresponding CI were identical to the unweighted point estimate and CI (Z^+ = 0.06, CI_{Z+} = 0.02-0.11). In addition, the width of both CIs corresponding to the unweighted and weighted mean effect sizes was 0.09, indicating that these effect sizes were precise estimates of the true population parameters. When employing the BESD, the results imply that MST improved outcome, on average, by 53 percent, compared to 47 percent for the comparison group when examining studies where data collection took place in 2000 or 2010.

When comparing the findings for each sub-category, the findings suggest that the larger mean effect size generated from studies where data collection began in 1980 or 1990 was significantly different from the mean effect size generated from studies where data collection began in 2000 or 2010. This is evident by the lack of overlap between the subgroups' associated confidence limits.

Fifth, Table 3.16 suggests that MST treatment effects were greater when the scholar(s) who wrote the article was involved in the implementation of MST (in addition to being the project evaluator[s]) (r = 0.18, SD = 0.20). This relationship, even further, was significant because the 95 percent confidence interval did not contain 0.00 (CI_r = 0.13-0.22). When examining the weighted mean effect size and corresponding CI, it can be seen that the point estimate was identical to the unweighted estimate (Z^+ = 0.18) and the CI was just slightly wider

(CI_{Z+} = 0.13-0.23). The widths of both CIs, however, were below the 0.10 precision threshold and can be considered precise (w_r = 0.09 and w_{Z+} = 0.10). The use of the BESD helps to better illustrate the difference between MST and comparison group youths when examining studies where the evaluator was involved. In particular, the 0.18 mean point estimate suggests that the success rate of MST represents an 18 percent improvement in outcome over the comparison group condition (or 59& versus 41%) when examining evaluator-involved studies collectively.

For studies where the evaluator(s) who wrote the article was not involved in the implementation of MST in some additional capacity, the average effect size and identical weighted mean effect size were considerably lower (r and Z^+ = 0.04, SD = 0.18) than the estimate produced for evaluator-involved studies. The confidence interval around both the unweighted mean estimate and weighted mean estimate contained 0.00, suggesting that treatment effects were not significantly related to overall outcomes when the evaluator was not involved in the implementation of MST (CI_r and CI_{Z+} = -0.01-0.10). These findings can be considered precise, as the margin of error is 0.09.

Taken together, the results indicate that there was a significant difference between subgroups for the involvement of the evaluator moderator, as evidenced by the lack of overlap between the CIs corresponding to both unweighted mean effect size and weighted mean effect size estimates. Specifically, studies where the researcher(s) was involved in the implementation of MST (in addition to being the evaluator on the project) yielded an average effect size estimate that was more than four times larger than the estimate yielded from studies where the researcher(s) did not play an additional role in the study's implementation.

The final moderator examined in Table 3.16 is related to the involvement of the MST developers. As can be seen, the average effect size estimate for studies when Henggeler and/or

his team were involved in MST implementation was 0.16 (SD = 0.20). This relationship was significant, as the confidence interval around the average estimate did not include 0.00 (CI_r = 0.12-0.21). When examining the average effect size and corresponding confidence limit when studies were weighted by sample size, it can be seen that the point estimate increased slightly to 0.17, but the CI remained the same (CI_{Z+} = 0.12-0.21). The application of the BESD helps to further illustrate the difference in MST effectiveness by involvement of MST developers. On average, there was a 16 percent difference between MST and comparison groups for the outcomes considered in the analyses when the developers were involved in study implementation (or 58% versus 42%). Note that these findings can be considered precise as the width of the confidence intervals for both unweighted and weighted point estimates were relatively narrow (w_r and w_{Z+} = 0.09).

The average effect size generated from studies where the MST developers were not involved in study implementation was 0.10 (SD = 0.20) for both weighted and unweighted computations. These effect sizes were considered significantly different from zero, as their corresponding CIs did not include 0.00 (CI_r and CI_{Z+} = 0.05-0.15). When translating the point estimate into the BESD, the results indicate that the MST group, on average, yielded a 10 percent improvement rate over the comparison group for outcomes of interest when Henggeler and/or other MST developers were *not* involved in study implementation (or 55% versus 45%). Importantly, the precision of these estimates can be considered satisfactory, as the width of the CIs is right at the 0.10 threshold for precision deemed appropriate in the field.

While the two effect sizes discussed above are significantly different from zero, their respective confidence limits overlap, indicating that they are not significantly different from one another. It is important to note, however, that the relationship was approaching significance, as

the CIs overlapped minimally from the upper bound of one category (no) to the lower bound of the other (yes). As such, it appears that MST may have been slightly more effective in improving outcome when Henggeler and/or other MST developers were involved, compared to when they were not involved in study implementation.

Sample Characteristics. Heterogeneity in effect sizes could also be attributed to differences across study subjects. As such, the current subsection examines the effectiveness of MST in relation to various sample characteristics. Table 3.17 presents the mean effect size estimates related to sample characteristics for MST and all outcomes considered in the present analyses. Enough data were available to compute moderating effects for four relationships, including: (1) type of youth, (2) average age, (3) race, and (4) arrest.

First, the table below indicates that MST was effective in improving outcome for all three types of youths examined in the current meta-analysis; however, effect size estimates varied by subgroup. In particular, the mean effect size was highest for sex offenders (r = 0.28, SD = 0.17), followed by youths with conduct problems (r = 0.14, SD = 0.21), and then youths diagnosed with mental health problems (r = 0.06, SD = 0.17). The confidence intervals around each point estimate suggest that the subgroups were significantly different from zero, as they do not include 0.00 (CI_r = 0.19-0.37, CI_r = 0.09-0.19, and CI_r = 0.02-0.11, respectively). Similar results were produced when studies' sample sizes were taken into consideration, with both mental health and conduct problem subgroups yielding identical weighted mean effect size estimates and CIs as their respective unweighted counterparts (Z^+ = 0.06, CI_{Z+} = 0.02-0.11 and Z^+ = 0.14, CI_{Z+} = 0.09-0.19, respectively). The weighted mean effect size for the sex offender subcategory was just 0.01 higher than its unweighted counterpart, producing an estimate of 0.29 (CI = 0.19-0.39). Importantly, the estimates generated from studies examining the effects of MST with youths

presenting conduct and mental health problems should be considered more precise than those produced from studies examining the effects of MST with sexual offenders. This is because the width of the CI around the estimates for conduct and mental health subgroups were much narrower ($w_{Z+} = 0.09$) than the width of the CI for the sexual offender subgroup ($w_{Z+} = 0.18$).

Assuming a 50 percent base rate, the findings discussed above can be translated into the BESD statistic. More specifically, for sexual offenders who participated in MST treatment, outcome improved, on average, by 64 percent, while outcome improved, on average, by 36 percent for sexual offenders who participated in comparison group treatment/services. Related to youths presenting conduct problems, outcome improved, on average, by 57 percent for MST participants, while it improved by 43 percent for comparison group youths. Finally, youths diagnosed with mental health issues who participated in MST improved outcome, on average, by 53 percent, compared to a 47 percent improvement rate for youths diagnosed with mental health issues who participated in comparison group treatment/services

Upon taking a closer look at the table and the relationships between subcategories, it can be gathered that the larger mean effect size estimate produced for the sexual offender subgroup was significantly different from the mean effect size estimate produced for the mental health subgroup. Although the average effect size generated from studies examining treatment effectiveness for sexual offenders was two times larger than the estimate generated from studies examining treatment effectiveness for youths presenting conduct problems, this relationship only approached significance. Specifically, the upper boundary of the CI estimate for the conduct problems subgroup touched the lower boundary of the CI for the sexual offender subgroup. Notably, the CIs for the conduct problems and mental health subcategories overlap considerably and thus, were deemed not significantly different from one another. The findings do suggest,

however, that MST was more effective for sexual offenders and conduct problem-youth, compared to youths experiencing mental health and psychological problems.

Second, Table 3.17 suggests that MST was effective in improving outcome for juveniles under the age of 15, as well as for juveniles aged 15 or older, producing average effect size estimates of 0.16 (SD = 0.19) and 0.10 (SD = 0.20), respectively. These relationships were significantly different from zero, as evidenced by their respective confidence intervals not containing 0.00. In particular, the CI around the point estimate for the under 15 subgroup was 0.12 to 0.21, while the CI around the point estimate for the 15 and older subgroup was between 0.04 and 0.15. Notably, the mean effect size and corresponding CI for studies where the average age was under 15 remained the same when weighted by sample size ($Z^{+} = 0.16$, $CI_{Z+} = 0.12$-0.21). While the weighted mean effect size remained the same for studies where the average age was 15 or older ($Z^{+} = 0.10$), the width of the confidence interval increased slightly ($CI_{Z+} = 0.04$-0.16). Ultimately, with a 0.09 margin of error, the estimates produced for the under age 15 subgroup can be considered precise. As such, we can be 95 percent confident that the mean effect size of 0.16 is contained within the confidence interval estimate. With a 0.12 margin of error, on the other hand, the estimates produced for the 15 years or over subgroup should be considered tentative, as it is more than the 0.10 threshold for precision.

Employing the BESD helps to better illustrate the differences between the point estimates discussed above. For studies where the average age of the sample was under age 15, MST participants improved on outcome by 58 percent, on average, while comparison group participants improved on outcome by 42 percent. For studies where the average age of the sample was 15 years of age or older, MST participants improved on outcome by 55 percent, on average, while comparison group participants improved on outcome by 45 percent.

Table 3.17
Mean Effect Sizes: Sample Characteristics–All Effect Sizes

	k	N	*r*	SD	95% CI	Z^+	95% CI
Type of Youth							
Conduct Problems	69	13,043	0.14	0.21	0.09–0.19	0.14	0.09–0.19
Mental Health	44	5,162	0.06	0.17	0.02–0.11	0.06	0.02–0.11
Sexual Offenders	14	1,138	0.28	0.17	0.19–0.37	0.29	0.19–0.39
Average Age							
Under 15	70	8,989	0.16	0.19	0.12–0.21	0.16	0.12–0.21
15 or Older	57	10,354	0.10	0.20	0.04–0.15	0.10	0.04–0.16
Race							
White (70.0% or more)	33	3,444	0.29	0.20	0.21–0.36	0.30	0.21–0.38
Non-White (70.0% or more)	36	6,213	0.04	0.18	-0.02–0.10	0.04	-0.02–0.10
Mix	56	9,219	0.10	0.18	0.06–0.14	0.10	0.06–0.14
At Least One Arrest							
No	33	7,381	0.05	0.21	-0.01–0.12	0.05	-0.01–0.12
Yes (70.0% or more)	48	6,201	0.21	0.20	0.15–0.28	0.22	0.15–0.28

While effect size estimates for both subgroups—under age 15 and age 15 or older—were significant, their respective confidence intervals overlapped with one another, suggesting that the estimates were not significantly different from one another. It is important to note, however, that the estimate generated from the studies where participants' average age was under 15 was larger than the estimate generated from studies where participants' average age was 15 or older. As such, it appears that MST did slightly better for younger participants.

Third, Table 3.17 demonstrates that differential treatment effects were found across racial subcategories; however, in all three cases, the unweighted and weighted mean effect sizes were in the right direction (i.e., MST participants did better than comparison group participants). The mean effect size was highest for studies where the majority of sample participants were White (r = 0.29, SD = 0.20), followed by studies where the racial composition of sample participants was mixed (r = 0.10, SD = 0.18), and then for studies where the majority of sample participants were non-White (r = 0.06, SD = 0.17). Confidence intervals around each effect size indicate that the point estimates for the White and mixed subgroups were significantly different from zero (CI_r = 0.21-0.36 and CI_r = 0.06-0.14, respectively); however, the estimate generated from studies where the majority of sample participants were non-White, did not reach a level of statistical significance (CI_r = -0.02-0.10). Results remained fairly consistent when weighted by sample size. While the weighted mean effect size increased slightly to 0.30 for the White subgroup, weighted estimates for the non-White and mixed subgroups remained at 0.04 and 0.10, respectively. Similarly, the confidence interval around the weighted mean effect size for the White subgroup widened to 0.21-0.38, while, the CIs for the non-White and mixed subgroups remained between -0.02-0.10 and 0.06-0.14, respectively. It is important to note that

the precision of the estimates was only satisfactory for the mixed subgroup (w_r and $w_{Z+} = 0.08$). Both the White and non-white subgroups yielded confidence intervals with widths greater than the 0.10 threshold for precision ($w_r = 0.15$ and $w_{Z+} = 0.17$ for the White subgroup; w_r and $w_{Z+} = 0.12$ for the non-White subgroup).

Assuming a 50 percent base rate, the findings discussed above can be translated into the BESD statistic. More specifically, MST yielded an average improvement rate of 29 percent over the comparison group condition (or 64.5% versus 35.5%) when the majority of the sample was White. When the sample's racial composition was mixed, MST, on average, improved outcome by 10 percent (or 55% versus 45%). Finally, when the majority of the sample was non-white, MST yielded an average improvement rate of just 4 percent over the comparison group condition (or 52% versus 48%)

With the differential effects identified above, it is not surprising that significant differences between racial subgroups were also found. When taking a closer look at the widths of the confidence intervals, more specifically, it can be seen that the higher mean effect size for the White subgroup was significantly different from the smaller estimates produced for the non-White and mixed subcategories. While the effect size for the mixed subgroup was more than two times larger than the estimate for the non-white group, this difference did not reach statistical significance.

The final relationship examined in Table 3.17 is related to sample participants' histories of arrest. Before the results are reported, it should be noted that the no/cannot tell and mixed categories were collapsed to allow for more cases to fall across the subgroups. As such, studies where the majority of sample participants (70% or more) had at least one prior arrest were coded as yes, while the remaining studies were coded as no. As can be seen from the table below,

treatment effects in favor of MST were found for studies where the majority of sample participants had at least one prior arrest, as well as for studies where the majority of the sample did not have at least one prior arrest. Only the point estimate for the subgroup where the majority of participants had at least one prior arrest, however, was deemed significantly different from zero (r = 0.21, SD = 0.20). This was evident by its confidence interval not including 0.00 (CI_r = 0.15-0.28). The CI around the estimate for the category where the majority of participants did not have one prior arrest contained 0.00, indicating that the mean effect size of 0.05 was not significant (CI_r = -0.01-0.12). Weighted mean effect sizes and corresponding confidence intervals for the no and yes categories were fairly similar to the unweighted estimates. The only difference was the mean effect size for the yes subgroup, which increased slightly to 0.22 when weighted by sample size. Importantly, the precision of both subgroups' estimates should be interpreted with caution, as the width of the CI for the no subgroup was 0.13 and the width for the yes subgroup was just slightly smaller, falling at 0.12.

With the average point estimate generated from studies where the majority of the sample had at least one prior arrest being more than four times larger than the estimate generated from studies where the majority of the sample did not have at least one prior arrest, it is not surprising that the estimates between subgroups were statistically significant (i.e., r = 0.21 for the yes subgroup was significantly different from r = 0.05 for the no subgroup). Another way to examine the magnitude of this effect is to employ the BESD statistic. Specifically, when the majority of the sample had at least one prior arrest, outcome improved, on average, by 60.5 percent for MST participants, compared to 39.5 percent for comparison group participants. When the majority of the sample did not have at least one prior arrest, outcome improved, on

average by just 52.5 percent for MST participants, compared to 47.5 percent for comparison group participants.

Treatment and Study Characteristics. Treatment characteristics, as well as factors related to the studies themselves could also help to explain some of the variability in the point estimates observed in the Main Effects subsection above. Table 3.18 explores these factors as they relate to MST treatment effectiveness and the outcomes of interest in the current study. Sufficient data were available for eight treatment characteristics, including: (1) comparison group received, (2) source of clients for treatment, (3) MST program completion rate, (4) average number of MST treatment days, (5) adherence to MST model, (6) number of study sites, (7) study design, and (8) study setting. Moderators related to study characteristics could only be explored for average follow-up period.

The variable for type of treatment and/or services comparison group youths received was collapsed into a dichotomous variable representing one specific treatment *or* service and multiple treatment/services. When examining the effect sizes for single and multiple subgroups, Table 3.18 reveals that higher point estimates were generated from studies where comparison group youths received a single treatment or service ($r = 0.16$, SD = 0.21) versus multiple treatment/services ($r = 0.12$, SD = 0.20). Confidence intervals around these mean estimates also revealed that both subgroups were significantly different from 0.00, however ($CI_r = 0.08$-0.24 and $CI_r = 0.09$-0.16, respectively). When estimates were weighted by sample size, average effect sizes for both subcategories increased minimally by 0.01 ($Z^+ = 0.17$ for the single subgroup and $Z^{+ =} = 0.13$ for the multiple subgroup). In a similar vein, corresponding confidence intervals widened moving to 0.08-0.25 for the single subgroup and 0.09-0.17 for the multiple subgroup. The results produced for the multiple subgroup can be considered precise, as the

width of its corresponding CI was under the 0.10 threshold considered by researchers in the field to be acceptable ($w_r = 0.07$ and $w_{Z+} = 0.08$). The findings related to the single subgroup, on the other hand, should be interpreted with caution as the width of its respective CI is much wider than the 0.10 limit ($w_r = 0.16$ and $w_{Z+} = 0.17$).

The application of the BESD helps to further understand how comparison group treatment/services moderated the effect of MST on outcome. MST youths improved on outcome at a rate of 16 percent when compared to comparison group youths who were referred to a single treatment or service (or 58% versus 42%), while outcome improved at a rate of 12 percent when compared to comparison group youths who were referred to multiple treatments and/or services (or 56% versus 44%).

While the findings suggest that larger effects, on average, were produced when MST youths were compared to youths who received a single treatment or service, the difference between subgroups was not enough to reach statistical significance. This is evident upon further exploration of the confidence intervals across subgroups. Notably, both weighted and unweighted confidence intervals for single and multiple subgroups overlap with one another.

The results in the table below also suggest differential treatment effects depending on the source of referral for MST and comparison group treatment/services. In particular, the mean effect size was highest for youths when referred to treatment via a criminal justice agency ($r = 0.19$, SD = 0.20) and lowest for youths when referred to treatment via a variety of social service and criminal justice agencies ($r = 0.05$, SD = 0.11). The average effect size produced from studies where youths were referred to treatment via a non-criminal justice agency (e.g., mental health agency) was slightly larger than the estimate generated from studies where youths were referred from a mix of referral sources ($r = 0.07$, SD = 0.20). While both the criminal justice and

non-criminal justice subgroups' estimates were significantly different from 0 (CI_r = 0.14-0.24 and CI_r = 0.02-0.12, respectively), the confidence interval around the mixed subgroup's point estimate included 0, suggesting that the average effect of 0.05 is not any better than what chance alone might find (Smith et al., 2009). Importantly, the width of the confidence interval around the mixed subgroup's point estimate is wider than than the accepted limit for precision (w_r = 0.13). Confidence intervals around the estimates for the criminal justice and non-criminal justice subgroups, on the other hand, were within the accepted limit, suggesting that the findings discussed above are precise. The findings did not change considerably when taking sample size into consideration. Specifically, the weighted mean effect size for the non-criminal justice agency subgroup, criminal justice agency subgroup, and mixed subgroup stayed at 0.07, 0.19, and 0.05, respectively.

Assuming a 50 percent base rate, the findings discussed above can be translated into the BESD statistic. More specifically, when referred to treatment via a criminal justice agency, outcome improved, on average, by 59.5 percent for the MST group, while outcome improved, on average, by 40.5 percent for the comparison group. Related to youths referred to treatment via a variety of social service and criminal justice agencies, outcome improved, on average, by 52.5 percent for MST participants, while it improved by 47.5 percent, on average, for comparison group participants. Finally, youths referred to treatment via a non-criminal justice agency improved outcome, on average, by 53.5 percent, compared to 46.5 percent for youths who participated in comparison group treatment/services.

When taking a closer look at Table 3.18 and the relationships between source of clients for treatment subcategories, it can be gathered that the larger mean effect size produced for the

Table 3.18
Mean Effect Sizes: Treatment and Study Characteristics–All Effect Sizes

	k	N	*r*	SD	95% CI	Z^+	95% CI
Comparison Group Received							
Single (i.e., treatment *or* services)	30	6,434	0.16	0.21	0.08–0.24	0.17	0.08–0.25
Multiple (i.e., treatment/services)	97	12,909	0.12	0.20	0.09–0.16	0.13	0.09–0.17
Source of Clients for Treatment							
Non-Criminal Justice Agency	49	5,727	0.07	0.20	0.02–0.12	0.07	0.02–0.13
Criminal Justice Agency	70	11,423	0.19	0.20	0.14–0.24	0.19	0.14–0.24
Mix	8	2,193	0.05	0.11	-0.02–0.11	0.05	-0.02–0.11
Program Completion—MST							
No	23	5,634	0.02	0.20	-0.06–0.10	0.02	-0.06–0.10
Yes (80% or more)	95	11,512	0.17	0.20	0.13–0.21	0.17	0.13–0.21
Average Number of MST Treatment Days (150 days or more)							
No	57	10,540	0.09	0.21	0.04–0.15	0.09	0.04–0.15
Yes	31	3,407	0.18	0.17	0.12–0.23	0.18	0.12–0.24
Adherence to MST Model							
Appeared Adherence was Measured	102	13,713	0.12	0.20	0.08–0.16	0.09	0.08–0.16
No Indication Adherence was Measure	25	5,630	0.18	0.20	0.11–0.25	0.24	0.11–0.26

Table 3.18
Mean Effect Sizes: Treatment and Study Characteristics–All Effect Sizes

	k	N	*r*	SD	95% CI	Z^+	95% CI
Number of Study Sites							
Single-Site	93	11,495	0.16	0.21	0.12–0.21	0.17	0.12–0.21
Multi-Site	34	7,848	0.06	0.17	0.00–0.12	0.06	0.00–0.12
Study Design							
Randomized	101	12,376	0.13	0.21	0.09–0.17	0.13	0.09–0.18
Quasi-Experimental	26	6,967	0.14	0.20	0.07–0.21	0.14	0.07–0.21
Study Setting							
Efficacy Trial	37	3,889	0.24	0.21	0.17–0.31	0.24	0.17–0.32
Effectiveness Study	90	15,454	0.09	0.18	0.05–0.13	0.09	0.05–0.13
Average Follow-Up Period—Delinquency (*k* = 26)							
0—12 Months	16	10,668	0.06	0.17	-0.01–0.13	0.06	-0.01–0.13
13—25^+ Months	10	5,119	0.19	0.21	0.08–0.30	0.19	0.08–0.31
Average Follow-Up Period—All Others (*k* = 101)							
0—12 Months	84	12,259	0.12	0.20	0.08–0.16	0.12	0.08–0.16
13—25^+ Months	17	2,164	0.24	0.23	0.10–0.37	0.24	0.10–0.38

criminal justice subgroup was significantly different from the mean effect size estimates produced for the non-criminal justice and mixed subgroups. Although the average effect size generated from studies examining treatment effectiveness for youths referred to MST/comparison group treatment/services via a non-criminal justice agency was slightly larger than the estimate generated from studies examining treatment effectiveness for youths referred to MST/comparison group treatment/services via a variety of referral sources, this CIs for these subgroups overlapped, suggesting that the subcategories are not significantly different from one another.

The next moderating variable explored in Table 3.18 relates to MST participants' treatment completion status. As can be seen, studies produced larger effect sizes, on average, when at least 80 percent of the sample completed MST versus studies where the majority of the sample did not complete MST (r = 0.17, SD = 0.20 versus r = 0.02, SD = 0.20). The relationship for completers and treatment effectiveness was significantly different from 0, yielding a confidence interval around the point estimate between 0.13 and 0.21. The confidence interval around the point estimate for the relationship between the less than 80 percent completers subgroup and treatment effectiveness was not deemed significant, however (CI_r = -0.02-0.11). It is important to notice that the width of the CI for the yes subgroup was within the 0.10 precision threshold, while the width of the CI for the no subgroup was considerably wider than the accepted limit. The results were identical when estimates were weighted by sample size. Further, when the BESD statistic was computed for the completers versus non-completers, the findings translated into a minimal improvement rate (2%) for MST participants over comparison group participants when the majority of the MST group did not complete treatment (or 51% versus 49%) and a considerably higher improvement rate (17%) for MST participants over

comparison group participants when the majority of the MST group did complete treatment (or 58.5% versus 41.5%).

The confidence intervals were examined once again to determine whether the larger effect size produced from studies where at least 80 percent of the participants completed MST was significantly different from the estimate produced from studies where less than 80 percent of the participants completed MST. Notably, the difference in the magnitude of the effect sizes was significant, as the lower bound of the CI for the yes subgroup was 0.03 away from the upper bound of the CI for the no subgroup (i.e., CIs did not overlap).

When examining the relationship between outcome and treatment effectiveness relative to the average number of MST treatment days, Table 3.18 reveals that the average effect size generated from studies where youths received 150 days or more of treatment was two times larger than the average effect size generated from studies where youths received less than 150 days of treatment (r = 0.18, SD = 0.17 versus r = 0.09, SD = 0.21). Further, the confidence interval around each subgroups' respective point estimate did not contain 0, suggesting that they were significantly different from 0 (CI_r = 0.12-0.23 and CI_r = 0.04-0.15, respectively). Estimates and corresponding CIs were almost identical when weighted by sample size, with only the CI for the yes subgroup widening slightly (CI_{Z+} =0.012-.024). Notably, the width of both subcategories' confidence intervals was greater than 0.10, signifying that the average effect size estimates were not precise. When translating the findings into the BESD statistic, the results indicate that the success rate of MST represented an 18 percent improvement in outcome, on average, when participants received at least 150 days of treatment (or 59% versus 41%) and represented only a 9 percent improvement in outcome, on average, when participants received less than 150 days of treatment (or 54.5% versus 45.5%).

While the point estimate for the yes subgroup was considerably larger than the point estimate for the no subgroup, the difference in the magnitude of these effect sizes only approached significance. More specifically, it appears that MST was more effective in changing outcome when youths received at least 150 days of treatment; however, this relationship was not found to be statistically significant.

The results presented in Table 3.18 demonstrate that the average effect generated from studies where MST adherence was not measured was larger than the average effect generated from studies where MST adherence was measured (r = 0.18, SD = 0.20 versus r = 0.12, SD = 0.20, respectively). It is important to note that when further analyses were explored, the results indicated that the majority of the studies falling into the "no" subcategory on this variable were early studies, conducted by Henggeler and the MST developers, and/or before the TAM instrument became a formal MST requirement. Regardless of whether it appeared that adherence was measured, however, the point estimates indicate that MST significantly improved outcome. That is, both subcategories' corresponding CIs did not include 0.00, and thus should be considered significantly different from zero or what would be expected by chance alone (CI_r = 0.11-0.25 for the no indication adherence was measured subgroup and CI_r = 0.08-0.16 for the appeared adherence was measured subgroup). Note that the mean r values and the Z^+ values were nearly identical in magnitude for both subgroups. The only notable difference was the width of the CI for the no indication adherence was measured subgroup (CI_{Z+} = 0.11-0.26). Further, while it appears that studies where there was no indication that adherence was measured did slightly better on outcome than those studies where it appeared that adherence was measured, the difference in the magnitude of the effect was not significantly different (i.e., the subgroups confidence intervals overlapped with one another). The results for the adherence subgroup

should be considered more precise than the results for the non-adherence subgroup, however. Finally, when translating the findings into the BESD statistic, the results indicate that the success rate of MST represented a 12 percent improvement in outcome (or 56% versus 44%) over the comparison group when it appeared adherence was measured. When it appeared adherence was not measured, on the other hand, the findings suggest that the success rate of MST represented an 18 percent improvement in outcome (or 59% versus 41%) over the comparison group.

Table 3.18 suggests that number of study sites was also significantly associated with outcome. Specifically, the mean effect size produced from single-site studies was considerably larger than the mean estimate produced from multi-site studies (r = 0.16, SD = 0.21 versus r = 0.06, SD = 0.17). Interestingly, the magnitude of the effect for the single-site subgroup was considered significantly different from 0, as evidenced by the CI not containing 0.00 (CI_r = 0.12-0.21); however, the magnitude of the effect for the multi-site subgroup was not deemed significantly different from 0.00 because it's corresponding confidence interval included 0.00 (CI_r = 0.00-0.12). Moreover, when comparing the estimates' respective confidence intervals, it can be determined that the difference in the magnitude of these effects only approached statistical significance. That is, the lower bound of the CI for the single-site subgroup just touches the upper bound of the CI for the multi-site subgroup, suggesting that MST may be more effective when conducted at a single site, as opposed to multiple sites. These findings should be interpreted with caution, however, as the difference between subgroups is not considered statistically significant. Related, only the width of the CI for the single-site subgroup is within an acceptable limit to deem the findings precise. To illustrate the results further, the BESD was computed, which yielded a 16 percent improvement rate in outcome for MST participants compared to comparison group participants when studies were conducted at a single site, versus

a minimal 6 percent improvement rate in outcome for MST participants compared to comparison group participants when studies were conducted across multiple sites.

When examining Table 3.18 related to study design, the findings demonstrate that, on average, randomized-controlled studies produced an effect size and weighted effect size of 0.13 and quasi-experimental studies produced an effect size and weighted effect size of 0.14. Both point estimates were significantly different from what we may expect from chance alone, as evidenced by their confidence intervals not containing 0.00 (CI_r = 0.09-0.17 and CI_{Z+} = 0.09-0.18 and CI_r and CI_{Z+} = 0.07-0.21, respectively). When comparing the confidence intervals between each subgroup, however, it can be noticed that the bounds overlap on both the lower and upper limits. Given that the point estimates are almost identical in magnitude, this nonsignificant finding is not surprising. In addition, it is important to point out that the width of the CI is quite substantial for the quasi-experimental subgroup (w_r and w_{Z+} = 0.14), indicating that its mean effect size of 0.14 may not be as precise as the mean effect size produced from randomized studies (w_r = 0.08 and w_{Z+} = 0.09).

Assuming a 50 percent base rate, the findings discussed above can be translated into the BESD statistic. More specifically, randomized-controlled studies improved outcome, on average, by 56.5 percent for the MST group, while outcome improved, on average, by 43.5 percent for the comparison group. Rates of improvement for both MST and comparison groups were just slightly higher for quasi-experimental studies—57 percent and 43 percent, respectively.

The results of moderator analyses also demonstrate that the setting in which the study was conducted was associated with differential treatment effects. Specifically, the average effect size generated from efficacy trials was 0.24, with a confidence interval between 0.17 and 0.31. Notably, because the CI does not contain 0, the findings suggest that efficacy trials were

significantly associated with outcome. In addition, the mean effect size produced from effectiveness studies was 0.09, with a confidence interval between 0.05 and 0.13. Similar to efficacy trials, these findings suggest that effectiveness studies were significantly associated with outcome because the CI does not contain 0. Note that the weighted mean effect sizes for both the efficacy trial and effectiveness study subgroups were identical in magnitude to their unweighted counterparts. The only notable difference was the width of the CI for the efficacy trial subgroup, which increased minimally by 0.01 (CI_{Z+} = 0.017-0.32).

When comparing the confidence intervals between the two subgroups discussed above, the findings suggest that the difference in the magnitude of the effect size is statistically significant. That is, while both effect sizes are significant, the lack of overlap between the confidence intervals suggests that efficacy trials were significantly more effective than effectiveness studies. To interpret the findings in an alternative manner, the BESD was computed, which indicates that, on average, there was a 24 percent difference between the MST group and comparison group on outcomes considered in the analysis for efficacy trials, while there was a 9 percent difference between the MST group and comparison group on outcomes considered in the analysis for effectiveness studies. Note that similar to findings discussed previously, the width of the CI for one subgroup (efficacy trial) was much larger than the 0.10 threshold for deeming effect size estimates precise (w_r = 0.31 and w_{Z+} = 0.32). As such, more confidence should be placed in the 0.09 estimate for effectiveness studies because its width is within the precision boundary (w_r and w_{Z+} = 0.09).

The final relationship in Table 3.18 examines treatment effectiveness by follow-up period. Consistent with previous analyses, follow-up period was separated for delinquency and all other outcomes. Note that categories were collapsed for both follow-up moderators to

represent 0-12 months post-treatment and 0-25+ months post-treatment. As can be seen, the mean effect and weighted mean effect sizes generated from studies that employed a follow-up period between 0 and 12 months was 0.06 (SD = 0.17) for delinquency and 0.12 (SD = 0.20) for all other outcomes of interest. While the point estimates for the "other outcomes" category were significantly different from 0 (CI_r and CI_{Z+} = 0.08-0.16), the presence of 0 in the confidence interval for the delinquency point estimate indicates that the mean effect size was not significant (CI_r and CI_{Z+} = -0.01-0.13). Regarding studies that employed a follow-up period between 13 and 25 months or more, the mean effect and weighted mean effect sizes generated was 0.19 (SD = 0.21) for delinquency and 0.24 (SD = 0.20) for all other outcomes of interest. Both unweighted and weighted point estimates for this longer follow-up period were significantly different from 0 for delinquency and other outcomes (CI_r = 0.08-0.30 and CI_{Z+} = 0.08-0.31; CI_r = 0.10-0.37 and CI_{Z+} = 0.10-0.38, respectively).

When comparing the shorter follow-up period to the longer follow-up period for delinquency, the findings suggest that the difference in the magnitude of the effect sizes is not statistically significant. Note, however, that the confidence intervals overlap only on the lower bound of the CI for the 13-25+ months subgroup and the upper bound of the CI for the 0-12 months subgroup. This suggests that longer follow-up periods may be associated with greater treatment effects. Given that the CIs did not overlap, and in addition, that the width of the CIs was greater than the 0.10 precision threshold, caution is advised in making definitive conclusions (Gendreau and Smith, 2007) (w_r and w_{Z+} = 0.14 for the 0-12 months subgroup and w_r and w_{Z+} = 0.23 for the 13-25+ months subgroup).

In line with the discussion above, the results also indicate that the difference in the magnitude of the effect sizes is not statistically significant when comparing the shorter follow-up

period to the longer follow-up period for all other outcomes. Interestingly, only the upper limit of the 0-12 months subgroup appears to overlap with the lower limit of the CI for the 13-25+ subgroup, which suggests that this relationship may be approaching significance. Again, these results should not be deemed conclusive, as the relationship did not reach statistical significance and the margin of error is considerably wide for the longer follow-up subgroup (w_r and w_{Z+} = 0.14 for the 0-12 months subgroup and w_r = 0.27 and w_{Z+} = 0.28 for the 13-25+ subgroup).

Moderating Effects—Delinquency

In the previous subsection, moderating effects were explored across all outcomes of interest in the current study (k = 127). The present subsection and the one to follow, attempt to uncover significant moderating effects for delinquency (k = 26) and psychopathology and mental health (k = 18), respectively. It is important to note that the number of effect sizes (k) in several of the categories were limited and/or disparate between subgroups. As such, the widths of the CIs associated with mean r and Z^+ for the majority of moderators explored in the analyses were greater than 0.10. It is therefore suggested that the results be interpreted with caution until future studies are able to replicate similar findings.

Publication and Study Context Characteristics. Table 3.19 presents the mean effect size estimates for publication and study context characteristics for delinquency (k = 26). Only one moderating effect explored in the table below reached statistical significance. Specifically, the magnitude of the unweighted and weighted effect sizes generated from studies where the evaluator(s) was involved in MST implementation were substantially larger than the unweighted and weighted effect sizes generated from studies where the evaluator(s) was not involved in MST implementation (r = 0.17 and Z^+ = 0.18 versus r and Z^+ = 0.02). As can be seen, the CIs about the mean r and Z^+ for the yes subgroup did not overlap (CI_r and CI_{Z+} = 0.09-0.26) with the

Table 3.19
Mean Effect Sizes: Publication and Study Context Characteristics–Delinquency

	k	N	*r*	SD	95% CI	Z^+	95% CI
Publication Type							
Unpublished	4	1,559	0.00	0.09	-0.08–0.08	0.00	-0.08–0.08
Published	22	5,785	0.13	0.18	0.06–0.20	0.13	0.06–0.20
Publication Decade							
1980 or 1990	6	662	0.18	0.21	0.02–0.33	0.18	0.02–0.34
2000 or 2010	20	6,682	0.09	0.15	0.02–0.15	0.09	0.02–0.15
Geographic Location							
U.S.	20	6,235	0.10	0.17	0.04–0.17	0.11	0.03–0.18
Outside the U.S.	6	1,109	0.11	0.19	-0.03–0.25	0.11	-0.01–0.23
Decade Data Collection Started							
1980 or 1990	10	1,309	0.21	0.18	0.10–0.32	0.22	0.10–0.34
2000 or 2010	14	5,753	0.06	0.13	-0.01–0.13	0.06	-0.01–0.13
Involvement of Evaluator							
No	11	5,373	0.02	0.12	-0.05–0.08	0.02	-0.05–0.08
Yes	15	1,971	0.18	0.17	0.09–0.26	0.18	0.09–0.26
Involvement of MST Developers							
No	15	6,177	0.07	0.15	0.00–0.14	0.07	0.00–0.14
Yes	11	1,167	0.16	0.20	0.06–0.26	0.17	0.06–0.27

mean *r* and Z^+ for the no subgroup (CI_r and CI_{Z+} = -0.05–0.08). These findings suggest that MST may be more effective on delinquency when the scholar(s) is involved in treatment evaluation *and* implementation.

While other publication and study context characteristics did not reach statistical significance, certain relationships are worth noting. For example, mean effect sizes generated from published studies were considerably larger in magnitude compared to mean effect sizes generated from unpublished studies (*r* and Z^+ = 0.13, SD = 0.18 versus *r* and Z^+ = 0.00, SD = 0.09). Although the confidence limits overlapped (CI_r and CI_{Z+} = 0.06-0.20 for published studies and CI_r and CI_{Z+}= -0.08-0.08 for unpublished studies), the commonality in the CIs' distributions was minimal. That is, only the lower bound of the CI for the non-published studies subgroup and the upper bound of the CI for the published studies subgroup overlapped. In this way, published studies may be associated with greater reductions in delinquency, compared to unpublished studies.

Table 3.19 also suggests that the decade in which data collection began may have been associated with larger treatment effects on delinquency. When comparing the confidence intervals about mean *r* and Z^+ for the 1980 or 1990 subgroup (CI_r = 0.10–0.32 and CI_{Z+} = 0.10–0.34), to the confidence intervals about mean *r* and Z^+ for the 2000 or 2010 subgroup (CI_r and CI_{Z+} = -0.01–0.13), it can be seen that the respective limits overlapped slightly on the lower limit of the former subgroup's CI and the upper limit of the latter subgroup's CI. In short, the decade data collection started did not have a significant moderating effect on delinquency; however, the findings suggest that studies where data collection began in the late 1900s may have had a greater impact on delinquency, compared to studies where data collection began in the early 2000s (*r* and Z^+ = 0.18, SD = 0.21 versus *r* and Z^+ = 0.09, SD = 0.15, respectively).

Finally, in examining delinquency across publication and study context characteristics, the results indicate a tendency for the two geographic location subgroups—U.S. and outside the U.S.—to have performed similarly to one another, judging by the considerable overlap in the CIs of their respective point estimates (r = 0.10, CI_r = 0.04–0.17 and Z^+ = 0.11, CI_{Z+} = 0.03–0.18 versus r and Z^+ = 0.10, CI_r = -0.03–0.25 and CI_{Z+} = -0.01–0.23, respectively). In a similar vein, there was substantial overlap in the confidence limits of the respective r and Z^+ values for publication decade subgroups—1980 or 1990 and 2000 or 2010 (r and Z^+ = 0.18, CI_r = 0.02–0.33 and CI_{Z+} = 0.02-0.34 versus r and Z^+ = 0.09, CI_r and CI_{Z+} = 0.02–0.15), as well as MST developers subgroups—no and yes (r and Z^+ = 0.07, CI_r and CI_{Z+} = 0.00-0.14 and r = 0.16, CI_r = 0.06–0.126 and Z^+ = 0.17, CI_{Z+} = 0.06-0.27. Note, however, that the differences in magnitude of the associated mean effect size estimates for publication decade subgroups and involvement of MST developers subgroups were substantially large, suggesting that earlier studies and involvement of the MST developers may have had a greater impact on delinquency, compared to studies written in later decades or studies where Henggeler and his development team were not involved in study implementation.

Sample Characteristics. Table 3.20 presents the results comparing mean effect size estimates across a number of sample characteristics. As indicated, the mean effect size significantly differed depending on type of youth and race. Specifically, the mean effect size was highest for studies exploring the effectiveness of MST for sex offenders (r and Z^+ = 0.14, CI_r = 0.14-0.42 and CI_{Z+} = 0.14-0.45) and for studies where the majority of participants (70% or more) were White (r and Z^+ = 0.25, CI_r = 0.15-0.34 and CI_{Z+} = 0.15-0.36). Note that these findings should not be taken to suggest that MST was ineffective on delinquency for youths presenting conduct or mental health issues or ineffective when the sample was predominantly

Table 3.20
Mean Effect Sizes: Sample Characteristics–Delinquency

	k	N	*r*	SD	95% CI	Z^+	95% CI
Type of Youth							
Conduct Problems	17	6,276	0.11	0.16	0.03–0.18	0.11	0.03–0.18
Mental Health	6	873	0.05	0.09	-0.03–0.12	0.05	-0.03–0.13
Sexual Offenders	3	195	0.29	0.14	0.14–0.42	0.29	0.14–0.45
Average Age							
Under 15	14	4,103	0.12	0.19	0.03–0.21	0.12	0.03–0.21
15 or Older	12	3,241	0.09	0.15	0.01–0.18	0.09	0.01–0.18
Race							
White (70.0% or more)	7	859	0.25	0.12	0.15–0.34	0.25	0.15–0.36
Non-White (70.0% or more)	7	2,565	0.01	0.10	-0.08–0.09	0.01	-0.08–0.09
Mix	11	3,511	0.10	0.18	0.01–0.19	0.10	0.01–0.19
At Least One Arrest							
No	8	2,175	0.05	0.15	-0.05–0.15	0.05	-0.05–0.15
Yes (70.0% or more)	9	2,289	0.17	0.14	0.07–0.26	0.17	0.07–0.26

non-White or a mix of White and non-White youths. In the majority of cases, the mean effect size and weighted mean effect size estimates were statistically significant (i.e., better than what we would expect by chance) and in all cases, in the right direction (i.e., MST participants had greater rates of improvement, compared to comparison group youths).

While type of youth and race had some impact on the mean effect size, average age, and at least one arrest did not reach statistical significance. The magnitude of the effect size produced from studies where the average age was less than 15 was slightly larger than the magnitude of the effect size produced from studies where the average age was 15 or older (*r* and $Z^+ = 0.12$ versus *r* and $Z^+ = 0.09$). Although the CI of the older youths subgroup overlapped with the CI of the younger youths subgroup (CI_r and CI_{Z+} = 0.01-0.18 and CI_r and CI_{Z+} = 0.03-0.21, respectively), these findings may suggest, at least preliminarily, that MST was marginally more effective for younger versus older youths on delinquency. Similarly, the magnitude of mean *r* and mean Z^+ effect sizes produced from studies where the majority of the sample (70% or more) had at least one arrest was more than three times larger than the mean estimates produced from studies where the majority of the sample did not have at least one arrest (*r* and $Z^+ = 0.17$ versus *r* and $Z^+ = 0.05$). While the confidence limits of the respective *r* and Z^+ values overlapped on the yes subgroup's lower bound and the no subgroup's upper bound (CI_r and CI_{Z+} = 0.07-0.26 and CI_r and CI_{Z+} = -0.05-0.15, respectively), these findings may suggest that MST was more effective for "riskier" juveniles compared to less "riskier" juveniles on delinquency.

Treatment and Study Characteristics. Table 3.21 presents the results comparing mean effect size estimates across treatment and study characteristics. While none of the mean effect sizes significantly differed across moderator categories, several relationships are worth noting. In particular, studies where youths were referred to treatment via a criminal justice agency were

associated with larger effect sizes (r and $Z^+ = 0.16$), in comparison to studies where youths were referred to treatment via a non-criminal justice agency (r and $Z^+ = 0.05$), or a mix of social service and criminal justice agencies ($r = 0.02$ and $Z^+ = 0.03$). While the confidence limits of the respective r and Z^+ values overlapped between criminal justice ($CI_r = 0.07\text{-}0.24$ and $CI_{Z+} = 0.07\text{-}0.25$), non-criminal justice (CI_r and $CI_{Z+} = -0.07\text{-}0.16$), and mixed ($CI_r$ and $CI_{Z+} = -0.06\text{-}0.09$) subgroups, these findings suggest that MST may be more effective on delinquency when youths are referred to treatment via a criminal justice referral source.

Substantially larger mean effect sizes were also generated from studies where 80 percent or more of the participants completed MST ($r = 0.14$ and $Z^+ = 0.15$) versus studies where less than 80 percent of the participants completed treatment (r and $Z^+ = 0.01$). Ultimately, the CI of the no subgroup's r and Z^+ values overlapped with the CI of the yes subgroup's r and Z^+ values (CI_r and $CI_{Z+} = -0.08\text{-}0.10$ and CI_r and $CI_{Z+} = 0.06\text{-}0.23$), deeming the relationship statistically nonsignificant; however, the findings indicate, at least preliminarily, that MST was more effective on delinquency when the majority of youths completed treatment. Similarly, the magnitude of the effect sizes generated from studies where youths participated in MST treatment for 150 days or more was two times larger than the estimate produced from studies where youths participated in MST treatment for less than 150 days (r and $Z^+ = 0.016$ versus r and $Z^+ = 0.08$). While the CIs of the respective r and Z^+ values overlapped between the no (CI_r and $CI_{Z+} = -0.01\text{-}0.17$) and yes ($CI_r = 0.08\text{-}0.24$ and $CI_{Z+} = 0.08\text{-}0.25$) subcategories, the findings may suggest that MST participants were more successful (i.e., achieved greater reductions in delinquency) when they received more dosage (i.e., more treatment) compared to participants who received less dosage.

Related to number of study sites, Table 3.20 indicates that the CIs of the respective *r* and Z^+ values for the multi-site subgroup overlapped marginally with the CIs of the respective *r* and Z^+ values for the single-site subgroup (CI_r and CI_{Z+} = -0.05-0.10 and CI_r and CI_{Z+} = 0.07-0.23). Nevertheless, studies where MST was conducted at a single site produced considerably larger effect sizes (*r* and Z^+ = 0.015), compared to studies where MST was conducted at multiple sites (*r* and Z^+ = 0.03). This finding may suggest that MST was slightly more effective on delinquent behavior when treatment was conducted at one site during the study duration, as opposed to multiple sites during the study duration. Related, efficacy trials were associated with larger effect sizes (*r* = 0.22 and Z^+ = 0.23), in comparison to effectiveness studies (*r* and Z^+ = 0.07). Similar to the moderator examining number of study sites, the CIs of the associated mean *r* and Z^+ overlapped between the effectiveness study subgroup (CI_r and CI_{Z+} = -0.01-0.13) and the efficacy trial subgroup (CI_r = 0.05-0.38 and CI_{Z+} = 0.05-0.40). Ultimately, however, the results suggest, at least preliminarily, that efficacy trials or studies where MST therapists were trained graduate students had a greater impact on delinquency, compared to effectiveness studies or studies where MST therapists were trained clinicians (*r* and Z^+ = 0.22 versus *r* and Z^+ = 0.07). Finally, in examining delinquency across treatment and study characteristics, the results indicate a tendency for the magnitude of the difference between the two comparison group subcategories—single and multiple—to perform similarly to one another (*r* = 0.13 and Z^+ = 0.14 versus *r* and Z^+ = 0.10). This was evidenced by the considerable overlap in the CIs for the multiple subgroup (CI_r and CI_{Z+} = 0.03-0.17) and the single subgroup (CI_r = 0.01-0.25 and CI_{Z+} = 0.01-0.26). Similarly, there was substantial overlap in the confidence limits of respective mean *r* and Z^+ for the two adherence to MST model subgroups—appeared adherence was measured and no indication that adherence was measured (*r* and Z^+ = 0.10, CI_r and CI_{Z+} = 0.03-0.17 versus *r*

Table 3.21
Mean Effect Sizes: Treatment and Study Characteristics–Delinquency

	k	N	*r*	SD	95% CI	Z^+	95% CI
Comparison Group Received							
Single (i.e., treatment *or* services)	8	4,037	0.13	0.20	0.01–0.25	0.14	0.01–0.26
Multiple (i.e., treatment/services)	18	3,307	0.10	0.16	0.03–0.17	0.10	0.03–0.17
Source of Clients for Treatment							
Non-Criminal Justice Agency	7	842	0.05	0.16	-0.07–0.16	0.05	-0.07–0.16
Criminal Justice Agency	16	5,585	0.16	0.18	0.07–0.24	0.16	0.07–0.25
Mix	3	917	0.02	0.05	-0.06–0.09	0.03	-0.06–0.09
Program Completion—MST							
No	8	3,665	0.01	0.20	-0.08–0.10	0.01	-0.08–0.10
Yes (80% or more)	15	2,180	0.14	0.17	0.06–0.23	0.15	0.06–0.23
Average Number of MST Treatment Days							
No	13	491	0.08	0.23	-0.01–0.17	0.08	-0.01–0.17
Yes (150 days or more)	5	575	0.16	0.11	0.08–0.24	0.16	0.08–0.25
Adherence to MST Model							
Appeared Adherence was Measured	20	3,521	0.10	0.17	0.03–0.17	0.10	0.03–0.17
No Indication Adherence was Measure	6	3,823	0.14	0.19	0.00–0.26	0.14	0.00–0.27

Table 3.21
Mean Effect Sizes: Treatment and Study Characteristics–Delinquency

	k	N	*r*	SD	95% CI	Z^+	95% CI
Number of Study Sites							
Single-Site	17	3,087	0.15	0.18	0.07–0.23	0.15	0.07–0.23
Multi-Site	9	4,257	0.03	0.13	-0.05–0.10	0.03	-0.05–0.10
Study Design							
Randomized	18	2,428	0.13	0.19	0.05–0.21	0.13	0.05–0.21
Quasi-Experimental	8	4,916	0.07	0.13	-0.02–0.15	0.07	-0.02–0.15
Study Setting							
Efficacy Trial	7	729	0.22	0.22	0.05–0.38	0.23	0.05–0.40
Effectiveness Study	19	6,615	0.07	0.13	0.01–0.13	0.07	0.01–0.13
Average Follow-Up Period							
0—12 Months	16	10,668	0.06	0.17	-0.01–0.13	0.06	-0.01–0.13
13—25^+ Months	10	5,119	0.19	0.21	0.08–0.30	0.19	0.08–0.31

and Z^+ = 0.14 and CI_r = 0.00-0.26 and CI_{Z+} = 0.00-0.27). Note that while the CIs for the randomized and quasi-experimental subgroups overlapped considerably (CI_r and CI_{Z+} = 0.05-0.21 versus CI_r and CI_{Z+} = -0.02-0.15), the magnitude of mean *r* and Z^+ for the randomized subcategory was substantially larger compared to the estimate produced for the quasi-experimental subcategory (*r* and Z^+ = 0.13 versus *r* and Z^+ = 0.07).

Moderating Effects—Psychopathology and Mental Health

The current subsection investigates potential moderating effects between treatment and psychopathology and mental health. In line with previous subsections, many moderators could not be tested with any meaningful interpretation because of the relatively small number of total effect sizes (*k* = 19) and missing data in many cases (Smith et al., 2009). In this way, the widths of the CIs associated with mean *r* and Z^+ for the majority of moderators explored in the analyses were greater than 0.10. It is therefore suggested that the results be interpreted with caution until future studies are able to replicate similar findings.

Publication and Study Context Characteristics. Table 3.22 presents the mean effect size estimates for publication and study context characteristics for MST and psychopathology and mental health. Notably, none of the moderating effects explored in the table below reached statistical significance, which, as previously mentioned, was likely due to small cell frequencies. Three relationships, nonetheless, are worth exploring, as the magnitude of the effect size estimate for one subgroup was two to three times larger than the magnitude of the effect size estimate for the other subgroup. These moderators include: publication type, publication decade, and decade data collection started.

Table 3.22 demonstrates that published studies produced an average effect size that was two times larger than the average effect size produced from unpublished studies (*r* and Z^+ = 0.16

Table 3.22
Mean Effect Sizes: Publication and Study Context Characteristics–Psychopathology and Mental Health

	k	N	*r*	SD	95% CI	Z^+	95% CI
Publication Type							
Unpublished	1	155	0.08	--	-0.10–0.25	0.08	-0.10–0.26
Published	17	2,075	0.16	0.19	0.07–0.24	0.16	0.07–0.24
Publication Decade							
1980 or 1990	3	390	0.32	0.30	-0.01–0.59	0.33	-0.01–0.67
2000 or 2010	15	1,840	0.11	0.15	0.04–0.18	0.11	0.04–0.18
Geographic Location							
U.S.	13	1,530	0.16	0.20	0.05–0.27	0.16	0.05–0.27
Outside the U.S.	5	700	0.13	0.13	0.02–0.23	0.13	0.02–0.24
Decade Data Collection Started							
1980 or 1990	9	1,024	0.22	0.22	0.08–0.35	0.22	0.08–0.36
2000 or 2010	9	1,206	0.07	0.12	0.00–0.15	0.07	0.00–0.15
Involvement of the Evaluator							
No	5	727	0.10	0.09	0.02–0.18	0.10	0.02–0.18
Yes	13	1,503	0.17	0.21	0.06–0.27	0.17	0.06–0.28
Involvement of MST Developers							
No	8	1,122	0.12	0.15	0.02–0.22	0.12	0.02–0.22
Yes	10	1,108	0.18	0.21	0.05–0.31	0.18	0.05–0.32

versus *r* and $Z^+ = 0.08$). The table below also indicates that earlier studies (i.e., 1980 or 1990), on average, generated effect size estimates that were almost three times larger than those generated from later studies (i.e., 2000 or 2010) ($r = 0.32$ and $Z^+ = 0.33$ versus *r* and $Z^+ = 0.11$). Similarly, when data collection started in the 1980s or 1990s, the magnitude of the effect was more than three times larger than the estimate generated from studies where data collection started in the 2000s or 2010s (*r* and $Z^+ = 0.22$ versus *r* and $Z^+ = 0.07$). Ultimately, these relationships were all in favor of MST and the magnitude of the effect sizes between subgroups differed substantially, however, the confidence intervals between variables' respective subgroups overlapped a sizable amount and thus could not be deemed statistically significant from one another. The findings may suggest, at least preliminarily, that MST was more effective on psychopathology and mental health when studies were written in the latter portion of the 1900s, versus the early portion of the 2000s, published, rather than non-published, and when data collection began in 1980 or 1990, rather than 2000 or 2010.

Sample Characteristics. Table 3.23 presents the results comparing mean effect size estimates across a number of sample characteristics. Note that the only moderator to reach statistical significance was the type of youth characteristic. In particular, mean *r* and Z^+ for the conduct problems subgroup was significantly different from mean *r* and Z^+ for the mental health subgroup (*r* and $Z^+ = 0.21$ versus *r* and $Z^+ = 0.02$). This was evidenced by the lack of overlap between the conduct problems subgroup's CI and the mental health subgroup's CI ($CI_r = 0.09$-0.33 and $CI_{Z+} = 0.09$-0.34 versus CI_r and $CI_{Z+} = -0.05$-0.08). Interestingly, although mean *r* and Z^+ for the sex offender subgroup was not significantly different from conduct problems or mental health subgroups, the magnitude of the effect for the sex offender subgroup was substantially larger than that of the mental health subgroup ($r = 0.26$ and $Z^+ = 0.27$ versus *r* and $Z^+ = 0.02$).

Table 3.23
Mean Effect Sizes: Sample Characteristics–Psychopathology and Mental Health

	k	N	*r*	SD	95% CI	Z^+	95% CI
Type of Youth							
Conduct Problems	9	1,183	0.21	0.20	0.09–0.33	0.22	0.09–0.34
Mental Health	7	868	0.02	0.12	-0.05–0.08	0.02	-0.05–0.08
Sexual Offenders	2	179	0.26	0.20	-0.03–0.51	0.27	-0.03–0.57
Average Age							
Under 15	10	1,108	0.18	0.21	0.05–0.30	0.18	0.05–0.31
15 or Older	8	1,122	0.13	0.15	0.03–0.22	0.13	0.03–0.23
Race							
White (70.0% or more)	5	497	0.26	0.14	0.12–0.38	0.26	0.12–0.40
Non-White (70.0% or more)	6	702	0.12	0.09	0.04–0.20	0.12	0.04–0.20
Mix	7	1,031	0.09	0.24	-0.07–0.25	0.09	-0.07–0.25
At Least One Arrest							
No	3	1,144	0.10	0.23	-0.10–0.29	0.10	-0.10–0.30
Yes (70.0% or more)	8	435	0.22	0.17	0.07–0.36	0.22	0.07–0.38

Taken together, the results suggest that MST was most effective on psychopathology and mental health for sex offenders and youth with conduct problems. The effectiveness of MST did not seem to vary significantly between subgroups for the remaining characteristics (average age, race, and at least one arrest), as evidenced by overlapping confidence intervals about mean *r* and Z^+ for each of the variables' respective subgroups. General patterns, however, can be gleaned. Specifically, the magnitude of the unweighted and weighted effect sizes produced from studies where the average age was less than 15 was slightly larger than the magnitude of the unweighted and weighted effect sizes produced from studies where the average age was 15 or older (*r* and $Z^+ = 0.18$ versus *r* and $Z^+ = 0.13$). This result may suggest that MST marginally improved psychopathology and mental health for younger juveniles compared to older juveniles.

Similarly, the findings indicate that youths' psychopathology and mental health symptoms improved differentially across racial subgroups. That is, the mean effect size was highest for studies where the majority of sample participants were White (*r* and $Z^+ = 0.26$), followed by studies where the majority of the sample was non-White (*r* and $Z^+ = 0.12$), followed by studies where the racial composition was mixed between White and non-White participants (*r* and $Z^+ = 0.09$). Note that all three subgroups produced mean effect size estimates that were in favor of MST and that both White and non-White groups yielded estimates that were significantly different from what we would expect from chance alone. Finally, the magnitude of the effect size produced from studies where the majority of the sample (70% or more) had at least one arrest was more than two times larger than the estimate produced from studies where the majority of the sample did not have at least one arrest (*r* and $Z^+ = 0.22$ versus *r* and $Z^+ = 0.10$). Ultimately, these results may suggest that MST was more effective for "riskier" juveniles compared to less "riskier" juveniles on psychopathology and mental health measures.

Treatment and Study Characteristics. Table 3.24 presents the results comparing mean effect size estimates across treatment and study characteristics. In line with the discussion above, several factors were in favor of MST (i.e., unweighted and weighted effect sizes are in the right direction) and the magnitude of the effect sizes between subgroups seemed to differ substantially; however due to the small number of effect sizes per category and in total, the widths of all confidence intervals around subcategories' *r* and Z^+ values overlapped considerably. In this way, the effectiveness of MST in improving youths' psychopathology and mental health symptoms did not appear to vary significantly between treatment or study characteristics. Nonetheless, several relationships are worth noting.

The magnitude of the effect differed considerably between comparison group subgroups. Specifically, the results indicate that in predicting rates of improvement in MST participants' psychopathology and mental health symptoms, the mean effect size generated from studies where the comparison group received a single treatment was greater than the mean effect size generated from studies where the comparison group received multiple treatment/services (*r* = 0.37 and $Z^+ = 0.35$ versus *r* and $Z^+ = 0.12$). Related to the source from which youths were referred to treatment, the findings suggest that MST may be slightly more effective for youths referred from criminal justice agencies (*r* and $Z^+ = 0.19$), followed by non-criminal justice agencies (*r* and $Z^+ = 0.11$), and then a mix of social service and criminal justice referral sources (*r* and $Z^+ = 0.03$).

In addition, larger effects for MST on psychopathology and mental health were generated from studies where 80 percent or more of the MST group completed treatment, compared to studies where less than 80 percent of the MST group completed treatment (*r* = 0.19 and $Z^+ =$ 0.18 versus *r* and $Z^+ = 0.11$). The amount of treatment MST youths received also appeared to be

a contributing factor to improvement in psychopathology and mental health-related behaviors. That is, the magnitude of the effect was larger for studies in which the average number of treatment days was 150 days or greater, compared to studies in which the average number of treatment days was less than 150 days (*r* and $Z^+ = 0.19$ versus *r* and $Z^+ = 0.11$). In short, the findings indicate, at least preliminarily, that MST was more effective on psychopathology and mental health when the majority of youths completed treatment and related, when they received more treatment (i.e., dosage), compared to youths who did not complete treatment and received less treatment.

The results in Table 3.24 also suggest that larger effects on youths' psychopathology and mental health symptoms were associated with studies conducted at a single site during the treatment period, compared to those studies conducted across multiple sites during the treatment period (*r* = 0.37 and $Z^+ = 0.35$ versus *r* and $Z^+ = 0.12$). In addition, larger effect sizes were generated from quasi-experimental studies, compared to randomized-controlled studies *r* and $Z^+ = 0.24$ versus *r* and $Z^+ = 0.13$). Finally, efficacy trials continued to outperform effectiveness studies on psychopathology and mental health outcomes (*r* and $Z^+ = 0.24$ versus *r* and $Z^+ = 0.11$).

Also related to Table 3.24, when there was no indication that adherence was measured, MST appeared to impact youths' psychopathology and mental health problems at a greater rate; however, in line with previous discussions, the majority of studies falling into this subgroup were early studies, conducted by Henggeler and the MST developers, and/or before the TAM instrument became a formal requirement in the MST protocol. Finally, the results indicate that in predicting psychopathology and mental health, the mean effect size for studies where the average follow-up period was 13 months or longer was larger than the mean effect size for studies where the follow-up period was 12 months or less (*r* and $Z^+ = 0.19$ versus *r* and $Z^+ =$

Table 3.24
Mean Effect Sizes: Treatment and Study Characteristics–Psychopathology and Mental Health

	k	N	*r*	SD	95% CI	Z^+	95% CI
Comparison Group Received							
Single (i.e., treatment *or* services)	2	235	0.37	0.42	-0.29–0.79	0.35	-0.30–1.07
Multiple (i.e., treatment/services)	16	1,995	0.12	0.15	0.05–0.19	0.12	0.05–0.19
Source of Clients for Treatment							
Non-Criminal Justice Agency	7	770	0.11	0.14	0.01–0.22	0.11	0.01–0.22
Criminal Justice Agency	10	1,204	0.19	0.22	0.06–0.31	0.19	0.06–0.32
Mix	1	256	0.03	--	-0.09–0.15	0.03	-0.09–0.15
Program Completion—MST							
No	7	317	0.11	0.08	-0.03–0.24	0.11	-0.03–0.24
Yes (150 days or more)	5	1,798	0.17	0.19	0.07–0.27	0.18	0.07–0.36
Average Number of MST Treatment Days							
No	7	431	0.11	0.12	0.01–0.21	0.11	0.01–0.21
Yes (150 days or more)	5	518	0.19	0.18	-0.01–0.37	0.19	-0.01–0.38
Adherence to MST Model							
Appeared Adherence was Measured	14	1,796	0.12	0.15	0.04–0.20	0.12	0.04–0.20
No Indication Adherence was Measure	4	434	0.26	0.27	0.00–0.48	0.26	0.00–0.53

Table 3.24
Mean Effect Sizes: Treatment and Study Characteristics–Psychopathology and Mental Health

	k	N	*r*	SD	95% CI	Z^+	95% CI
Number of Study Sites							
Single-Site	13	1,483	0.14	0.21	0.03–0.25	0.14	0.07–0.28
Multi-Site	5	747	0.17	0.11	0.07–0.28	0.18	0.03–0.25
Study Design							
Randomized	14	1,737	0.13	0.14	0.06–0.20	0.13	0.06–0.20
Quasi-Experimental	4	493	0.24	0.30	-0.07–0.50	0.24	-0.07–0.55
Study Setting							
Efficacy Trial	6	664	0.24	0.25	0.04–0.42	0.24	0.01–0.04
Effectiveness Study	12	1,566	0.11	0.14	0.03–0.19	0.11	0.03–0.19
Average Follow-Up Period							
0—12 Months	15	1,944	0.14	0.18	0.06–0.23	0.14	0.06–0.23
13—25^+ Months	3	286	0.19	0.21	-0.05–0.41	0.19	-0.05–0.44

0.14). As noted, however, the CIs of the respective *r* and Z^{+} values for all of these relationships overlapped considerably and as such, should not be considered precise.

CONCLUSION

The present chapter presented the results of several analyses aimed at evaluating whether MST improved outcome overall for nine dependent variables and moreover, what factors or circumstances led to the strongest treatment effects for the program. Significant main effects were demonstrated for six of the nine outcomes, including: (1) delinquency, (2) problem behavior, (3) psychopathology and mental health, (3) family functioning and relationships, (5) parent functioning and relationships, and (5) service utilization. Although the results were in favor of MST for the remaining three outcomes (peer relationships, school performance, and substance abuse), these relationships failed to reach statistical significance. The findings also suggested that MST effectiveness is moderated by a number of characteristics, including: publication type, involvement of the evaluator, source of clients for treatment, type of youth, race, previous arrest, study setting, and adherence to MST model. The next chapter summarizes the findings, discusses the limitations of the current study, and outlines potential policy and practical implications.

CHAPTER 4

DISCUSSION

MST was invented in 1978 when Scott Henggeler and his graduate students at Memphis State University were asked to develop a diversion program for juvenile offenders (Cullen, 2005). At the time of the program's founding, Martinson's (1974) nothing works doctrine continued to ring loud and clear throughout the correctional atmosphere, making the already daunting task of creating an intervention program for at-risk youths a very treacherous feat. While criminologists continued to "declare rehabilitation dead" (Cullen & Jonson, 2012, p. 35), psychologists, such as Henggeler, declared that rehabilitation *could* be alive if based on the correct model or theoretical underpinnings (Cullen, 2005).

Notably, treatment models during the 1970s were rooted in psychodynamic theory, rather than sociology (Cullen & Jonson, 2012). As such, clients were prompted to go back in time to determine how past events and actions could have caused or contributed to the current problems they were experiencing (Burns et al., 2000). With a background in community psychology, Henggeler understood, however, that "most kids in the justice system not only were personally troubled but also were drawn from troubled contexts or *systems*" (emphasis in the original) (Cullen & Jonson, 2012, p. 195). In this way, the multiple systems in which they were enmeshed had to be addressed if rehabilitation was going to "work."

Departing from traditional psychotherapeutic approaches, the MST theory of change is consistent with Brofenbrenner's (1979) theory of social ecology in that it targets youths' observable behavioral problems, rather than focusing on unconscious forces from the past (Van Voorhis & Braswell, 2007). The intervention strategy, more specifically, assumes that antisocial

behavior is driven by the interplay of individual risk factors associated with the multiple systems in which youths are embedded. In this context, antisocial behavior can only be changed when interventions take into account the factors that are driving the youth to engage in a particular antisocial behavior, as well as the various social contexts in which they dwell. In addition, the theory argues that youths' parents/caregivers are the main conduits for change. Thus, parents/caregivers are seen "as full collaborators in treatment" (Henggeler et al., 2009, p. 25) and the family network becomes a central focus for achieving treatment success.

Since its inception, MST has expanded exponentially. In particular, the treatment strategy has been implemented in 34 states and 15 countries, and has treated more than 23,000 youths per year (MST Services, Inc., 2016). Given the program's dissemination across the United States and abroad, it is important to understand its overall treatment effects. Although many individual studies and a small number of comprehensive reviews have been carried out to examine MST's effectiveness in changing juvenile behavior, the results have been mixed and have recently called attention to the importance of moderating factors or characteristics that may help to explain why we are seeing such mixed results. In this context, the overall goal of the current study was to advance the literature on the effectiveness of MST and determine more in depth what factors or circumstances lead to the strongest treatment effects for the program. The objectives of the dissertation, more specifically, were threefold.

First, this dissertation involved the analysis of 44 unpublished and published studies, providing the most comprehensive review of studies on the effectiveness of MST to date. Second, the effectiveness of MST for youths and their families on nine dependent variables was examined, including: (1) delinquency, (2) problem behavior, (3) psychopathology and mental health, (4) family functioning and relationships, (5) peer relationships, (6) school performance,

(7) parent functioning and relationships, (8) substance abuse, and (9) service utilization. Third, to determine the effectiveness of MST on the nine outcome variables, a meta-analysis was conducted. From this quantitative synthesis process, the overall mean effect size, weighted mean effect size, and corresponding confidence intervals were calculated for the independent variable (i.e., MST) and each dependent variable under review. In addition, the impact of several moderating variables was explored. Moderators were examined across five general categories, including: (1) publication characteristics, (2) study context characteristics, (3) sample characteristics, (4) treatment characteristics, and (5) study characteristics. The current chapter summarizes the major findings presented in the Results chapter and subsequently uses the major theory of correctional treatment—risk-need-responsivity as prism to illuminate why MST does not have larger effects despite being a carefully designed intervention.

DOES MULTISYSTEMIC THERAPY WORK?

Table 4.1 presents a summary of the main effect size estimates as they relate to MST and the nine dependent variables explored in the current dissertation. In general, the findings suggest that MST worked across all outcome measures (i.e., all point estimates were in the direction favoring the MST group over the comparison group). Specifically, small, but significant effects were found for MST on six outcomes, including: (1) delinquency, (2) problem behavior, (3) psychopathology and mental health, (4) family functioning and relationships, (5) parent functioning and relationships, and (6) service utilization. In addition, the findings indicate that MST had a nominal effect on youths' peer relationships, school performance, and use of illicit substances; however, these results did not reach statistical significance. The following section summarizes the main effects of the current study and compares them to the findings demonstrated in previous systematic reviews that employed different coding schemes and

methodologies. The section also discusses the influence of heterogeneity in effect sizes and the results generated from moderator analyses from the current study.

A Summary of the Main Effects and Comparison of the Results with Previous Meta-Analyses

As discussed above, modest, but significant treatment effects were found on the primary outcome delinquency. As such, Table 4.1 suggests that MST reduced delinquent behavior, on average, by 11 percent over comparison group treatment/services. The estimate differs from earlier meta-analyses that found slightly larger effects on delinquency (e.g., Curtis et al., 2004; Farrington & Welsh, 2003). In particular, Farrington and Welsh (2003) found that the average effect of MST on delinquency was $d = 0.41$, or an average 20 percent decrease in recidivism for the MST group compared to the comparison group, in their systematic review of nearly 40

Table 4.1
Summary of the Effects

Outcome	*k*	Z^+	95% CI
Delinquency	26	0.11	0.05–0.17
Problem Behavior	19	0.15	0.05–0.25
Psychopathology and MH	18	0.15	0.07–0.23
Family Functioning and Relationships	14	0.14	0.01–0.27
Peer Relationships	9	0.09	0.00–0.17
School Performance	9	0.16	-0.13-0.45
Parent Functioning and Relationships	14	0.15	0.05–0.26
Substance Abuse	8	0.04	-0.10–0.17
Service Utilization	10	0.21	0.14–0.35

family-based crime prevention program studies. A somewhat larger estimate was generated when Curtis and her colleagues (2004) conducted a meta-analysis of MST almost one year later. In their study, the average effect of MST was $d = 0.55$, or an average reduction in delinquent behavior of 27 percent for the MST group compared to the comparison group. Recall, however, that both Farrington and Welsh (2003) and Curtis et al.'s (2004) systematic reviews were conducted with a small number of independent published studies, the majority of which were efficacy trials conducted by Henggeler and/or other MST developers ($k = 8$ and $k = 11$, respectively). In this way, there may have been variability in the effects generated from earlier meta-analyses because the current meta-analysis included more recent studies that were both published and unpublished, as well as conducted by dependent *and* independent authors.

The average 11 percent reduction in delinquent behavior found in the current study also differs from Littell et al.'s (2005) previous meta-analysis of MST. In this case, however, the effect was larger than that reported by these authors. Recall that in Littell and colleagues' (2005) study, rigorous methods and statistical techniques prevented the inclusion of over 20 studies that had been written between 1985 and 2003; the time period in which the reviewers included for data collection. Ultimately, these restrictive criteria led them to conclude that MST was not consistently more effective than other juvenile programs.

Despite the mixed findings discussed above, the current study's results were in line with those demonstrated in the most recent synthesis of MST research conducted by the Netherlands scholars (van der Stouwe et al., 2014). In their study, more specifically, the average effect for MST on juvenile delinquency was $d = 0.20$, or an average reduction in delinquency of 10 percent for the MST group over the comparison group. More generally, the results of the current study indicate that MST is about as effective as correctional programs overall. For example, in a meta-

analysis examining whether interventions following the core principles of effective intervention were more effective at reducing rates of recidivism, compared to those interventions that did not follow these core principles, Andrews and his colleagues found an effect size of $r = 0.10$ across all programs (Andrews, Zinger, Hoge, Bonta, Gendreau, & Cullen, 1990). As further evidence, Lösel (1995) found that the mean effect size of studies ranged between $r = 0.05$ and $r = 0.18$ in his review of 13 meta-analyses evaluating offender rehabilitation programs. When taken together, his results indicated that "the mean effect size of all assessed studies" was "probably about 0.10" (Lösel, 1995, p. 89). Finally, when comparing the effect of MST on delinquency from the current study to the results demonstrated in Lipsey and Wilson's (1998) meta-analysis evaluating the efficacy of correctional intervention on serious juvenile offenders, the findings indicate that MST did slightly better than juvenile interventions in general ($d = 0.12$ or $r = 0.06$).

As mentioned in the introduction to this section, several modest, but significant effects were also found for secondary outcomes in the current study. Specifically, Table 4.1 suggests that MST reduced problem behavior by 15 percent, as well as improved psychopathology and mental health symptoms by 15 percent, family functioning and relationships by 14 percent, parent functioning and relationships by 15 percent, and service utilization by 21 percent. The table also indicates that although treatment effects were in favor of MST for peer relationships, school performance, and substance abuse, the results failed to reach statistical significance. Note, however, that the number of effect sizes for each of these categories was limited, which may explain why the width of the confidence intervals was substantially large, and subsequently, why the point estimates were not deemed statistically significant.

Generally, the results yielded from secondary analyses in the current study align with those demonstrated in Curtis et al. (2004) and van der Stouwe et al.'s (2014) systematic reviews

of MST. That is, both studies produced results that were in favor of MST across all outcomes included for analysis.[4] In the early evaluation conducted by Curtis et al. (2004), however, larger average effect sizes for family relationships were found for what they called, individual adjustment variables (e.g., youth symptoms, parent symptoms, youth behavior problems, and hospitalization). In addition, measures of family relations demonstrated larger average effect sizes compared to measures of individual adjustment. By contrast, Van der Stouwe and colleagues (2014) did not find larger effect sizes for family relationships over individual adjustment variables and even further, the effect of MST on peer relationships "diminished altogether" (2014, p. 476).

More in line with the Netherland's study, the current study found similar nominal effects for MST on individual adjustment outcomes, as well as outcomes related to family and peer relationships. One caveat to this in the current study was that the mean effect size for service utilization was slightly larger than the estimates produced for the other outcomes. When examining Table 4.1 further, however, it can be seen that all associated confidence intervals about the point estimates overlapped, suggesting that none of the effect size estimates related to secondary outcomes (or the primary outcome, for that matter) were significantly different from one another. In short, MST worked to change a variety of juvenile behaviors, as evidenced by the results of the current study that produced minimal, but significant treatment effects on most outcomes. Importantly, these results are in line with the overall goals of the MST treatment philosophy—to provide treatment to youths targeting their individual risk factors, as well as factors associated with the multiple systems in which they are enmeshed.

[4] Note, one outcome examined in van der Stouwe et al.'s (2014) study that was not examined in the current study (skills and cognitions) yielded negative treatment effects; however, this relationship was not statistically significant.

Summary of the Moderating Effects

The subsection above demonstrates that MST was effective in changing a variety of juvenile behaviors; however, the magnitude of the mean effect sizes varied. The results generated from independent studies across all outcomes of interest were substantially heterogeneous. For example, across the 26 studies examining MST's effectiveness on delinquency, individual effect sizes ranged from -0.11 to 0.49. Even across the small number of studies examining MST's effectiveness on school performance ($k = 9$), the distribution of the effects ranged from -0.59 to 0.57. When outliers were removed from the distribution, significant variation in many of the outcomes' distributions still remained, which suggested that systematic differences between cases (i.e., studies) existed.

Heterogeneity in effect sizes may occur for two reasons. First, differences may be generated because of the methodologies employed in individual studies. Second, program-related characteristics may attribute to variability in effect sizes. For instance, the treatment modality, length of treatment, target population, and/or quality of the treatment may be different across studies (Cullen and Gendreau, 2000). In line with this discussion, the current study examined a number of publication, study context, sample, treatment, and study characteristics that could potentially explain the observed variability across all outcomes, as well as for the primary outcome delinquency and secondary outcome psychopathology and mental health. From these analyses, several important variables were found that appeared to condition (Cullen & Gendreau, 2000) the magnitude of the mean effect size.

Note that due to the limited number of effect sizes in several categories and/or considerable differences between the number of effect sizes between subgroups, the widths of the confidence intervals associated with mean effect sizes for the majority of moderators

explored in the analyses related to delinquency and psychopathology were considerably wide (i.e., imprecise). In this way, several moderating characteristics on delinquency and psychopathology and mental health did not reach statistical significance. The following discussion, however, will focus on a general pattern of findings that are both practically and theoretically relevant.

Methodological Considerations. First, several characteristics related to the publication and context in which the study was conducted were associated with delinquency and psychopathology and mental health. For example, greater effects were associated with published studies versus unpublished studies. This trend is not surprising, given that research has shown that published research tends to favor significant results while unpublished research tends to find nonsignificant or null results (Glass et al., 1981; Rothstein, Sutton, & Borenstein, 2005). Notably, the difference in the magnitude of the mean effect size for studies conducted in the U.S. versus studies conducted outside of the U.S. approached significance when examining all outcomes together; however, the point estimates between subgroups were fairly similar when examining this relationship for delinquency studies and psychopathology and mental health studies, respectively. This finding differs from van der Stouwe and colleagues' (2014) study that found larger effects for studies carried out in the United States. Specifically, their study suggested that the transport of MST to other countries may not be as effective as others have suggested (e.g., Ogden et al., 2008). As such, an investigation using a larger sample of MST evaluations will be needed in the future to consider this issue further.

Studies where data collection began earlier (1980 or 1990) and thus that were written earlier also yielded larger mean effect size estimates compared to studies where data collection began later (2000 or 2010) and thus were written later. The study setting (efficacy trial versus

effectiveness study) treatment characteristic was also associated with greater effect size estimates, which may be related to the time period in which data collection began and ultimately, when the study was written. The current study found, in particular, that efficacy trials were significantly more effective than effectiveness studies on outcome. While these relationships did not reach statistical significance when examining the effect of MST on delinquency and on psychopathology and mental health separately, the same pattern held true (i.e., larger point estimates were associated with efficacy trials compared to effectiveness studies).

Recall from Chapter 1 that Henggeler and his team conducted several trials of MST that helped to establish support for the program early on. These efficacy trials set the stage for later effectiveness studies, as well as for research on the transportability and dissemination of MST. Note that these studies not only were conducted in the late 1900s but also were conducted in "research friendly" environments where factors that typically pose challenges for researchers were not an issue. In addition, close clinical supervision and oversight was provided to graduate students who served as therapists. Thus, it is not surprising that moderating variables related to publication decade, decade data collection started, and study setting were strongly associated with differences in the magnitude of effect sizes.

The results from the current meta-analysis also suggest that the average effect size generated from studies where MST adherence was *not* measured was larger than the average effect size generated from studies where MST adherence was measured. This relationship was found across all outcomes and for delinquency and psychopathology and mental health, respectively. There is likely an explanation for this curious finding. Specifically, this study coded for whether adherence to the MST model was measured that went beyond the normal QA/QI process required of all provider organizations (e.g., audiotapes with feedback, use of the

TAM instrument; protocols that were developed after the first clinical trials of MST). In line with the above discussion, then, studies where it did not "appear" that adherence was measured were almost all written in earlier decades (1980 or 1990) and thus were efficacy trials where Henggeler and the MST development team were involved in evaluation, implementation, and quality assurance. To this end, there might have been *more* fidelity checks in place for these studies compared to other studies, even though it was not explicitly stated in the article that adherence was measured.

The relationship between treatment effectiveness and outcome also seemed to be moderated by whether the evaluator of the study was involved in some additional capacity (e.g., technical assistance, quality assurance). Across all outcomes, studies where the researcher(s) was involved in the implementation of MST (in addition to being the evaluator[s]) yielded an average effect size estimate that was significantly larger than the estimate yielded from studies where the researcher did not play an additional role in the study's implementation. This particular moderator was not significant for delinquency or psychopathology and mental health, but the general pattern of findings was seen in the analyses related to both outcomes.

When investigating the relationship between treatment and outcome when the MST developers were involved in the study, the results indicated that studies had a marginally larger effect, on average, when Henggeler and/or his team was involved in the evaluation and/or implementation of the project. However, this relationship did not reach statistical significance. The same results were found when delinquency and psychopathology and mental health outcomes were examined on their own. These findings are in line with previous research suggesting that "demonstration" programs (i.e., researcher-involvement) yielded better results compared to "practical" programs (i.e., non-researcher-involvement) (Lipsey, 1999). However,

it appears that having the evaluator involved in general, rather than the evaluator being Henggeler and/or other MST developers, was the relationship that mattered the most.

Program Considerations. The results of the current study also indicate that particular sample and treatment characteristics moderated the relationship between MST and outcome. Specifically, treatment effects on both delinquency and psychopathology and mental health were highest when MST was conducted with sexual offenders and youths manifesting conduct problems (i.e., juvenile offenders, youths diagnosed with conduct disorder or oppositional defiant disorder, and/or youths at-risk for delinquency). In contrast, the results indicated that MST was effective for youths suffering from psychological and mental health disorders (i.e., youths with serious emotional disturbances or mental health issues, substance-abusing or substance-dependent youths, maltreating families, and dual-diagnoses); however, these effects were minimal compared to those demonstrated for justice-involved youths. Related, studies where the majority of the sample had been arrested at least once prior to referral to MST treatment yielded larger effect size estimates compared to studies where the majority of the sample had not been arrested at least once prior to referral to MST treatment.

As further evidence to the discussion above, differential treatment effects were found when youths were referred to MST via a criminal justice agency versus a non-criminal justice agency or a mix of social service and criminal justice agencies. That is, studies where youths were referred to MST via a criminal justice agency had lower rates of delinquency and improved psychopathology and mental health outcomes compared to studies where youths were referred to MST via a non-criminal justice agency or a mix of agencies. In addition, larger mean effect sizes were associated with studies where the majority of participants completed treatment and studies where the majority of participants received at least 150 days of treatment.

Two other sample characteristics appeared to moderate the relationship between MST and delinquency and MST and psychopathology and mental health—race and average age. Specifically, MST was significantly associated with larger effects on delinquency when the majority of studies' samples were White, followed by studies where the racial composition of the samples was mixed between White and non-White youths. The program, on the other hand, appeared to be least effective on delinquency when the majority of the studies' samples were non-White. While limited cell frequencies likely prevented race from reaching statistical significance on the outcome of psychopathology and mental health, disparate effects were found between subgroups for this relationship as well. Similar to delinquency, it appeared that improvement on psychopathology and mental health issues were greatest when the majority of the sample was White. In contrast to the delinquency outcome, however, it appeared that MST was least effective on psychopathology and mental health when studies' samples were mixed between White and non-White youths. As discussed in the Results chapter, it is important to note that these findings should not be taken to suggest that MST was ineffective when the sample was predominantly non-White or a mix of White and non-White youths. In the majority of cases, the mean effect size and weighted mean effect size estimates were statistically significant (i.e., better than what we would expect by chance) and in all cases, in the right direction (i.e., MST participants had greater rates of improvement, compared to comparison group youths). These results were also demonstrated in van der Stouwe et al.'s (2014) synthesis of the research, which may indicate that certain cultural barriers exist that prevent studies from yielding similar results across racial subgroups.

Notably, it appeared that the point estimates produced from studies where the majority of the studies' samples were White were very similar between the primary outcome and the

secondary outcome. In other words, the magnitude of the effect sizes for the White subgroup were almost identical for delinquency and psychopathology and mental health (r and $Z^+ = 0.25$ and r and $Z^+ = 0.26$, respectively). This was also the case for the mixed subgroup. Specifically, when the racial composition of studies' samples was mixed between White and non-White youths, the average effect on delinquency was 0.10 and almost identical for the average effect on psychopathology and mental health (r and $Z^+ = 0.09$). The estimates between delinquency and psychopathology and mental health for the non-White subgroup, on the other hand, were substantially different from one another, approaching statistical significance (r and $Z^+ = 0.01$ and r and $Z^+ = 0.12$, respectively). In short, the results may indicate that MST was equally as effective for White and mixed subgroups on delinquency and psychopathology and mental health, but more effective for the non-White subgroup on psychopathology and mental health compared to delinquency.

Average age of the youths participating in MST also appeared to moderate the relationship between treatment effectiveness and delinquency and psychopathology and mental health outcomes. Specifically, larger estimates were generated from studies where the sample of participants were younger (i.e., less than 15 years of age), compared to estimates generated from studies where the sample of participants were older (i.e., 15 years of age or older). It is important to note that this finding was consistent with the Netherland's study (van der Stouwe et al., 2014).

From the current section, we can conclude, that greater effects for MST on outcome are associated with certain methodological considerations, as well as particular program characteristics. Even after examining potential moderating effects, however, the results suggest

that MST is ineffective under some conditions and highly effective under other conditions. In this way, the next section discusses why differential effects continue to be generated.

ASSESSING THE EFFECTS OF MULTISYSTEMIC THERAPY: IMPLICATIONS FROM THE RNR MODEL

The current section attempts to "make sense" of why the effects of MST are modest and explores potential reasons why there is heterogeneity in the estimates discussed above. One way to approach this issue is to use the major theory of correctional treatment—the risk-need-responsivity (RNR) model—as a prism to illuminate why MST does not have larger effects, despite being a carefully designed intervention. The following subsections briefly describe the RNR model and also discuss how MST conforms (and does not conform) to these principles.

The Risk-Need-Responsivity Model

Research is clear that some correctional programs are more effective at reducing recidivism and changing offender behavior than others (Latessa, 1999). Thanks to a group of Canadian psychologists, research is also clear, that agencies that follow the principles of effective intervention (i.e., RNR) when developing and implementing programs, have a higher chance of success than those that do not (Andrews, et al., 1990; Gendreau, Smith, & French, 2006; Lipsey, 1999). Importantly, these findings have been replicated across various correctional populations, including juvenile offenders (see e.g., Andrews et al., 1990; Dowden & Andrews, 1999a, 1999b, 2000). To set the context for the next subsection, a brief overview of the RNR model is provided below.

First, the risk principle states that interventions should be used primarily with higher-risk offenders and that the intensity of treatment and level of services must be matched to this level of risk. As such, agencies should devote the majority of resources to higher risk offenders,

targeting their criminogenic needs or dynamic risk factors that can be changed (Lowenkamp & Latessa, 2004). The risk principle is especially important in that it is the first step in the correctional or juvenile justice process—that is, it helps agencies decide *who* to target for change (Lowenkamp & Latessa, 2005). In this way, "agencies must assess offenders with standardized and objective risk assessment instruments" (Lowenkamp & Latessa, 2004, p. 8) (e.g., the Youth Level of Service / Case Management Inventory) and subsequently, supervision and treatment should be provided based on level of risk.

Once we understand *who* to target for change, we must determine *what* factors about the individual are causing them to commit crime or driving them to behave antisocially. In this way, the second principle in the RNR model is the need principle which states that interventions should target for change offenders' criminogenic needs or the known predictors of crime (Andrews & Bonta, 2010). Criminogenic needs include, for example, antisocial personality factors (e.g., impulsivity, risk-seeking), antisocial attitudes, values, and beliefs, antisocial peers, substance abuse, leisure activities, education/employment, and family, and when identified through the use of actuarial risk assessments, "have a role in preventing, rather than simply predicting, offending." (Hannah-Moffat & Maurutto, 2003, p. 2). Along with prior criminal history, Andrews and Bonta (2010) call the characteristics listed above the "Central Eight," with the first four or the "Big Four" having the greatest impact on recidivism, and the second four having a slightly less, but still impactful relationship with future criminal behavior.

Finally, the responsivity principle tells us *how* to intervene with offenders so that we "maximize [their] ability to learn from a rehabilitative intervention" (Bonta & Andrews, 2007, p. 1). There are two parts to the responsivity principle. The general responsivity principle indicates that behavioral interventions are effective in changing a variety of criminal and non-

criminal behaviors. In addition to using social learning techniques (e.g., modeling, graduated practice, role playing, reinforcement, and cognitive restructuring), behavioral or cognitive-behavioral therapy (CBT) interventions have been shown to be the most effective types of interventions. As such, treatments that are behavioral in nature have the ability to target and change for the better the variety of criminogenic needs underlying recidivism. The specific responsivity principle, on the other hand, states that interventions should be tailored to match individuals' learning styles, skills and abilities (i.e., strengths), and motivation to change (Andrews & Bonta, 2010).

MST's Compliance with the RNR Principles

While MST was developed outside of the Canadians' RNR paradigm, it is loosely aligned with some of its core features—a reason why it has likely garnered support over recent years. On the other hand, "loosely aligning" with the principles of effective intervention, rather than conforming to them wholeheartedly or being developed in line with them initially, may explain why there is variability across individual studies and/or why moderating effects do not "tell us the whole story." In this way, the following subsection evaluates MST through the lens of the RNR model and highlights how it implicitly meets each principle of effective intervention, while at the same time falling short.

Recall that the risk principles tells us *who* to target for change. That is, interventions should be used with higher risk offenders targeting their criminogenic needs and subsequently, provide "increasing levels of service as risk level increases" (Makarios, Sperber, & Latessa, 2014, p. 334). MST meets the risk principle in that the program was originally designed to target high-risk youths displaying "serious behavior disorders" and/or "violent and chronic juvenile offenders" (Henggeler, 1997, pp. 1, 6) at "high-risk of out-of-home placement" (Henggeler,

1998, p. 3). In addition, each clinician within the MST team carries a small caseload that allows them to provide intensive services to each family (Bernfeld et al., 2001). Families receive, on average, between 40 and 60 hours of direct clinical contact over the course of a three- to five-month timeframe (Burns et al., 2000; Henggeler et al., 2009) and sessions can be held as often as daily when the treatment initially begins and then takes place less often over the course of treatment (Burns et al., 2000). These features of the program suggest that the model attempts to match intensity of treatment to level of risk.

Notably, although preliminary, this systematic review found evidence to support the risk principle and MST treatment. Specifically, several factors that might be viewed as indicators of risk were related to MST achieving more success. Thus, stronger treatment effects were associated with justice-involved youths (e.g., sex offenders and conduct problem youths) compared to youths suffering from psychological and mental health disorders, as well as for youths who were arrested at least one time prior to being referred to MST services and youths who received at least five months of treatment or more.

The difficulty, however, is that although designed for high-risk youths, the MST model does not ensure that those receiving the treatment are, in fact, all high risk. The failure to clearly identify the risk levels of subjects in evaluation studies may account in part for the heterogeneity in effect sizes that are manifested. Thus, it is now well established that the most effective strategy for assessing an individual's risk to reoffend is to use an actuarial-based assessment instrument. Research has shown not only in corrections but also in other fields related to human behavior that the incorporation of validated instruments that use quantitative data to make informed decisions are able to predict behavior more accurately, compared to relying on clinical judgment alone (Ayres, 2007; Cullen & Jonson, 2012). While the MST model requires that

therapists use a series of conceptual aids during the intervention process, including a background information form, strengths and needs assessment, case summary for supervision and consultation, and fit circles, these documents are not validated nor are they based on actuarial data.

Once *who* we target for change is identified, *what* we target for change must be determined. That is, we must identify criminogenic needs (i.e., dynamic factors that can change) and target them through treatment and intervention. Instructively, MST meets the need principle in that clients are assessed individually with a focus on addressing the known causes of delinquency (Henggeler, 1998), and subsequently the therapist spends time with the youth and their family to link these risk factors to each contextual system the youth and his family are enmeshed in (e.g., family, school, peers, individual, and neighborhood) (Henggeler & Schoenwald, 1998). The treatment model deviates from the need principle, however, in that once again a validated, actuarial risk assessment is not used to identify these factors. Further, when risk factors are identified, they are grouped by the systems in which the youth is enmeshed, rather than placing precedence on one set of factors over another.

Recall that the RNR model argues that some needs are more salient than others, as they have been found to have a greater impact on recidivism, compared to the "moderate four" needs. These factors include criminal history, antisocial personality, antisocial attitudes, values, and beliefs, and antisocial peers. If you think about a dandelion growing in your lawn, for example, the "moderate four" needs are the factors that can be seen when you walk through your yard; the ugly flower that is easy to pull so that it is not such an eye sore any more. What happens, then, when you just pull the dandelion out of the ground and do not attend to the root? It grows back! In this sense, the "Big four" criminogenic risk factors are like the dandelion in your lawn;

difficult to identify and difficult to change because they are not always tangible. Take, as a further example, a youth who has poor attitudes about going to school. They do not value an education, think selling drugs on the street corner is a better use of time than sitting in a classroom all day, and ultimately, do not show up for school three or four days out of the week. School in this scenario is the "dandelion"—the criminogenic need that we can see waving loud and clear in front of us. We think that if we can just get the youth back into school, he or she will clean up his or her behavior, but we are forgetting about the "root" (i.e., his or her poor attitudes, values, and beliefs about going to school in the first place).

In line with the examples above, it appears that in the MST model, therapists are identifying and focusing on many of the bottom-tier criminogenic needs because they are more apparent and easier to change, rather than focusing on the more difficult need areas that tend to have the greatest impact. Ultimately, it is not that MST has identified irrelevant risk factors—they are all empirically demonstrated—but that it encourages therapists to look at a wide range of factors that are not prioritized correctly. Even further, because an actuarial risk assessment is not required, it is unlikely that MST interventions are based on youths' most risky behaviors or criminogenic needs.

Finally, recall from the previous subsection that the responsivity principle tells us *how* offenders respond to treatment and focuses on delivering services in a manner, that have been shown through research, to change behavior (general responsivity) and takes into account individual differences (individual responsivity). One of MST's biggest areas of strength is that it meets both parts of the responsivity principle. Specifically, MST meets the general responsivity principle in that it is rooted is social psychological theory and is empirically based. In this manner, "MST interventions integrate techniques from those pragmatic and problem-focused

child psychotherapy approaches that have at least some empirical support, including pragmatic family therapies (e.g., strategic, structural, behavioral family systems approaches), cognitive-behavioral techniques, and behavioral parent training" (Schoenwald et al., 1999, p. 5). In addition, the interventions are present-focused and action-oriented, targeting specific and well-defined problems. From an individual responsivity perspective, MST provides services to youths and their families that are individualized to meet their unique needs as much as possible and "brings the intervention to the troubled youngster" (Cullen & Jonson, p. 196) by going to their home, school, or other community setting.

While MST integrates intervention strategies that are in line with CBT and social learning approaches, it does not appear that there is a set curriculum or set of "model programs" that therapists chose from when targeting criminogenic needs. In this manner, one of its biggest strengths—that treatment interventions can be chosen that are responsive to changing targeted risk factors for each individual youth and their family—can also be one of its biggest detriments because of the flexibility afforded to therapists.

CONCLUSION: THE FUTURE OF MULTISYSTEMIC THERAPY

MST is a carefully designed program that is used widely and, according to the current dissertation and other systematic reviews, reduces delinquency and problem behavior, while at the same time, improving a variety of other juvenile behaviors. It is modestly successful overall, and in some instances more than modest effects have been demonstrated. The challenge for MST is how to proceed in the future. This issue is all the more salient because the inventor of MST, Scott Henggeler, is near the end of his career.

One promising avenue is for MST to embrace the RNR model and create a new version that is a hybrid of the two. This would be done by incorporating the results of validated actuarial

risk assessment instruments to determine *who* is most appropriate for treatment and to ensure that the original intent of the program—to provide treatment to high-risk youths—is not lost in translation. The incorporation of risk assessment information could be achieved by requiring the referring agency to provide the results of the assessment or by ensuring therapists are trained and certified to conduct the risk assessment tool when the results from referring agencies are unavailable. The use of actuarial-based risk assessment instruments will not only help therapists to identify *what* criminogenic needs are driving youths' antisocial behavior, but also help to ensure that identified factors are targeted in a way places precedence on the "Big Four" criminogenic needs. Finally, it is suggested that MST Services, Inc. build a library of treatment strategies and "model programs" to ensure that the general responsivity principle is followed more closely. This library will allow the program to remain flexible, while at the same time making sure that criminogenic needs are targeted in the most effective way possible across all implementation sites.

REFERENCES

*References marked with an asterisk (*) indicate studies included in the current meta-analysis.

Achenbach, T. M. (1991). Manual for the Child Behavior Checklist and 1991 profile. Burlington: Department of Psychiatry, University of Vermont.

Albarracin, D., Fishbein, M., Johnson, B. T., & Muellerleile, P. (2001). Theories of reasoned action and planned behavior as models of condom use: A meta-analysis. *Psychological Bulletin, 127,* 142-161.

Alston, J. M., Chan-Kang, C., Marra, M. C., Pardey, P. G., & Wyatt, T. J. (2000). *A meta-analysis of rates of return to agricultural R&D: Ex pede herculem?* Washington DC: International Food Policy Research Institute.

Andrews, D. A., & Bonta, J. (2010). *The psychology of criminal conduct* (5th ed.). New Providence, NJ: Anderson/LexisNexis.

Andrews, D. A., Bonta, J., & Hoge, R. (1990). Classification for effective rehabilitation: Rediscovering psychology. *Criminal Justice and Behavior, 17*, 19-52.

Andrews, D. A., Zinger, I., Hoge, R. D., Bonta, J., Gendreau, P., & Cullen, F. T. (1990). Does correctional treatment work? A clinically relevant and psychologically informed meta-analysis. *Criminology, 28,* 369-404.

Aos, S., Lee, S., Drake, E., Pennucci, A., Klima, T., Miller, M., Anderson, L., Mayfield, J., & Burley, M. (2011). Return on investment: Evidence-based options to improve statewide outcomes (Document No. 11-07-1201). Olympia: Washington State Institute for Public Policy.

Aos, S., Lieb, R., Mayfield, J., Miller, M., Pennucci, A. (2004). Benefits and costs of prevention and early intervention programs for youth. Olympia: Washington State Institute for Public Policy.

Aos, S., Phipps, P., Barnoski, R., & Lieb, R. (2001). The comparative costs and benefits of programs to reduce crime: A review of research findings with implications for Washington State. In B. C. Welsh, D. P. Farrington, & L. W. Sherman (Eds.), *Costs and benefits of preventing crime* (pp. 149-175). Boulder, CO: Westview Press.

Armstrong, J. S. (1994). The fertile field of meta-analysis: Cumulative progress in agricultural forecasting. *International Journal of Forecasting, 10,* 140-147.

*Asscher, J. J., Dekovic, M., Manders, W. A., van der Laan, P. H., & Prins, P. J. M. (2013). A randomized controlled trial of the effectiveness of multisystemic therapy in the Netherlands: Post-treatment changes and moderator effects. *Journal of Experimental Criminology, 9,* 169-187.

*Asscher, J. J., Dekovic, M., Manders, W. A., van der Laan, P. H., Prins, P. J. M., & van Arum, S. (2014). Sustainability of the effects of multisystemic therapy for juvenile delinquents in the Netherlands: Effects on delinquency and recidivism. *Journal of Experimental Criminology, 10,* 227-243.

Babor, R. F., de la Fluente, J., Saunders, J., & Grant, M. (1992). AUDIT: The Alcohol Use Disorders Identification Test: Guidelines for use in primary health care. Geneva, Switzerland: World Health Organization.

*Baglivio, M. T., Jackowski, K., Greenwald, M. A., & Wolff, K. T. (2014). Comparison of multisystemic therapy and functional family therapy effectiveness: A multiyear statewide propensity score matching analysis of juvenile offenders. *Criminal Justice and Behavior, 41,* 1033-1056.

Bengtsson, J., Ahnstrom, J., & Weibull, A. C. (2005). The effects of organic agriculture on biodiversity and abundance: A meta-analysis. *Journal of Applied Ecology, 42,* 261-269.

Berleman, W. C., & Steinburn, T. W. (1969). The value and validity of delinquency prevention experiments. *Crime & Delinquency*, *15,* 471-478.

Berman, A. H., Bergman, H., Palmstierna, T., & Schlyter, F. (2005). Evaluation of the Drug Use Disorder Identification Test (DUDIT) in criminal justice and detoxification settings in a Swedish population sample. *European Addiction Research, 11,* 22–31.

Bernfeld, G. A., Jennings, V, & Corriveau, A. (2001). 'The road less traveled' in community corrections: Intensive ecological approaches and the struggle for treatment integrity. *Journal on Community Corrections, 16,* 51-65.

Blaske, D. M., Mann, B. J., & Henggeler, S. W. (1989). Individual, family and peer characteristics of adolescent sex offenders and assaultive offenders. (5 *Developmental Psychology, 25*, 846-855.

Blueprints Programs. (2014). About Us. Retrieved December 8, 2014, from http://www.blueprintsprograms.com/about.php.

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

Borduin, C. M., Blaske, D. M., Cone, L., Mann, B. J., & Hazelrigg, M. D. (1989). Development and validation of a measure of peer relations: The Missouri Peer Relations Inventory. Unpublished manuscript, Department of Psychology, University of Missouri—Columbia.

*Borduin, C. M., Henggeler, S. W., Blaske, D. M. & Stein, R. (1990). Multisystemic treatment of adolescent sexual offenders. *International Journal of Offender Therapy and*

Comparative Criminology, 35, 105-114.

*Borduin, C. M., Mann, B. J., Cone, L. T., Henggeler, S. W., Fucci, B. R., Blaske, D. M., & Williams, R. A. (1995). Multisystemic treatment of serious juvenile offenders: Long-term prevention of criminality and violence. *Journal of Consulting and Clinical Psychology, 53,* 569—578.

*Borduin, C. M., Schaeffer, C. M., & Heiblum, N. (2009). A randomized clinical trial of multisystemic therapy with juvenile sexual offenders: Effects on youth social ecology and criminal activity. *Journal of Consulting and Clinical Psychology, 77,* 26-37.

Borenstein M, Hedges, L. V., Higgins J. P. T., & Rothstein H. R. (2005). Comprehensive Meta-Analysis (Version 2) [Computer software]. Engelwood, NJ: Biostat. Available from https://www.meta-analysis.com/pages/brochure.php.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis.* United Kingdom: Wiley.

Bronfenbrenner, U. (1979). *The ecology of human development: Experiments in nature and design.* Cambridge, MA: Harvard University Press.

*Brown, T. L., Henggeler, S. W., Schoenwald, S. K., Brondino, M. J., & Pickrel, S. G. (1999). Multisystemic treatment of substance abusing and dependent juvenile delinquents: Effects on school attendance at posttreatment and 6-month follow-up. *Children's Services: Social Policy, Research, and Practice, 2,* 81-93.

*Brunk, M., Henggeler, S. W., & Whelan, J. P. (1987). A comparison of multisystemic therapy and parent training in the brief treatment of child abuse and neglect. *Journal of Consulting and Clinical Psychology, 55,* 311-318.

Burns, B. J., Schoenwald, S. K., Burchard, J. D., Faw, L., & Santos, A. B. (2000). Comprehensive community-based interventions for youth with severe emotional disorders: Multisystemic therapy and the wraparound process. *Journal of Child and Family Studies, 9,* 283-314.

Butler, S., Baruch, G., Hickley, N., & Fonagy, P. (2011). A randomized controlled trial of MST and a statutory therapeutic intervention for young offenders. Journal of the American Academy of Child & Adolescent Psychiatry, 50,* 1220-1235.

Card, N. (2012). *Applied meta-analysis for social science research*. New York, NY: Guilford.

Chalmers, T. C., & Lau, J. (1994). What is meta-analysis? *Emergency Care Research Institute*, *12*, 1-5.

Cressey, D. R. (1958). The nature and effectiveness of correctional techniques. *Law and Contemporary Problems, 23,* 754-771.

Cullen, F. T. (2005). The twelve people who saved rehabilitation: How the science of criminology made a difference—The American Society of Criminology 2004 presidential address. *Criminology, 43,* 1-42.

Cullen, F. T., & Gendreau, P. (2000). Assessing correctional rehabilitation: Policy, practice, and prospects. In J. Horney (Ed.), *Policies, processes, and decisions of the criminal justice system*: *Criminal justice 2000* (Vol. 3, pp. 109-175). Washington, DC: U.S. Department of Justice, National Institute of Justice.

Cullen, F. T., & Gendreau, P. (2001). From nothing works to what works: Changing professional ideology in the 21st century. *The Prison Journal*, *81,* 313-338.

Cullen, F. T., & Gilbert, K. E. (2014). *Reaffirming rehabilitation* (2nd ed.). Waltham, MA: Anderson.

Cullen, F. T., & Jonson, C. L. (2011). Rehabilitation and treatment programs. In J. Q. Wilson & J. Petersilia (Eds.), *Crime and public policy* (pp. 293-344). New York, NY: Oxford University Press.

Cullen, F. T., & Jonson, C. L., (2012). *Correctional theory: Context and Consequences*. Thousand Oaks, CA: Sage.

Cullen, F. T., & Smith, P. (2011). Treatment and rehabilitation. In M. Tonry (Ed.), *The Oxford handbook of crime and criminology*. (pp. 156-178). New York, NY: Oxford University Press.

Cumming, G., & Finch, S. (2005). Inference by eye: Confidence intervals and how to read pictures of data. *American Psychologist, 60,* 170-180.

*Cunningham, A. (2002) One step forward: Lessons learned from a randomized study of multisystemic therapy in Canada. London, Canada: Center for Children and Families in the Justice System. Retrieved December 8, 2014 from http://www.lfcc.on.ca/One_Step_ Forward.pdf.

Curtis, N. M., Ronan, K. R., & Borduin, C. M. (2004). Multisystemic treatment: A meta-analysis of outcome studies. *Journal of Family Psychology, 8,* 411-419.

Curtis, N. M., Ronan, K. R., Heiblum, N., & Crellin, K. (2009). Dissemination and effectiveness of multisystemic treatment in New Zealand: A benchmarking study. *Journal of Family Psychology, 23,* 119-129.

Derogatis, L. R. (1993). *Brief Symptom Inventory: Administration, scoring, and procedures manual*. Minneapolis, MN: National Computer Systems, Inc.

Derogatis, L. R., Lipman, R. S., & Covi, L. (1973). SCL-90: An outpatient psychiatric rating scale—Preliminary report. *Psychopharmacological Bulletin, 9*,13-27.

Dopp, A. R., Borduin, C. M., Wagner, D. V., & Sawyer, A. M. (2014). Economic impact of multisystemic therapy through midlife: A cost-benefit analysis with serious juvenile offenders and their siblings. *Journal of Consulting and Clinical Psychology.*

Dowden, C., & Andrews, D. A. (1999). What works for female offenders: A meta-analytic review. *Crime and Delinquency, 45,* 438-452.

Dowden, C., & Andrews, D. A. (2000). Effective correctional treatment and violent reoffending: A meta-analysis. *Canadian Journal of Criminology, 42,* 449-467.

Drake, E. K., Aos, S., & Miller, M. G. (2009). Evidence-based public policy options to reduce crime and criminal justice costs: Implications from Washington State. *Victims and Offenders, 4,* 170-196.

Durlak, J. A., & Lipsey, M. W. (1991). A practitioner's guide to meta-analysis. *American Journal of Community Psychology, 19,* 291-332.

Elliott, D. S. (1998). Editor's introduction In S. W. Henggeler, *Blueprints for violence prevention: Multisystemic therapy* (pp. xi-xxi). Boulder, CO: Institute of Behavioral Science, University of Colorado.

Ellis, D. A., Frey, M. A., Naar-King, S., Templin, T., Cunningham, P. B., & Cakan, N. (2005). Use of multisystemic therapy to improve regimen adherence among adolescents with type 1 diabetes in poor metabolic control: A randomized controlled trial. *Diabetes Care, 28,* 1604-1610.

Ellis, D. A., Naar-King, S., Frey, M. A., Templin, T., Rowland, M., & Greger, N. (2004). Use of multisystemic therapy to improve regimen adherence among adolescents with type 1 diabetes in poor metabolic control: A pilot study. *Journal of Clinical Psychology in Medical Settings, 11,* 315-324.

*Fain, T., Greathouse, S. M., Turner, S. F., & Weinberg, H. D. (2014). Effectiveness of multisystemic therapy for minority youth: Outcomes over 8 years in Los Angeles County. *OJJDP Journal of Juvenile Justice, 3,* 24-37.

Farrington, D. P., & Welsh, B. C. (2003). Family-based prevention of offending: A meta-analysis. *Australian and New Zealand Journal of Criminology, 36,* 127-151.

Farrington, D. P., & Welsh, B. C. (2007). *Saving children from a life in crime: Early risk factors and effective interventions*. New York: Oxford University Press.

Finckenauer, J. O. (1982) *Scared straight! and the panacea phenomenon*. Englewood Cliffs, NJ: Prentice-Hall.

Fixsen, D. L., Naoom, S. F., Blasé, K. A., Friedman, R. M., & Wallace, F. (2005). *Implementation Research: A Synthesis of the Literature.* Tampa: Univ. So. Florida. Louis de la Parte Florida Mental Health Inst. Natl. Implement. Res. Netw. FMHI Publ. 231.

Fonagy, P., Cottrel, D., Phillips, J., Bevington, D., Glaser, D., & Allison, E. (2015). *What works for whom? A critical review of treatments for children and adolescents* (2nd ed.). Guildford Press: New York.

Gendreau, P. (1996). The principles of effective intervention with offenders. In A. T. Harland (Ed.), *Choosing correctional options that work: Defining the demand and evaluating the supply* (pp. 117-130). Thousand Oaks, CA: Sage.

Gendreau, P., Goggin, C., Cullen, F. T., & Andrews, D. A. (2000). The effects of community sanctions and incarceration on recidivism. *Forum on Corrections Research, 12,* 10-13.

Gendreau, P., Little, T., & Goggin, C. (1996). A meta-analysis of the predictors of adult offender recidivism: What works! *Criminology, 34,* 575-607.

Gendreau, P. & Ross, R. R. (1979). Effective correctional treatment: Bibliotherapy for cynics. *Crime and Delinquency, 25,* 463-489.

Gendreau, P., & Ross, R. R. (1987). Revivification of rehabilitation: Evidence from the 1980s. *Justice Quarterly, 4,* 349-407.

Gendreau, P., & Smith, P. (2007). Influencing the "people who count": Some perspectives on the reporting of meta-analytic results for prediction and treatment outcomes with offenders. *Criminal Justice and Behavior, 34,* 1536-1559.

Gendreau, P., Smith, P., & French, S. (2006). The theory of effective correctional intervention: Empirical status and future directions. In F. T. Cullen, J. P. Wright, & K. R. Blevins (Eds.). *Taking stock: The status of criminological theory—advances in criminological theory—Volume 15.* New Brunswick, NJ: Transaction.

*Giles, M. J. (2003). *Gender differences among adolescents with conduct disorder in response to day treatment or multisystemic therapy* (Unpublished doctoral dissertation). Walden University, Minneapolis, MN.

Glass, G. V. (1976). Primary, secondary, and meta-analysis. *Educational Researcher, 5,* 3-8.

Glass, G. V., McGaw, B., & Smith, M. L. (1981). *Meta-analysis in social research.* Beverly Hills, CA: Sage.

Gold, M. 1974. A time for skepticism. *Crime and Delinquency, 20,* 20-24.

Greenwood, P. W. (2006). *Changing lives: Delinquency prevention as crime-control policy*. Chicago: University of Chicago Press.

Gresham, F. M., & Elliott, S. N. (1990). Social skills rating system. Circle Pines, MN: American Guidance Service.

Hanson, R., & Bussiere, M. (1998). Predicting relapse: A meta-analysis of sexual offender recidivism studies. *Journal of Consulting and Clinical Psychology, 66,* 348-362.

Harrison, F. (2010). Getting started with meta-analysis. *Methods in Ecology and Evolution, 2,* 1-10.

He, J., Vupputuri, S., Allen, K., Prerost, M. R., Hughes, J., & Whelton, P. K. (1999). Passive smoking and the risk of coronary heart disease: A meta-analysis of epidemiologic studies. *The New England Journal of Medicine, 340,* 920-926.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. New York: Academic Press.

Henggeler, S. W. (1997). *Treating serious anti-social behavior in youth: The MST approach.* Washington, D.C., U.S. Department of Justice, Office of Justice Programs Office of Juvenile Justice and Delinquency Prevention.

Henggeler, S. W. (1998). *Blueprints for violence prevention: Multisystemic therapy.* Boulder, CO: Institute of Behavioral Science, University of Colorado.

Henggeler, S. W. (1999). Multisystemic therapy: An overview of clinical procedures, outcomes, and policy implications. *Child Psychology and Psychiatric Review, 4,* 2-10.

Henggeler, S. W. (2011). Efficacy studies to large-scale transport: The development and validation of multisystemic therapy programs. *The Annual Review of Clinical Psychology,7,* 351-381.

Henggeler, S. W., & Borduin, C. M. (1992). *Multisystemic Therapy Adherence Scales.* Unpublished Instrument, Department of Psychiatry and Behavioral Sciences, Medical University of South Carolina.

Henggeler, S. W., Borduin, C. M., Schoenwald, S. K., Huey, S. J., & Chapman J. E. (2006). *Multisystemic Therapy Adherence Scale-Revised (TAM-R).* Unpublished Instrument, Department of Psychiatry and Behavioral Sciences, Medical University of South Carolina.

*Henggeler, S. W., Clingempeel, W. G., Brondino, M. J., & Pickrel, S. G., (2002). Four-year follow-up of multisystemic therapy with substance-abusing and substance-dependent juvenile offenders. *Journal of Consulting and Clinical Psychology, 74,* 42-54.

*Henggeler, S. W., Halliday-Boykins, C. A., Cunningham, P. B., Randall, J., Shapiro, S. B., & Chapman, J. E. (2006). Juvenile drug court: Enhancing outcomes by integrating evidence-based treatments. *Journal of Consulting and Clinical Psychology, 74,* 42-54.

*Henggeler, S. W., Letourneau, E. J., Chapman, J. E., Borduin, C. M., Schewe, P. A., & McCart, M. R. (2009). Mediators of change for multisystemic therapy with juvenile sexual offenders. *Journal of Consulting and Clinical Psychology, 77,* 451-462.

*Henggeler, S. W., Melton, G. B., Brondino, M. J., Scherer, D. G., & Hanley, J. H. (1997). Multisystemic therapy with violent and chronic juvenile offenders and their families: The role of treatment fidelity in successful dissemination. *Journal of Consulting and Clinical Psychology, 65,* 821-833.

*Henggeler, S. W., Melton, G. B., & Smith, L. A. (1992). Family preservation using multisystemic therapy: An effective alternative to incarcerating serious juvenile offenders. *Journal of Consulting and Clinical Psychology, 60,* 953-961.

Henggeler, S. W., Melton, G. B., Smith, L. A., Schoenwald, S. K., & Hanley, J. H. (1993). Family preservation using multisystemic treatment: Long-term follow-up to a clinical trial with serious juvenile offenders. *Journal of Child and Family Studies, 2,* 293-293.

Henggeler, S. W., Mihalic, S. F., Rone, L. Thomas, C., & Timmons-Mitchell, J. (1998). *Multisystemic therapy: Blueprints for violence prevention, Book six*. Boulder: Center for the Study and Prevention of Violence, University of Colorado at Boulder.

*Henggeler, S. W., Pickrel, S. G., & Brondino, M. J. (1999). Multisystemic treatment of substance abusing and dependent delinquents: Outcomes, treatment fidelity, and transportability. *Mental Health Services Research, 1,* 171-184.

Henggeler, S. W., Rodick, J. D., Borduin, C. M., Hanson, C. L., Watson, S. M. & Urey, J. R. (1986). Multisystemic treatment of juvenile offenders: Effects on adolescent behavior and family interactions. *Developmental Psychology, 22*, 132-141.

Henggeler, S. W., Rowland, M. D., Halliday-Boykins, C., Sheidow, A. J., Ward, D. M., Randall, J., Pickrel, S. G., & Edwards, J. (2003). One-year follow-up of multisystemic therapy as an alternative to the hospitalization of youths in psychiatric crisis. *Journal of the American Academy of Child and Adolescent Psychiatry, 42,* 543-551.

*Henggeler, S. W., Rowland, M. D., Randall, J., Ward, D. M., Pickrel, S. G., Cunningham, P. B., Miller, S. L., Edwards, J., Zealberg, J. J., Hand, L. D. & Santos, A. B. (1999). Home based multisystemic therapy as an alternative to the hospitalization of youths in psychiatric crisis: Clinical outcomes. *Journal of the American Academy of Child and Adolescent Psychiatry, 38,* 1331-1339.

Henggeler, S. W., & Schaeffer, C. M. (2010). Treating serious emotional and behavioural problems using multisystemic therapy. *Australian and New Zealand Journal of Family Therapy, 31,* 149-164.

Henggeler, S. W., & Schoenwald, S. K. (1998). *Multisystemic therapy supervisory manual: Promoting quality assurance at the clinical level.* Charleston, SC: MST Institute.

Henggeler, S. W., Schoenwald, S. K., Borduin, C. M., Rowland, M. D., & Cunningham, P. B. (1998). *Multisystemic treatment of antisocial behavior in children and adolescents*. New York: Guilford.

Henggeler, S. W., Schoenwald, S. K., Borduin, C. M., Rowland, M. D., & Cunningham, P. B. (2009). *Multisystemic therapy for antisocial behavior in children and adolescents* (2nd ed.). New York, NY: Guilford Press.

Henggeler, S. W., Schoenwald, S. K., Rowland, M. D., & Cunningham, P. B. (2002). *Serious emotional Disturbance in children and adolescents: Multisystemic therapy. New York: Guilford Press.*

Hodges, K., & Wong, M. M. (1996). Psychometric characteristics of a multidimensional measure to assess impairment: The Child and Adolescent Functional Assessment Scale. *Journal of Child and Family Studies*, 5, 445–467.

Huey, S. J., Jr., Henggeler, S. W., Brondino, M. J., Pickrel, S. G., (2000). Mechanisms of change in multisystemic therapy: Reducing delinquent behavior through therapist adherence and improved family and peer functioning. *Journal of Consulting and Clinical Psychology, 68,* 451-467.

*Huey, S. J., Jr., Henggeler, S. W., Rowland, M. D., Halliday-Boykins, C. A., Cunningham, P. B., Pickrel, S. G., & Edwards, J. (2004). Multisystemic therapy effects of attempted suicide by youth presenting psychiatric emergencies. *Journal of the American Academy of Child and Adolescent Psychiatry, 43,* 183-190.

Huey, S. J., Jr., Henggeler, S. W., Rowland, M. D., Halliday-Boykins, C. A., Cunningham, P. B., & Pickrel, S. G. (2005). Predictors of treatment response for suicidal youth referred for emergency psychiatric hospitalization. *Journal of Clinical Child and Adolescent Psychology, 34,* 582-589.

Hunt, M. (1997). *How science takes stock: The story of meta-analysis.* New York: Russell Sage.

Hunter, J. E., & Schmidt, F. L. (1990). *Methods of meta-analysis.* Newburk Park, CA: Sage.

Hunsley, J. (2007). Addressing key challenges in evidence-based practice in psychology. *Professional Psychology: Research and Practice, 39,* 113-121.

Hunsley, J., Elliott, K., & Therrien, Z. (2013). *The efficacy and effectiveness of psychological treatments.* Ottawa, ON: Canadian Psychological Association.

LeFrance, M., Hecht, M. A., & Paluck, E. L. (2003). The contingent smile: A meta-analysis of sex differences in smiling. *Psychological Bulletin, 129,* 305-334.

Leschied, A., & Cunningham, A. (2002). *Seeking effective interventions for serious young offenders: Interim results of a four-year randomized study of multisystemic therapy in Ontario, Canada.* London, CA: Centre for Children and Families in the Justice System.

*Letourneau, E. J., Henggeler, S. W., Borduin, C. M., Schewe, P. A., McCart, M. R., Chapman, J. E., & Saldana, L. (2009). Multisystemic therapy for juvenile sexual offenders: 1-year results from a randomized effectiveness trial. *Journal of Family Psychology, 23*, 89–102.

Levine, M., Perkins, D. D., & Perkins, D. V. (2005). *Principles of community psychology: Perspectives and applications* (3rd ed.,). New York, NY: Oxford University Press.

Light, R. J., & Pillemer, D. B. (1984). *Summing up: The science of reviewing research.* Cambridge, MA: Harvard University Press.

Lipsey, M. W. (1992). Juvenile delinquency treatment: A meta-analytic inquiry into the variability of effects. In T. D. Cook, G. Cooper, D. S. Cordray, H. Hartmann, L. V. Hedges, R. J. Light, T. A. Louis, & F. Mosteller (Eds)., *Meta-analysis for explanation: A casebook* (pp. 83-127). New York: Russell Sage.

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

Littell, J. H. (2005). Lessons from a systematic review of effects of multisystemic therapy. *Children and Youth Services, 27,* 445-463.

Littell, J. H., Popa, M., & Forsythe, B. (2005). *Multisystemic therapy for social, emotional, and behavioral problems in youth aged 10-17*. Campbell Systematic Reviews.

Logan, C. H., & Gaes, G. G. (1993). Meta-analysis and the rehabilitation of punishment. *Justice Quarterly, 10,* 245-263.

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

*Manders, W. A., Dekovic, M., Asscher, J. J., van der Laan, P. H., & Prins, P. J. M. (2013). Psychopathy as predictor and moderator of multisystemic therapy outcomes among adolescents treated for antisocial behavior. *Journal of Abnormal Child Psychology, 41,* 1121-1132.

*Mann, B. J., Borduin, C. M., Henggeler, S. W., & Blaske, D. M. (1990). An investigation of systemic conceptualizations of parent-child coalitions and symptom change. *Journal of Consulting and Clinical Psychology, 58,* 336-334.

Martinson, R. (1974). What works? Questions and answers about prison reform. *The Public Interest, 35*(Spring), 22-54.

*Mayfield, J. (2011). Multisystemic therapy outcomes in an evidence-based practice pilot.

Olympia: Washington State Institute for Public Policy.

McClure, E. B. (2000). A meta-analytic review of sex differences in facial expression and their development in infants, children and adolescents. *Psychological Bulletin, 126,* 424-453.

McCubbin, H. I., Patterson, I. M., & Wilson, L. R. (1985). FILE: Family Inventory of Life Events and Changes. In D. H. Olson, H. I. McCubbin, H. Barnes, A. Larsen, M. Muxen, & M. Wilson (Eds.), *Family inventories* (pp. 82-114). St. Paul, MN: University of Minnesota Press.

Merton, R. K. (1973). *The sociology of science: Theoretical and empirical investigations*. Norman K. Storer (Ed.). Chicago, IL: University of Chicago Press.

*Mitchell-Herzfeld, S., Shady, T. A., May, J., Kim, D. H., Marsh, K., Dorabawila, V., Rees, F. (2008). *Effects of multisystemic therapy (MST) on recidivism among juvenile delinquents in New York State.* Rensselaer, NY: New York State Office of Children and Family Services.

Morgan, R. D., Gendreau, P., Smith, P., Gray, A. L., Labrecque, R. M., MacLean, N., Van Horn, S., Bolanos, A. D., Batastini, A. B., & Mills, J. F. (2016). Quantitative synthesis of the effects of administrative segregation on inmates' well-being. Unpublished paper, Texas Tech University.

MST Services, Inc.. (2016). About Our History. Retrieved March 5, 2016, from http://mstservices.com/mst-service/our-history.

MST Services, Inc.. (2010). MST treatment model. Retrieved June 5, 2014, from http://www.mstservices.com/mst_treatment_model.php?print=true.

Mullen, B. (1989). *Advanced basic meta-analysis.* Hillsdale, NJ: Lawrence Erlbaum Associates, Inc., Publishers.

Naar-King, S., Ellis, D., Kolmodin, K., Cunningham, P., Jen, K. L. C., Saelens, B., & Brogan, K. (2009). A randomized pilot study of multisystemic therapy targeting obesity in African-American adolescents. *Journal of Adolescent Health, 45,* 417-419.

*Oakley, T. L. (2000). *Multisystemic therapy for high risk young offenders: An exploration of school outcomes* (Unpublished master of science thesis). University of Guelph, Guelph, Ontario.

*Ogden, T., & Hagen, K. A. (2006). Multisystemic therapy of serious behavior problems in youth: Sustainability of therapy effectiveness two years after intake. *Journal of Child and Adolescent Mental Health, 11,* 142-149.

*Ogden, T., Hagen, K. A., & Anderson, O. (2006). Sustainability of the effectiveness of a programme of multisystemic treatment (MST) across participant groups in the second year of operation. *Journal of Children's Services, 2,* 4-14.

*Ogden, T., & Halliday-Boykins, C. A. (2004). Multisystemic treatment of antisocial adolescents in Norway: Replication of clinical outcomes outside of the U.S. *Child and Adolescent Mental Health, 9*(2), 77-83.

Olson, D., Portner, J., & Lavee, Y. (1985). *FACES III.* Minneapolis: Department of Family Social Service University of Minnesota.

Orwin, R. G. (1983). A fail-safe N for the effect size in meta-analysis. *Journal of Educational Statistics, 8,* 157-159.

*Painter, K. (2009). Multisystemic therapy as community-based treatment for youth with severe emotional disturbance. *Research on Social Work Practice, 19,* 314-324.

Palmer, T. (1975). Martinson revisited. *Journal of Research in Crime and Delinquency, 12,* 133-152.

Palmer, T. (1992). *The re-emergence of correctional intervention.* Newbury Park, CA: Sage.

Pratt, T. C. (2001). *Assessing the relative effects of macro-level predictors of crime: A meta-analysis.* Ph.D. dissertation, Department of Criminal Justice, University of Cincinnati.

Quay, H. C., & Peterson, D. R. (1987). *Manual for the Revised Problem Behavior Checklist.* Coral Gables, FL: University of Miami.

Rhodes, D. R., Yu, J., Shanker, K., Deshpande, N., Varambally, R., Ghosh, D., Barrette, T., Pandey, A., & Chinnaiyan, A. M. (2004). Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression. *Proceedings of the National Academy of Sciences, 101,* 9309-9314.

Rothstein, H. R., Sutton, A. J., & Borenstein, M. (eds.). (2005). Publication bias in meta-analysis: Prevention, assessment, and adjustments. Chichester, UK: Wiley.

Rosenberg, M.S. (2005). The file-drawer problem revisited: A general weighted method for calculating fail-safe numbers in meta-analysis. *Evolution, 59*, 464-468.

Rosenthal, R. (1979). The 'file drawer' problem and tolerance for null results. *Psychological Bulletin, 86,* 638-641.

Rosenthal, R. (1984). *Meta-analytic procedures for social research.* Beverly Hills, CA: Sage.

Rosenthal, R. (1991). Meta-Analytic Procedures for Social Research (revised ed.). Newbury

Park: Sage Publications.

Rounsaville, B. J., Carroll, K. M., & Onken, L. S. (2001). A stage model of behavioral therapies research: Getting started and moving on from Stage 1. *Clinical Psychology: Science and Practice, 8,* 133–142.

*Rowland, M. D., Halliday-Boykins, C. A., Henggeler, S. W., Cunningham, P. B., Lee T. G., Kruesi, M. J. P., & Shapiro, S. B. (2005). A randomized trial of multisystemic therapy with Hawaii's Felix Class youths. *Journal of Emotional and Behavioral Disorders, 13,* 13–23.

Scammacca, N., Roberts, G., & Stuebing, K. K. (2014). *Review of Educational Research, 84*(4), 328–364.

Scammacca, N., Roberts, G., Vaughn, S., & Stuebing, K. (2015). A meta-analysis of interventions for struggling readers in grades 4–12: 1980–2011. *Journal of Learning Disabilities, 48,* 369–490.

*Schaeffer, C. M., Swenson, C. C., Tuerk, E. H., & Henggeler, S. W. (2013). Comprehensive treatment for co-occurring child maltreatment and parental substance abuse: Outcomes from a 24-month pilot study of the MST-Building Stronger Families program. *Child Abuse and Neglect, 37*, 596-607.

Schoenwald, S. K. (1998). *Multisystemic therapy consultation manual.* Charleston, SC: MST Institute.

Schoenwald, S. K. (2010). From policy pinball to purposeful partnerships: The policy contexts of multisystemic therapy transport and dissemination. In J. R. Weisz, A. E. & Kazdin (Eds.), *Evidence-Based Psychotherapies for Children and Adolescents* (pp. 560-576).

Schoenwald, S. K. (2001). *The MST Consultant Adherence Measure.* Charleston: Family Services Research Center, Medical University of South Carolina.

Schoenwald, S. K., Brown, T. L., & Henggeler, S. W. (2000). Inside multisystemic therapy: Therapist, supervisory, and program practices. *Journal of Emotional and Behavioral Disorders, 8,* 113-127.

Schoenwald, S. K., Chapman, J. E., & Sheidow, A. J. (2006). *Implementation fidelity in MST. In S.K. Schoenwald & J. Reid, Co-Chairs, Community-based model programs panel: Implementing with fidelity.* Blueprints Conference, 2006: Evidence-Based Programs, Research-to-Practice Conference, Denver, CO.

Schoenwald, S. K., Henggeler, S. W., Brondino, M. J., & Rowland, M. D. (2000). Multisystemic therapy: Monitoring treatment fidelity. *Family Process, 39,* 83-103.

Schoenwald, S. K., Henggeler, S. W., & Edwards, D. (1998). *MST Supervisor Adherence Measure.* Charleston, SC: MST Institute.

Schoenwald, S. K., Ward, D. M., Henggeler, S. W., Pickrel. S. G., & Patel, H. (1996). MST treatment of substance abusing or dependent adolescent offenders: Costs of reducing incarceration, inpatient and residential placement. *Journal of Child and Family Studies, 5,* 431-444.

*Schoenwald, S. K., Ward, D. M., Henggeler, S. W., & Rowland, M. D. (2000). MST vs. hospitalization for crisis stabilization of youth: Placement outcomes 4 months post-referral. *Mental Health Services Research, 2,* 3-12.

Schmidt, F. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods 1,* 115-129.

Sherman, L. W. 2011. Criminology as invention. In M. Bosworth & C. Hoyle (Eds.), *What is criminology?* (pp. 423-438). Oxford, UK: Oxford University Press.

*Sheidow, A. J., Bradford, W. D., Henggeler, S. W., Rowland, M. D., Halliday-Boykins, C., Schoenwald, S. K., & Ward, D. M. (2004). Treatment costs for youths in psychiatric crisis: Multisystemic therapy versus hospitalization. *Psychiatric Services, 55,* 548-554.

Sheidow, A. J., Henggeler, S. W., & Schoenwald, S. K. (2003). Multisystemic therapy. In T. L. Sexton, G. R. Weeks, & M. S. Robbins (Eds.), *Handbook of family therapy* (pp. 303-322). New York: Brunner-Routledge.

Sheidow, A. J., Schoenwald, S. K., Wagner, H. R., Allred, C.A., & Burns, B. J. (2006). Predictors of workforce turnover in a transported treatment program. *Administration and Policy in Mental Health and Mental Health Services Research, 1,* 1-12.

Simpson D. D., & McBride, A. A. (1992). Family, friends, and self (FFS) assessment scales for Mexican American youth. *Hispanic Journal of Behavioral Sciences, 14,* 327-340

Skinner, H. A., Steinhauer, P. D., & Santa-Barbara, J. (1983). The Family Assessment Measure. *Canadian Journal of Community Mental Health, 2,* 91-105.

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

Smith, M. L., & Glass, G. V. (1977). Meta-analysis of psychotherapy outcome studies. *American Psychologist, 32,* 752–760.

*Smith-Boydston, J. M., Holtzman, R. J., & Roberts, M. C. (2014). Transportability of multisystemic therapy to community settings: Can a program sustain outcomes without MST Services oversight? *Child Youth Care Forum, 43,* 593-605.

*Smith-Toles, M. D. (2003). *Mental health and related behavioral problems in children and adolescents: Modified multisystemic versus traditional therapy in residential treatment* (Unpublished doctoral dissertation). Walden University, Minneapolis, MN.

Snook, B., Eastwood, J., Gendreau, P., Goggin, C., & Cullen, F. T. (2007). Taking stock of criminal profiling: A narrative review and meta-analysis. *Criminal Justice and Behavior*, *34*, 437-453.

Sohn, D. (1995). Meta-analysis as a means of discovery. *American Psychologist, 50*, 108-110.

Spence, S. H. (1995). *Social skills training: Enhancing social competence with children and adolescents*. Windsor, Berkshire, United Kingdom: NFER-Nelson.

Stambaugh, L. F., Mustillo, S. A., Burns, B. J., Stephens, R. L., Baxter, B., Edwards, D., & DeKraai, M. (2007). Outcomes from wraparound and multisystemic therapy in a center for mental health services system-of-care demonstration site. *Journal of Emotional and Behavioral Disorders, 15,* 143-155.

Stanne, M. B., Johnson, D. W., & Johnson, R. T. (1999). Does competition enhance or inhibit motor performance? *Psychological Bulletin, 125,* 133-154.

Strazzullo, P., D'Elia, L., Kandala, N-B., & Cappuccio, F. P. (2009). Salt intake, stroke, and cardiovascular disease: Meta-analysis of prospective studies. *British Medical Journal, 339,* 45-67.

Strother, K. B., Swenson, M. E., & Schoenwald, S. K. (1998). *Multisystemic therapy organizational manual.* Charleston, SC: MST Institute.

*Sundell, K., Hansson, K., Lofholm, C. A., Olsson, T., Gustle, L. H., & Kadesjo, C. (2008). The transportability of multisystemic therapy to Sweden: Short-term results from a randomized trial of conduct disordered youth. *Journal of Family Psychology, 22,* 550-560.

Sutton, A. J. (2009). Publication bias. In H. Cooper, L. V. Hedges, & J. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed.) (pp. 435-452). New York, NY: Russell Sage Foundation.

Swenson, M. & Duncan, M. (1998). *Multisystemic therapy: How is it done?* [Fact Sheet]. Retrieved April 12, 2015 from http://mstservices.com/files/howitsdone.pdf.

Swenson, C. C., Penman, J., Henggeler, S. W., & Rowland, M. D. (2010*). Multisystemic therapy for child abuse and neglect.* Charleston, SC: Family Services. Research Center, Medical University of South Carolina.

*Swenson, C. C., Schaeffer, C., Henggeler, S. W., Faldowski, R., & Mayhew, A. M. (2010). Multisystemic therapy for child abuse and neglect: A randomized effectiveness trial. *Journal of Family Psychology, 24,* 497-507.

Texas Department of Mental Health Mental Retardation. (2004). User's manual for the child and adolescent Texas recommended authorization guidelines. Retrieved April 12, 2015 from http://www.dshs.state.tx.us/mhprograms/RDMCAtrag.shtm.

Thiam, A., Bravo-Ureta, B. E., & Rivas, T. E. (2005). Technical efficiency in developing country agriculture: A meta-analysis. *Agricultural Economics, 25,* 235-243.

*Timmons-Mitchell, J., Bender, M. B., Kisna, M. A., & Mitchell, C. C. (2006). An independent effectiveness trial of multisystemic therapy with juvenile justice youth. *Journal of Clinical Child and Adolescent Psychology, 35,* 227-236.

van der Stouwe, T., Asscher, J. J., Stams, G. J. J. M., Dekovic, M., & vad der Laan, P. H. (2014). The effectiveness of multisystemic therapy (MST): A meta-analysis. *Clinical Psychology Review, 34,* 468-481.

Van Voorhis, P., & Braswell, M. (2009). Family therapy. In Van Voorhis, P., Braswell, M., & Lester, D. (Eds.). *Correctional Counseling and Rehabilitation* (7th ed.). Cincinnati, OH: Lexis/Nexis.

*Wagner, D. V., Borduin, C. M., Sawyer, A. M., & Dopp, A. R. (2014). Long-term prevention of criminality in siblings of serious and violent juvenile offenders: A 25-year follow-up to a randomized clinical trial of multisystemic therapy. *Journal of Conulting and Clinical Psychology, 82,* 492-499.

*Weiss, B., Han, S., Harris, V., Catron, T., Ngo, V. K., Caron, A., Gallop, R., & Guth, C. (2013). An independent randomized clinical trial of multisystemic therapy with non-court-referred adolescents with serious conduct problems. *Journal of Consulting and Clinical Psychology, 81,* 1027-1039.

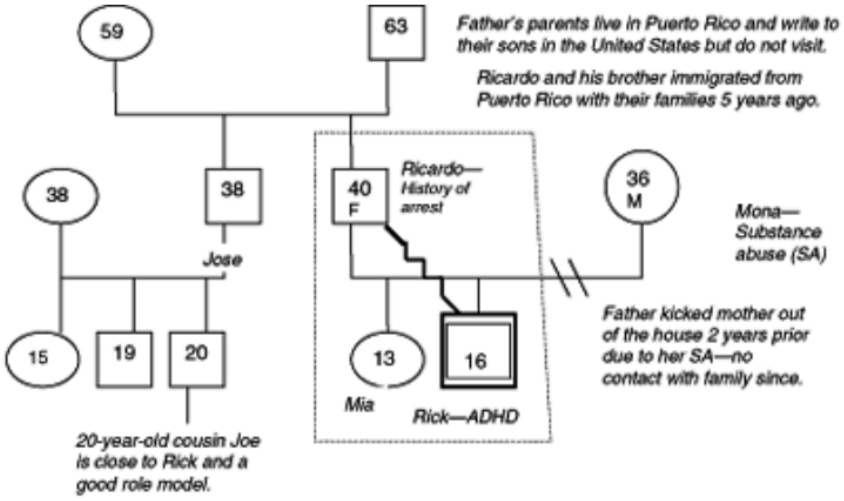
Whitehead, J.T. & Lab, S. P. (1989). A meta-analysis of juvenile correctional treatment. *Journal of Research in Crime and Delinquency, 26*, 276-95.

Wilson, B. J., & Giguere, V. (2008). Meta-analysis of human cancer microarrays reveals GATA3 is integral to the estrogen receptor alpha pathway. *Molecular Cancer, 7,* 49-57.

Winters, K. C., & Henly, G. (1989). *The Personal Experiences Inventory*. Los Angeles, CA: Western Psychological Services.

Wolf, F. M. (1986). *Meta-analysis: Quantitative methods for research synthesis.* Beverly Hills, CA: Sage.

APPENDIX A–EXAMPLE OF BACKGROUND INFORMATION AND STRENGTHS AND NEEDS ASSESSMENT



Referral Behaviors

Behavior	Frequency	Intensity	Duration
Marijuana use	3–4 times a week	1–2 blunts shared with 2 peers	Approx. 16 months
Truancy	2–3 times/ week	Skips entire day	Started last school year (12 months ago)
Physical aggression, fights with peers	Two known incidents	The client had to receive stitches for laceration, black eye	Incidents were 3 and 8 months prior to intake
Father–son conflict—both are verbally aggressive	1–2 times per week	Both scream, youth uses rude language, lasts 1–2 minutes	Approximately 2 years

Desired Outcomes

Participant	Goal
Rick	Get off probation, get Dad to stop nagging me
Ricardo Naaves (father)	Rick to go to school, get a job, make me proud of him
Mia (sister)	Do good, stop getting into trouble, stop fighting with Dad
Jose Naaves (uncle)	Stay away from bad kids, stop smoking dope
Mr. Johnson (probation officer)	Stay in school, stop using drugs, no violence

Source: Henggeler, Schoenwald, Borduin, Rowland, & Cunningham, 2009, p. 72

Background Case Information		
Family: Smith	Therapist: T. Knight	Date: 2/28/2014

Systemic Strengths	Systemic Weaknesses/Needs
Individual	
Athletic—enjoys football	Impulsive—acts before he thinks
Social—leader amongst group of friends	Temper—hot-headed, short fuse
Relationship skills—wants to improve relationship with both mom and dad	Anger—gets angry easily, displays many angry attitudes
Admits substance abuse problems—knows he has a problem using marijuana	School—failing most classes
Takes medications as prescribed—history of ADD (takes appropriate medication)	Substance abuse—does not believe he can stop using marijuana
Ready to work—interested in finding a part-time job	Conflict with family and friends—does not readily admit that he plays a significant role in situations when there is conflict
	Peers—cannot stop hanging out with antisocial friends and negative influencers
Family	
Father and Mother—committed to helping the youth	Father—authoritarian parenting style
Siblings—prosocial and could be role-models to the youth	Father and son—demonstrate high level of contact typically
	Mother—history (6+ years ago) of substance abuse, but has been sober since then

School	
Sports—plays on football team	Sports—could get kicked off the football team if grades do not improve
Teachers—appear to care about youth and have his best interests in mind – want to help him do better in school and do not want him dropping out/failing out	Parents—tend to avoid interacting with school personnel, even though they try to reach out to parents
Grades—wants to improve grades so he does not get kicked out / have to drop out	
Peers	
Parents—recall positive friends in youth's life 2 years ago – they still go to youths school	Friends—has not had positive influencers in his life for about 2 years; most of his current friends skip school / are failing out
Siblings—1 sister and 1 brother that youth is close with go, to the same school	Drugs—most friends youth hangs out with use drugs / drink
Football team—most kids on team are pro-social and could show youth how to do better	
Community	
Neighborhood—offers sports and lots of other recreational activities for youths to get involved in	Neighborhood—on the outskirts of a bad area with high crime / lots of drugs
Neighbor—watches youth and his siblings sometimes when parents are working – positive role-model for the kids and their family	Police—known youth and his friends and many of the things they do together that get them into trouble
Church—family goes to church, youth expressed interest in getting involved in youth activities here	

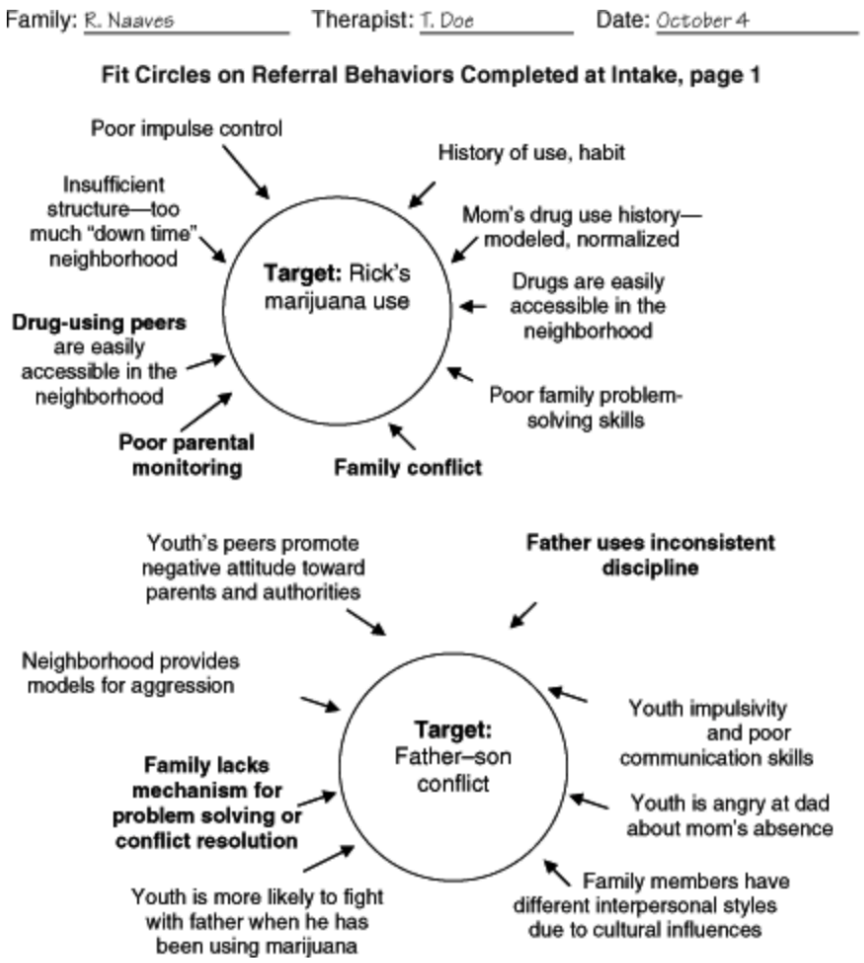
Source: Adapted from Henggeler, Schoenwald, Borduin, Rowland, & Cunningham, 2009, p. 73

APPENDIX B–CASE SUMMARY AND CONSULTATION FORM

Case Summary for Supervision and Consultation – Description and Examples	
Case Summary Category	Example
Overarching / Primary MST Goals	Joe will stop using marijuana as evidenced by clean urine screens for 10 consecutive weeks, per report of parents, therapist, and probation officer
Previous Intermediary Goals	Therapist and father to attend Joe's IEP conference at school to determine next steps for Joe's education
Barriers to Intermediary Goals	Joe left school early on Tuesday and Thursday last week (new information discovered during school meeting)
Advances in Treatment	Teachers and father agree that Joe may need to switch to a more vocational track in school; guidance counselor is investigating this option
Assessment of "Fit" Between Identified Problems and Their Broader Systemic Context	Regarding why Joe left school early--Joe reports attitude of "why should I try—I'm still going to fail"; Joe's friends were leaving school too
New Intermediary Goals for Next Week	Therapist to check in with father concerning school daily report card

Source: Adapted from Henggeler, Schoenwald, Borduin, Rowland, & Cunningham, 2009, pp. 35-37

APPENDIX C–EXAMPLE OF A FIT CIRCLE



Source: Henggeler, Schoenwald, Borduin, Rowland, & Cunningham, 2009, p. 42

APPENDIX D–THERAPIST ADHERENCE MEASURE—REVISED (TAM-R)

Multisystemic Therapy Institute
Therapist Adherence Measure-Revised (TAM-R)

Current Supervisor: _____
Therapist Code: _____
Family Code: _____
Respondent _____ Caregiver _____
Date Form Completed (mm-dd-yyyy): _____

Please answer the following two questions:

How many times has the therapist met with your family within the last week? _____
If zero, then when did the therapist last see anyone in the family (mm/dd/yyyy)? _____

Regarding your last 2-3 sessions:

	Did Not Respond	Not at All	A Little	Some	Pretty Much	Very Much
1. The therapist tried to understand how my family's problems all fit together.						
2. My family and the therapist worked together effectively.						
3. My family knew exactly which problems we were working on.						
4. The therapist recommended that family members do specific things to solve our problems.						
5. The therapist's recommendations required family members to work on our problems almost every day.						
6. The therapist understood what is good about our family.						
7. My family and the therapist had similar ideas about ways to solve problems.						
8. The therapist tried to change some ways that family members interact with each other.						
9. The therapist tried to change some ways that family members interact with people outside the family.						
10. My family and the therapist were honest and straightforward						

	Did Not Respond	Not at All	A Little	Some	Pretty Much	Very Much
with each other.						
11. The therapist's recommendations should help the children to mature.						
12. Family members and the therapist agreed upon the goals of the session.						
13. My family talked with the therapist about how well we followed her/his recommendations from the previous session.						
14. My family talked with the therapist about the success (or lack of success) of her/his recommendations from the previous session.						
15. We got much accomplished during the therapy session.						
16. My family was sure about the direction of treatment.						
17. The therapist's recommendations made good use of our family's strengths.						
18. My family accepted that part of the therapist's job is to help us change certain things about our family.						
19. The therapist's recommendations should help family members to become more responsible.						
20. The therapist talked to family members in a way we could understand.						
21. Our family agreed with the therapist about the goals of treatment.						
22. The therapist checked to see whether homework was completed from the last session.						
23. The therapist did whatever it took to help our family with tough situations.						
24. The therapist helped us to						

	Did Not Respond	Not at All	A Little	Some	Pretty Much	Very Much
enforce rules for the child.						
25. The therapist helped family members talk with each other to solve problems.						

	Did Not Respond	Not an Issue	Not at All	A Little	Some	Pretty Much	Very Much
26. The therapist helped us keep our child from hanging around with troublesome friends.							
27. The therapist helped us improve our child's behavior at school.							
28. The therapist helped us get our child to stay in school every day.							

Source: MSTInstitute.org

APPENDIX E–SUPERVISOR ADHERENCE MEASURE (SAM)

Form ID Organization Team Supervisor Name Therapist Current Consultant/System Supervisor Date Form Completed (mm/dd/yyyy)					
Please consider your supervision sessions over the past two months as you complete the following items.					
1. When the supervisor recommended changes in my course of action, the rationale for the recommendation was described in terms of one or more of the MST principles.	Never	Rarely	Sometimes	Often	Almost Always
2. You could tell that the supervisor was in charge of the sessions.	Never	Rarely	Sometimes	Often	Almost Always
3. Team members took a long time to describe the details of cases before the supervisor spoke.	Never	Rarely	Sometimes	Often	Almost Always
4. The supervisor asked clinicians for evidence to support their hypotheses about the causes of problems targeted for change or of barriers to intervention success.	Never	Rarely	Sometimes	Often	Almost Always
5. The supervisor asked clinicians how descriptions of this week's case developments pertained to identification of barriers to success.	Never	Rarely	Sometimes	Often	Almost Always
6. When clinicians talked about events in the distant past, the supervisor recommended that current interactions within the family and between family members and others be examined first.	Never	Rarely	Sometimes	Often	Almost Always
7. When clinicians reported on a variety of interventions tried during the week, the supervisor asked for clarification regarding which intermediary goals the interventions aimed to address.	Never	Rarely	Sometimes	Often	Almost Always
8. The supervisor followed up on recommendations made in previous supervision sessions.	Never	Rarely	Sometimes	Often	Almost Always
9. When interventions were not successful, discussion focused on identifying the barriers to success and actions the clinician should take to overcome them.	Never	Rarely	Sometimes	Often	Almost Always
10. I have the skills to implement all of the recommendations made in supervision.	Never	Rarely	Sometimes	Often	Almost Always
11. Interventions that were discussed targeted sequences of interaction between family members.	Never	Rarely	Sometimes	Often	Almost Always

12. Clinicians received positive feedback during the sessions.	Never	Rarely	Sometimes	Often	Almost Always
13. The supervisor asked clinicians how descriptions and questions about case developments pertained to "fit" assessment.	Never	Rarely	Sometimes	Often	Almost Always
14. It was easy for team members to acknowledge frustrations, mistakes, and failures.	Never	Rarely	Sometimes	Often	Almost Always
15. When a clinician presented information about events that transpired during the week, the supervisor asked the clinician and team to clarify the relevance of the information to one or more steps of the analytical process.	Never	Rarely	Sometimes	Often	Almost Always
16. Weekly case summaries were referred to during the discussion of cases.	Never	Rarely	Sometimes	Often	Almost Always

Source: MSTInstitute.org

APPENDIX F–CONSULTANT ADHERENCE MEASURE (CAM)

Organization

Team

Consultant/System Supervisor

Therapist or Supervisor Name

Respondent

Last MST Consultation Date

For questions 1 -19, please think about your last MST consultation session

	Never	Rarely	Sometimes	Usually	Almost Always	Always
1. The consultant explained how to implement specific intervention strategies for a case						
2. You could tell the consultant had case-specific ideas about barriers to success and how to overcome them						
3. The consultant helped when the team was "stuck" on some aspect of a case						
4. The consultant described interventions in sufficient detail that clinicians could carry them out						
5. The consultant addressed clinician behaviors that facilitate engagement or treatment progress in specific cases						
6. The consultant was competent at his/her job						
7. The consultant really listened when clinicians talked						
8. The consultant gave positive feedback to clinicians						
9. The consultant conveyed a sense that she/he and the team are "in it together."						
10. You could tell the consultant had the best interests of the client and clinicians at heart						
11. The consultant gave supportive feedback to clinicians when needed						
12. The consultant referred to specific MST principles when discussing cases						
13. The consultant helped generate a more comprehensive understanding of the "fit" of a problem						
14. The consultant explained what he/she was doing and why						
15. The consultant helped clinicians prioritize problems and intervention targets						
16. The consultant tried to gauge clinician "buy in" to his/her recommendations						
17. Consultation was well structured						
18. The consultant conveyed a "can do" attitude						
19. You could tell the consultant was well prepared						

Please answer the following questions about your consultant overall	Not at all	A Little	Somewhat	Quite	Very	Extremely
20. How knowledgeable do you think your consultant is in the theory of MST?						
21. How skilled do you think your consultant is in treatment modalities used in MST such as strategic, structural, behavioral, cognitive-behavioral, therapies?						
22. How skilled do you think your consultant is in implementing MST interventions?						
23. How skilled do you think your consultant is in teaching clinicians to do MST?						

Source: MSTInstitute.org

APPENDIX G–CODEBOOK

The Effectiveness of Multisystemic Therapy: A Review of the Evidence

Meta-Analysis Codebook

Identification #: _____

Title: _____
_____

_____

Authors: _____

Outcomes: _____
_____
_____
_____

I. PUBLICATION CHARACTERISTICS

A. Publication Type **[Pubtype]** _____

1 = Book
2 = Book chapter
3 = Federal report
4 = State or local report
5 = Conference paper
6 = Journal
7 = Thesis/Dissertation
8 = Other (specify) _____

B. Publication Decade **[Decade]** _____

1 = 1980
2 = 1990
3 = 2000
4 = 2010

C. Primary Author Affiliation **[Affiliation]** _____

1 = University
2 = State/Local agency
3 = Federal agency
4 = Other (specify) _____
5 = Missing

D. Primary Author Discipline **[Discipline]** _____

1 = Criminal justice/Criminology
2 = Psychology
3 = Sociology
4 = Social work
5 = Other (specify) _____
6 = Missing

II. STUDY CONTEXT CHARACTERISTICS

A. Type of Funding Agency **[Funding]** _____
- 1 = Unfunded
- 2 = Agency funded
- 3 = State funded
- 4 = Federally funded
- 5 = Other (specify) _____
- 6 = Missing

B. Geographic Location **[GeoLocation]** _____
- 1 = USA
- 2 = Canada
- 3 = UK
- 4 = Australia
- 5 = New Zealand
- 6 = Other (specify) _____
- 7 = Missing

C. Decade Data Collection Began **[Gathered]** _____
- 1 = 1980
- 2 = 1990
- 3 = 2000
- 4 = 2010
- 5 = Missing

D. Involvement of the Evaluator **[Involvement]** _____
- 0 = No
- 1 = Yes

E. Scott Henggeler Involved **[Henggeler]** _____
- 0 = No
- 1 = Yes

III. SAMPLE CHARACTERISTICS

A. Target Population **[TargetPop]** _____

1 = Serious juvenile offenders
2 = Youths with serious conduct problems
3 = Substance-abusing or –dependent juveniles
4 = Juvenile sexual offenders
5 = Youths presenting serious emotional disturbances or mental health issues
6 = Maltreating families
7 = Dual diagnosis
8 = At-risk for delinquency

B. Mean Age **[MeanAge]** _____

C. Gender (N & %) **[Males]** _____

[Females] _____

D. Race

N (%) Black **[RaceBlack]** _____

N (%) White **[RaceWhite]** _____

N (%) Hispanic **[RaceHispanic]** _____

N (%) Other (specify) _____ **[RaceOther]** _____

E. Race Dichotomous

N (%) White **[RaceWhite]** _____

N (%) Nonwhite **[RaceNonWhite]** _____

F. N (%) with at least one prior arrest **[PriorArrest]** _____

G. N (%) with at least one prior incarceration **[PriorIncarcer]** _____

H. Measurement of Risk **[RiskMeasure]** _____

1 = Uses valid psychometric instrument
2 = Author reported risk level/number only
3 = Coder defined (based on youth's criminal history)
4 = Cannot tell (not reported)

I. Risk Level **[RiskLevel]** _____

1 = Low
2 = Moderate (midpoint on scale)
3 = High
4 = Other (specify) _____
5 = Cannot tell (not reported)

IV. TREATMENT CHARACTERISTICS

A. Source of Clients for Treatment **[Source]** _____

1 = Non-criminal justice agency
2 = Criminal justice agency
3 = Multiple sources

B. Completion Rate **[TotalRate]** _____

N (%) Successfully Completed MST **[MSTComplete]** _____

N (%) Successfully Completed Comparison Treatment **[CompComplete]** _____

C. Was attrition a problem (i.e., did more than 20% drop out of MST or Comparison group)? **[Attrition]** _____

0 = No
1 = Yes
2 = Unknown

D. Type of Comparison Group Condition **[CompCondition]** _____

1 = No treatment/services
2 = Individual counseling/therapy
3 = Non-criminal justice services only (e.g., MH services)
4 = Traditional probation
5 = Traditional probation + individual or group treatment
6 = Group or family counseling
7 = Historical data
8 = Mixed
9 = Other (specify) _____
10 = Missing

E. Was treatment fidelity measured for MST? **[FidelityMST]** _____

0 = No / Cannot tell
1 = Yes

F. If treatment fidelity was measured, what was average adherence score for MST? **[AdherenceScore]** _____

G. Sample Size

Total sample size	**[TotalSampSize]**	_____
MST sample size	**[MSTSampSize]**	_____
Comparison group sample size	**[CompSampSize]**	_____

H. Number of Study Sites **[StudySites]** _____

1= Single site
2 = Multisite

I. Average number of treatment hours

MST	**[HoursMST]**	_____
Comparison	**[HoursComp]**	_____

J. Average number of days

MST	**[DaysMST]**	_____
Comparison	**[DaysComp]**	_____

K. Study Setting **[Setting]** _____

1 = Efficacy trial
2 = Effectiveness study

L. Study Design **[Design]** _____

1 = Randomized-experimental
2 = Quasi-experimental

V. STUDY CHARACTERISTICS

A. Follow-Up Period (months) – Delinquency Measure **[FUDelinq]** _____

B. Follow-Up Period (months) – Other Measures **[FUOthers]** _____

_____

C. Source of Data (list for each outcome if applicable)

1 = Self-report
2 = Official record
3 = Observation
4 = Mixed
5 = Other (specify) _____
6 = Cannot tell

[DataSource1] _____
[DataSource2] _____
[DataSource3] _____
[DataSource4] _____
[DataSource5] _____
[DataSource6]

D. Direction of Treatment/Comparison (list for each outcome if applicable)

1 = Favors treatment group (i.e., MST did better)
2 = Favors comparison group (i.e., comparison did better)
3 = Favors neither (i.e., equal)

[Direction1] _____
[Direction2] _____
[Direction3] _____
[Direction4] _____
[Direction5] _____
[Direction6]

E. Type of Analysis (list for each outcome if applicable)

1 = Frequency/percent
2 = Mean/standard deviation
3 = Other (specify) _____
4 = Mixed

[Analysis1] _____
[Analysis2] _____
[Analysis3] _____
[Analysis4] _____
[Analysis5] _____
[Analysis6] _____

APPENDIX H—LISTING OF AUTHOR(S), PUBLICATION YEAR, EFFECT SIZE, & SAMPLE SIZE—DELINQUENCY

Appendix H
Listing of Author(s), Publication Year, Effect Size, & Sample Size—Delinquency

Author(s)	Pub. Year	*r*	N
Asscher, Dekovic, Manders, van der Laan, Prins, & van Arum	2014	0.00	256
Baglivio, Jackowski, Greenwald, & Wolff	2014	-0.04	2,203
Borduin, Henggeler, Blaske, & Stein	1990	0.49	16
Borduin, Mann, Cone, Henggeler, Fucci, Blaske, & Williams	1995	0.41	176
Borduin, Schaeffer, & Heiblum	2009	0.37	48
Butler, Baruch, Hickey, & Fonagy	2011	0.38	108
Cunningham	2006	0.00	409
Fain, Greathouse, Turner, & Weinberg	2014	0.09	1137
Henggeler, Halliday-Boykins, Cunningham, Randall, Shapiro, & Chapman	2006	-0.05	161
Henggeler, Letourneau, Borduin, Schewe, & McCart	2009	0.21	127
Henggeler, Melton, Brondino, Scherer, & Hanley	1997	0.16	155
Henggeler, Melton, & Smith	1992	0.25	84
Henggeler, Pickrel, & Brondino	1999	0.04	118

Appendix H
Listing of Author(s), Publication Year, Effect Size, & Sample Size—Delinquency

Author(s)	Pub. Year	*r*	N
Henggeler, Rowland, Randall, Ward, Pickrel, Cunningham, Miller, Edwards, Zealberg, Hand, & Santos	1999	-0.07	116
Mayfield	2011	0.09	202
Mitchell-Hertzfeld, Shady, Mayo, Kim, Marsh, Dorabawila, & Rees (boys)	2008	0.05	269
Mitchell-Hertzfeld, Shady, Mayo, Kim, Marsh, Dorabawila, & Rees (girls)	2008	-0.11	629
Ogden & Hagen	2006	0.13	75
Ogden, Hagen, & Andersen	2007	0.32	105
Painter	2009	0.15	174
Rowland, Halliday-Boykins, Henggeler, Cunningham, Lee, Kruesi, & Shapiro	2005	-0.06	31
Smith-Boydston, Holtzman, & Roberts	2014	0.08	147
Sundell, Hanson, Lofholm, Olsson, Gustle, Kadesjo	2008	-0.07	156
Timmons-Mitchell, Bender, Kishna, & Mitchell	2006	0.29	93
Wagner, Borduin, Sawyer, & Dopp	2014	0.21	110
Weiss, Han, Harris, Catron, Ngo, Caron, Gallop, & Guth	2013	-0.07	164

APPENDIX I—LISTING OF AUTHOR(S), PUBLICATION YEAR, EFFECT SIZE, & SAMPLE SIZE—PROBLEM BEHAVIOR

Appendix I
Listing of Author(s), Publication Year, Effect Size & Sample Size—Problem Behavior

Author(s)	Pub. Year	*r*	N
Asscher, Dekovic, Manders, van der Laan, Prins, & van Arum	2014	0.22	256
Borduin, Mann, Cone, Henggeler, Fucci, Blaske, & Williams	1995	0.58	176
Borduin, Schaeffer, & Heiblum	2009	0.43	48
Butler, Baruch, Hickey, & Fonagy	2011	-0.02	108
Giles (boys)	2003	0.03	30
Giles (girls)	2003	0.01	30
Henggeler, Clingempeel, Brondino, & Pickrel	2002	-0.08	118
Henggeler, Letourneau, Borduin, Schewe, & McCart	2009	0.17	127
Henggeler, Melton, & Smith	1992	-0.04	84
Henggeler, Melton, Brondino, Scherer, & Hanley	1997	0.07	155
Ogden & Hagen	2006	0.19	75
Ogden, Hagen, & Andersen	2007	0.18	105
Painter	2007	0.09	174

Appendix I
Listing of Author(s), Publication Year, Effect Size & Sample Size—Problem Behavior

Author(s)	Pub. Year	*r*	N
Rowland, Halliday-Boykins, Henggeler, Cunningham, Lee, Kruesi, & Shapiro	2005	0.02	31
Sheidow, Bradford, Henggeler, Rowland, Halliday-Boykins, Schoenwald, & Ward	2004	0.02	115
Smith-Toles	2003	0.09	155
Sundell, Hanson, Lofholm, Olsson, Gustle, Kadesjo	2008	-0.04	156
Timmons-Mitchell, Bender, Kishna, & Mitchell	2006	0.55	93
Weiss, Han, Harris, Catron, Ngo, Caron, Gallop, & Guth	2013	0.03	164

APPENDIX J—LISTING OF AUTHOR(S), PUBLICATION YEAR, EFFECT SIZE, & SAMPLE SIZE—PSYCHOPATHOLOGY AND MENTAL HEALTH

Appendix J
Listing of Author(s), Publication Year, Effect Size & Sample Size—Psychopathology and Mental Health

Author(s)	Pub. Year	*r*	N
Borduin, Mann, Cone, Henggeler, Fucci, Blaske, & Williams	1995	0.04	176
Borduin, Schaeffer, & Heiblum	2009	0.41	48
Butler, Baruch, Hickey, & Fonagy	2011	0.00	108
Henggeler, Clingempeel, Brondino, & Pickrel	2002	-0.06	118
Henggeler, Halliday-Boykins, Cunningham, Randall, Shapiro, & Chapman	2006	0.02	161
Henggeler, Melton, Brondino, Scherer, & Hanley	1997	0.24	155
Letourneau, Henggeler, Borduin, Schewe, McCart, & Chapman	2009	0.14	127
Manders, Dekovic, Asscher, van der Laan, & Prins	2013	0.03	256
Mann, Borduin, Henggeler, & Blaske	1990	0.63	45
Ogden & Hagen	2006	0.28	75
Ogden, Hagen, & Andersen	2007	0.27	105
Painter	2009	-0.08	174
Rowland, Halliday-Boykins, Henggeler, Cunningham, Lee, Kruesi, & Shapiro	2005	0.12	31

Appendix J
Listing of Author(s), Publication Year, Effect Size & Sample Size—Psychopathology and Mental Health

Author(s)	Pub. Year	*r*	N
Sheidow, Bradford, Henggeler, Rowland, Halliday-Boykins, Schoenwald, & Ward	2004	0.04	115
Smith-Toles	2003	0.08	155
Sundell, Hanson, Lofholm, Olsson, Gustle, Kadesjo	2008	0.14	156
Swenson, Schaeffer, Henggeler, Faldowski, & Mayhew	2010	0.30	90
Timmons-Mitchell, Bender, Kishna, & Mitchell	2006	0.33	93

APPENDIX K—LISTING OF AUTHOR(S), PUBLICATION YEAR, EFFECT SIZE, & SAMPLE SIZE—FAMILY FUNCTIONING AND RELATIONSHIPS

Appendix K
Listing of Author(s), Publication Year, Effect Size, & Sample Size—Family Functioning and Relationships

Author(s)	Pub. Year	*r*	N
Borduin, Mann, Cone, Henggeler, Fucci, Blaske, & Williams	1995	0.21	176
Borduin, Schaeffer, & Heiblum	2009	0.22	48
Brunk, Henggeler, & Whelan	1987	0.12	43
Butler, Baruch, Hickey, & Fonagy	2011	-0.03	108
Henggeler, Melton, & Smith	1992	0.17	84
Henggeler, Melton, Brondino, Scherer, & Hanley	1997	-0.01	155
Henggeler, Rowland, Randall, Ward, Pickrel, Cunningham, Miller, Edwards, Zealberg, Hand, & Santos	1999	-0.05	116
Mann, Borduin, Henggeler, & Blaske	1990	0.34	45
Ogden & Halliday-Boykins	2004	-0.10	100
Painter	2007	0.03	174
Rowland, Halliday-Boykins, Henggeler, Cunningham, Lee, Kruesi, & Shapiro	2005	0.07	31
Smith-Toles	2003	0.18	155
Timmons-Mitchell, Bender, Kishna, & Mitchell	2006	0.66	93
Weiss, Han, Harris, Catron, Ngo, Caron, Gallop, & Guth	2013	0.00	164

APPENDIX L—LISTING OF AUTHOR(S), PUBLICATION YEAR, EFFECT SIZE, & SAMPLE SIZE—PEER RELATIONSHIPS

Appendix L
Listing of Author(s), Publication Year, Effect Size, & Sample Size—Peer Relationships

Author(s)	Pub. Year	*r*	N
Asscher, Dekovic, Manders, van der Laan, & Prins	2013	0.11	256
Borduin, Mann, Cone, Henggeler, Fucci, Blaske, & Williams	1995	0.02	176
Borduin, Schaeffer, & Heiblum	2009	0.51	48
Butler, Baruch, Hickey, & Fonagy	2011	0.11	108
Henggeler, Letourneau, Borduin, Schewe, & McCart	2009	0.07	127
Henggeler, Melton, & Smith	1992	0.02	84
Henggeler, Melton, Brondino, Scherer, & Hanley	1997	0.00	155
Henggeler, Rowland, Randall, Ward, Pickrel, Cunningham, Miller, Edwards, Zealberg, Hand, & Santos	1999	-0.04	116
Sundell, Hanson, Lofholm, Olsson, Gustle, Kadesjo	2008	0.09	156

APPENDIX M—LISTING OF AUTHOR(S), PUBLICATION YEAR, EFFECT SIZE, & SAMPLE SIZE—SCHOOL PERFORMANCE

Appendix M
Listing of Author(s), Publication Year, Effect Size, & Sample Size—School Performance

Author(s)	Pub. Year	*r*	N
Borduin, Schaeffer, & Heiblum	2009	0.53	48
Brown, Henggeler, Schoenwald, Brondino, & Pickrel	1999	0.18	118
Henggeler, Rowland, Randall, Ward, Pickrel, Cunningham, Miller, Edwards, Zealberg, Hand, & Santos	1999	0.23	116
Oakley	2000	0.31	58
Painter	2007	-0.05	174
Rowland, Halliday-Boykins, Henggeler, Cunningham, Lee, Kruesi, & Shapiro	2005	0.22	31
Sundell, Hanson, Lofholm, Olsson, Gustle, Kadesjo	2008	-0.59	156
Timmons-Mitchell, Bender, Kishna, & Mitchell	2006	0.57	93
Weiss, Han, Harris, Catron, Ngo, Caron, Gallop, & Guth	2013	0.01	164

APPENDIX N—LISTING OF AUTHOR(S), PUBLICATION YEAR, EFFECT SIZE, & SAMPLE SIZE—PARENT FUNCTIONING AND RELATIONSHIPS

Appendix N
Listing of Author(s), Publication Year, Effect Size, & Sample Size—Parent Functioning and Relationships

Author(s)	Pub. Year	*r*	N
Asscher, Dekovic, Manders, van der Laan, & Prins	2013	-0.01	256
Borduin, Mann, Cone, Henggeler, Fucci, Blaske, & Williams	1995	0.15	176
Borduin, Schaeffer, & Heiblum	2009	0.40	48
Brunk, Henggeler, & Whelan	1987	0.41	43
Butler, Baruch, Hickey, & Fonagy	2011	0.01	108
Henggeler, Letourneau, Borduin, Schewe, & McCart	2009	-0.02	127
Henggeler, Melton, & Smith	1992	-0.14	84
Henggeler, Melton, Brondino, Scherer, & Hanley	1997	0.08	155
Huey, Henggeler, Rowlland, Halliday-Boykins, Cunningham, Pickrel, & Edwards	2004	0.00	116
Mann, Borduin, Henggeler, & Blaske	1990	0.50	45
Schaeffer, Swenson, Tuerk, & Henggeler	2013	0.45	43
Sundell, Hanson, Lofholm, Olsson, Gustle, Kadesjo	2008	-0.06	156
Swenson, Schaeffer, Henggeler, Faldowski, & Mayhew	2010	0.23	90
Weiss, Han, Harris, Catron, Ngo, Caron, Gallop, & Guth	2013	0.32	164

APPENDIX O—LISTING OF AUTHOR(S), PUBLICATION YEAR, EFFECT SIZE, & SAMPLE SIZE—SUBSTANCE ABUSE

Appendix O
Listing of Author(s), Publication Year, Effect Size, & Sample Size—Substance Abuse

Author(s)	Pub. Year	*r*	N
Henggeler, Halliday-Boykins, Cunningham, Randall, Shapiro, & Chapman	2006	0.24	161
Henggeler, Letourneau, Borduin, Schewe, & McCart	2009	0.29	127
Henggeler, Pickrel, & Brondino	1999	-0.04	118
Henggeler, Rowland, Randall, Ward, Pickrel, Cunningham, Miller, Edwards, Zealberg, Hand, & Santos	1999	-0.03	116
Rowland, Halliday-Boykins, Henggeler, Cunningham, Lee, Kruesi, & Shapiro	2005	-0.34	31
Sundell, Hanson, Lofholm, Olsson, Gustle, Kadesjo	2008	-0.02	156
Timmons-Mitchell, Bender, Kishna, & Mitchell	2006	0.28	93
Weiss, Han, Harris, Catron, Ngo, Caron, Gallop, & Guth	2013	-0.15	164

APPENDIX P—LISTING OF AUTHOR(S), PUBLICATION YEAR, EFFECT SIZE, & SAMPLE SIZE—SERVICE UTILIZATION

Appendix P
Listing of Author(s), Publication Year, Effect Size, & Sample N—Service Utilization

Author(s)	Pub. Year	*r*	N
Henggeler, Halliday-Boykins, Cunningham, Randall, Shapiro, & Chapman	2006	0.12	113
Henggeler, Pickrel, & Brondino	1999	0.07	118
Mayfield	2011	-0.14	75
Ogden & Hagen	2006	0.20	31
Ogden, Hagen, & Andersen	2007	0.42	161
Rowland, Halliday-Boykins, Henggeler, Cunningham, Lee, Kruesi, & Shapiro	2005	0.41	105
Schaeffer, Swenson, Tuerk, & Henggeler	2013	0.20	156
Schoenwald, Ward, Henggeler, & Rowland	2000	0.52	90
Sundell, Hanson, Lofholm, Olsson, Gustle, Kadesjo	2008	0.15	43
Swenson, Schaeffer, Henggeler, Faldowski, & Mayhew	2010	0.28	202