

One Step Forward



Lessons Learned from a Randomized Study of Multisystemic Therapy in Canada

Alison Cunningham, Director of Research and Planning

Five years ago, a committed and energetic group of people in four southern Ontario communities embarked upon a process that brought a promising intervention for serious young offenders to Canada. Multisystemic Therapy (MST) had attracted attention in the United States where two randomized studies showed dramatic success in reducing arrests and incarceration.

Ontario's Ministry of Community and Social Services supported the MST project because it promised to be a cost-efficient way of reducing youth crime. Reductions in offending would, in turn, reduce both losses to crime victims and costs associated with criminal justice processing.

The National Crime Prevention Centre (NCPC) supported the evaluation to learn if MST could work in Canada as well as it had in the United States. The follow-up will end in 2004, and few research questions will be left unanswered.

There are two parts to this document. The first is a background of the MST project including interim research results on 407 youth. The second is a discussion of "lessons learned" and the related recommendations for policy makers, funding bodies, and researchers. This discussion begins with a description of 10 different ways the wrong conclusion could have been made about the effectiveness of MST in Canada, had a less rigorous methodology been used.

...continued page 2

Highlights

- The focus here is a 4 year randomized study of MST in four Ontario communities
- 409 youth participated. Half received MST and half continued with the usual services in the local system
- the experimental design was used to answer this question: will MST be followed by lower levels of criminal conviction than existing services in Ontario?
- the 2 groups are not distinguishable on any outcome measure suggesting there is no treatment effect
- after 3 years, 79% of youth had been convicted at least once
- there is a promising trend of a lower number of subsequent convictions for MST recipients
- the 3 year follow-up ends in 2004
- without the randomized design, the wrong conclusion would probably be made about the effectiveness of MST in Ontario
- many lessons were learned. This study should be the first step in using high-quality research to find effective interventions for young offenders in Canada

This document and the MST evaluation were funded by the National Crime Prevention Centre, www.crime-prevention.org. The views expressed here do not necessarily reflect the opinions of the National Crime Prevention Centre or the Department of Justice.

Additional copies can be downloaded from www.lfcc.on.ca

© 2002 Centre for Children and Families in the Justice System, London Family Court Clinic Inc. 200 - 254 Pall Mall St., London, Ontario, Canada N6A 5P6.

Various observations and recommendations flow from the lessons learned. The biggest lesson is clear: the time and effort spent on rigorous research pays off in information that informs the search for effective interventions.

Conversely, research that falls short of accepted standards of scientific rigour – unfortunately the norm in Canada – could be justifying the status quo when better interventions should be sought. It might even be pushing practice in the wrong direction.

We can look to the United States for examples of how randomized field studies are contributing to the crime prevention knowledge base. While “evidence-based practice” has become a common buzz word, there is little Canadian evidence that can reliably inform our choices of program models. This study suggests caution in assuming

American results will replicate in Canada. Even in the United States, crime prevention is driven more by rhetoric than reality because current research results should really be viewed with no more than cautious optimism.[1]

Some may be tempted to label this study a failure because we are not able to recommend the adoption of MST in Canada. Quite the opposite. We learn a great deal from finding out what does not work. MST is probably not the answer for this client group, but the current interventions did not fare well either. It would be a mistake to take these results as proof that existing practice is effective.

This study puts us one step forward in the journey to find effective interventions for serious young offenders. It is a worth-while trip because the goal is community safety.

The MST Project

With therapy teams in London, Mississauga, Simcoe County and Ottawa, about 200 families received MST between 1997 and 2001. At the same time, about 200 families continued with the usual services available through the local youth justice and social service systems. These services most typically took the form of probation supervision augmented as seen necessary by referral to specialized programming.

Group assignment was accomplished by random assignment – like flipping a coin – so the two groups were equivalent at the outset. That being true, the behaviour of the usual services group reflects the behaviour of the MST youth, had they not received MST. In this way, any post-intervention differences can be attributed to the MST.

The experimental design was needed to answer this question: Will MST be more effective in reducing criminal behaviour in serious young offenders than the services already available in these four areas of Ontario?

Rationale for the Project

Both the federal and Ontario governments agreed to support the project for these reasons:

- **to reduce the costs associated with custody**
If MST could reduce crime, it would reduce the need for custody sentences. The majority of Ontario’s youth corrections budget is spent on custody, leaving little funding for community-based programs. For example, for our sample of 409 youths, one third of whom had no prior criminal record, \$3.5 million had already been

spent on direct custody costs at referral (excluding detention). In the follow-up period, so far, the 53% with convictions collectively spent almost 12,000 days in open custody and more than 10,000 days in secure custody, for an estimated cost of over \$6.5 million (excluding detention and adult prison sentences). Extrapolate these numbers to the provincial level, add non- institutional correctional costs such as probation supervision, and it is clear that a great deal of money is spent on a strategy with no evidence of effectiveness.

- **to find an alternative to custody**
The federal government wants to discourage the use of custody, in part by promoting alternative sentencing options. In Ontario youth courts, 40% of convictions in 1999/00 ended in a custody sentence. While many members of the public might be supportive, there is no evidence to suggest existing custody programs can ameliorate the underlying causes of criminal behaviour, when offending is linked to family and emotional problems. Conversely, anecdotal evidence suggests that custody stays may be harmful and could actually increase recidivism. If reducing crime is the goal, we need to find and validate ways to deliver intensive services in the community.
- **to examine cost-efficiency**
The NCPC wants to document how money spent on interventions with youth will reduce down-stream costs associated with criminal behaviour. Also of interest was if the cost of MST, relatively high compared with other community programs (agency costs of \$6,000 to \$7,000 per family plus payment to MST Services Inc. for training, supervision and licensing), would be financially

recouped by fewer or shorter custody sentences. Funders need to know if cheaper interventions can achieve the same outcomes.

- **to find an intervention for serious or chronic offenders**

Consistent with the spirit of the new Youth Criminal Justice Act, there is a need to differentiate our responses to youth crime on two fronts. One is by devising a range of front-end options that divert minor offences from formal processing. The second is by creating intensive interventions for cases at the opposite end of the spectrum, those involving serious or chronic offenders. While much attention has been paid, and rightly so, to prevention and early intervention programs, there is also a need to intervene with youth who are already deeply involved with criminal behaviour. In this sample of 409 youth, examining their histories of criminal convictions before, during and after the study, about one quarter have been prosecuted four or more times. What can we do for them?

Why MST?

MST was selected as an intervention for study in Ontario because two randomized studies carried out by its developers at the Medical University of South Carolina suggested it might be a cost-efficient, community-based option to reduce crime and keep high-risk youth out of residential placements such as custody. These data were widely reported and made MST a stand-out among delinquency intervention efforts, a field not characterized by overwhelming success. While some meta-analyses of research were showing that some treatment programs can be effective with some youths, improvements were small and translation of “best-practice” models to field environments had proved challenging. Moreover, few were claiming success with the most hard-to-serve youth, whose anti-social behaviour seemed intractable and who were on a trajectory that could well take them to the adult penal system.

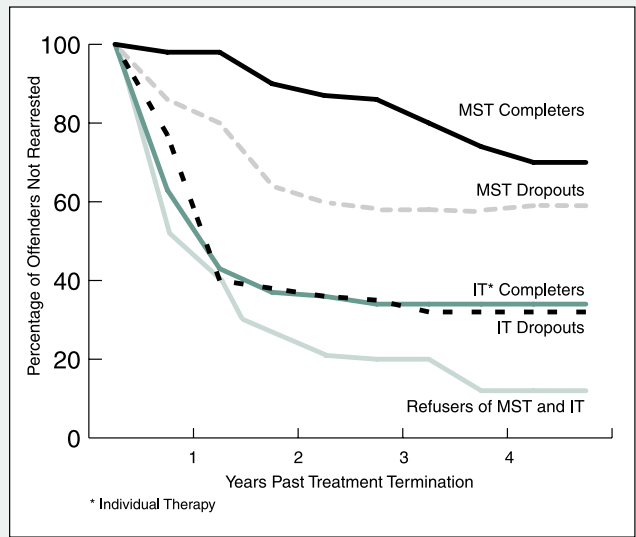
Evidence of MST’s potential to reduce youth crime came from two randomized studies:

- in Simpsonville, South Carolina, 84 violent or chronic juvenile offenders at imminent risk of out-of-home placement were randomly assigned to MST or to the usual services of the Department of Juvenile Justice. At 59-weeks, the MST youth averaged .87 arrests while the usual services group averaged 1.52. There was also a lower average length of incarceration (5.8 versus 16.2 weeks). In a 2.4 year follow-up, MST doubled the percentage of youth who were not re-arrested, in comparison with usual services.[2]

- in Columbia, Missouri, 176 juvenile offenders aged 12 to 17 were randomly assigned to MST or individual counselling. Also included were 24 who refused to participate. After four years, the re-arrest rates of five groups were compared: MST completers (n=77), MST drop outs (n=15), therapy completers (n=63), therapy drop outs (n=21) and the 24 who refused to participate. Results are illustrated in Figure 1, where it can be seen that dramatic differences were evident.[3]

Modest but not statistically significant differences in arrest were found in Charleston[4] and a multi-site study in South Carolina.[5] A fifth study, on 16 sex offenders, should be interpreted with caution because of the small sample.[6]

Figure 1
Survival Curve for Columbia, Missouri MST Study (n=200)



Source: S.W. Henggeler (1997). *Treating Serious Anti-Social Behavior in Youth: The MST Approach*. *OJJDP Juvenile Justice Bulletin*, May, at 5.

What is MST?

MST is an intensive, home-based intervention for serious young offenders that adopts the family preservation modality of intervention. Services are delivered to the whole family (rather than the “identified client” of the youth), targeted at those with the most serious criminal behaviour, time-limited (one to three months), flexibly scheduled to meet the family’s needs, delivered in the home, tailored to the needs of family members, provided in the context of a family’s values, beliefs and culture, and available 24 hours a day, seven days a week. MST therapists carry small caseloads of four to six families and will probably make several visits each week to the family home for an average of two to 15 contact hours each week, more in the early weeks and less as the case nears closure.

The assumptions underlying the development and use of MST are:

- youthful criminal behaviour is influenced by many factors in their social world (family, peers, school, etc.) so treating the youth in isolation of these factors will be ineffective
- the social world, or ecology, should be the “identified client” rather than the individual
- most youthful criminal behaviour is transitory and will desist with the passage of time so intensive interventions should be targeted at those who are most likely to continue offending into adulthood
- most justice interventions are ineffective in reducing criminal behaviour in youths most likely to continue offending into adulthood
- most residential treatment programs are ineffective in addressing the needs that prompted admission
- most justice interventions place youths in close contact with other criminally involved youth which can increase their likelihood of re-offending
- every youth is unique so an intervention should be tailored to individual needs and the circumstances of their social world that create criminal behaviour and serve as barriers to its reduction
- traditional focus on problem areas should be augmented with attention to the identification and amplification of strengths in both youths and their ecology
- social service intervention is always episodic so the parent or parent surrogate is the key agent of change and should be

empowered to make and sustains the gains

The MST process begins with the identification of problem behaviours, a task which involves the whole family. While the focus is on elimination of problems, this is accomplished in great measure by identifying and building on strengths. Next comes an assessment of the factors in the youth’s ecology supporting the continuation of problem behaviours and the obstacles to their elimination. These factors may be found in any sphere of the youth’s ecology or the linkages among them so therapists go to the school, spend time with the peer group, or speak with the extended family.

By identifying the “fit” between the problems and the broader systemic context, MST workers are defining both the targets of intervention and the indicators of effectiveness of the measures undertaken. Depending upon the family, interventions might include improving caregiver discipline practices, enhancing family affective relations, decreasing association with deviant peers, increasing association with pro-social peers, improving school or vocational performance, engaging youth in pro-social recreation, and/or developing an indigenous support network of extended family, neighbours, and friends to help caregivers achieve and maintain such changes.[7]

A therapeutic strategy should produce observable results in behaviour or the strategy is revised. In other words, desired change in the behaviour (e.g., school attendance) shows the intervention (e.g., parent contacting the school daily) is on the right track. Failure to achieve change requires a reassessment of the “fit” and plainly indicates the need to try a new tactic. The therapist is ultimately accountable for overcoming barriers to change. Blaming language such as

The Nine Principles of Multisystemic Therapy

- 1 The primary purpose of assessment is to understand the “fit” between the identified problems and their broader context**
- 2 Therapeutic contacts should emphasize the positive and should use systemic strengths as levers for change.**
- 3 Interventions should be designed to promote responsible behaviour and decrease irresponsible behaviour among family members.**
- 4 Interventions should be present-focussed and action-oriented, targeting specific and well-defined problems.**
- 5 Interventions should target sequences of behaviour within or between multiple systems that maintain the identified problems.**
- 6 Interventions should be developmentally appropriate and fit the developmental needs of the youth.**
- 7 Interventions should be designed to require daily or weekly effort by family members.**
- 8 Intervention efficacy is evaluated continuously from multiple perspectives with providers assuming accountability for overcoming barriers to successful outcomes.**
- 9 Interventions should be designed to promote treatment generalization and long-term maintenance of therapeutic change by empowering care givers to address family members’ needs across multiple systemic contexts.**

“sabotage” and “resistance” is not permitted. In fact, diagnostic labels of any type are discouraged in favour of a perspective that focuses on challenges and strengths.

MST was designed for chronic or violent young offenders and their families. Most youth had a constellation of other presenting problems including school refusal, aggression, substance abuse, non-compliance, risk taking, or severe parent/child conflict. An intensive training regime overseen by MST Services Inc. of Charleston, South Carolina, begins with a one-week orientation to the nine treatment principles. This is followed by on-going weekly consultations on each case as well as quarterly booster training sessions. There is stringent and continuous monitoring of adherence to the method. It takes about one year before even an experienced therapist becomes proficient with MST and MST-specific supervision must be provided thereafter to maintain fidelity to the model.

Partners in the Project

The Ontario MST project was made possible by two levels of government and by the cooperation of eight community agencies. Funding came from the Ministry of Community and Social Services. They also funded the initial training, consultation and supervision of MST Services Inc. The Ministry of Corrections joined the project in the last year.

MST services were provided by or with the cooperation of these agencies:

Simcoe County

- Kinark Child and Family Services (Barrie)
- New Path Child & Family Counselling Services of Simcoe County

London

- Craigwood Youth Services (lead agency)
- Madame Vanier Children’s Services

Mississauga Area

- Associated Youth Services of Peel

Ottawa

- Crossroads Children’s Centre
- Eastern Ontario Young Offender Services (lead agency)
- Youth Services Bureau

All are either children’s mental health centres or agencies that counsel youth. They are private, non-profit agencies which operate largely with government funding. All were funded at the same level for the MST provision but for some agencies this was new money and for others it was a redistribution within an existing budget.

The evaluation was designed and carried out by the Centre for Children and Families in the Justice System of the

London Family Court Clinic. The first year of the evaluation was funded by the Ministry of Community and Social Services and the last three years were funded by the National Crime Prevention Centre in Ottawa.

The Evaluation Strategy

The American studies suggested that MST might contribute to public safety in Ontario by reducing criminal offending which translates into cost savings in two important sectors: 1) state-paid justice services; and 2) the losses of crime victims. Those two studies suggested MST was an efficacious intervention (see Table 1). Still unknown was the issue of effectiveness. Could MST be implemented under conditions typically found in the field, and would the South Carolina results be replicated? The NCPC was also interested in its efficiency or, more specifically, its cost efficiency.

It was decided early in the developmental phase that use of MST in Ontario had to occur within the context of a randomized study, in order to answer questions of effectiveness and efficiency. While rarely if ever used in Canadian corrections, this methodology was fast becoming the norm in other countries.[8] A randomized field trial was necessary to test the effectiveness of MST in Ontario compared with existing services, to determine the generalizability of American findings to Canada, and to address the need for an evaluation conducted independently of the developers of MST.

Other important features of the study were intake screening against inclusionary and exclusionary criteria, a large sample, a valid measure of outcome, and long-term follow-up. The data collection strategy was specifically designed to answer research questions posed by stakeholder groups. Considerable care and expense were expended to ensure fidelity to the treatment model. The outcome measure involved real behaviour in the community, not in-program changes in attitudes or clinical symptoms. The research was designed and conducted by investigators independent of the method’s developer, the funder, and the agencies delivering the program.

The multi-site nature of the project permitted comparisons across different communities under variable conditions of implementation. The intent was to implement the same intervention across the sites. All teams had the same training, were supervised by the same consultant, and met quarterly for boosters. A standard research protocol was used.

An important feature of the study was monitoring the

Table 1

The Three E's of Outcome Evaluation

EFFICACY	this intervention can work under rigorous conditions of implementation (e.g., program is supervised within the context of a well-funded research study)
EFFECTIVENESS	this intervention can work when implemented in the “real world”
EFFICIENCY	this intervention achieved the same (or better) outcomes at less cost per unit of outcome when compared with other interventions

process of implementation. Being a field trial, the level of success in implementing MST was part of the evaluation.

Participating Youth & Families

At each site, unique referral paths were devised to match local funding arrangements. In two sites (Simcoe County and Mississauga) referrals came only from probation officers of the Ministry of Community and Social Services. All youth had prior convictions or were before the courts for the first time. In London and Ottawa, referral was open to any youth who met the eligibility criteria of age (12 or over) and assessed risk for future criminal behaviour. In Ottawa, referrals were also accepted for 10 and 11 year olds and these referrals typically came from the youth bureau of the Ottawa Police or from a children’s mental health centre.

In all, 64% were probation referrals, including 11 from the Ministry of Correctional Services. Referrals were excluded if a youth was at risk to commit sexual offences, if nobody was available to act in a parental role, if the family was successfully engaged in therapy elsewhere, if the youth would not be safe in the home, if the youth was acutely psychotic, or if a therapist would be at risk of injury in the family home.

The 409 families came from London (122), the Mississauga area (100), Simcoe County (94) and Ottawa Carleton (93). Characteristics of the youth varied considerably among the sites, a topic discussed at length in the full report. In the sample as a whole, 74% were males. The average age at referral was 14.6, including 27 Ottawa youth who were 10 or 11. Thirteen percent self-identified as Aboriginal. The median family income was \$20,000 to \$30,000 although 36% of families were supported by welfare in the previous year. The rate of lone-parent families was 47%.

About one third of youth had no record of prior criminal convictions at referral although there had to be evidence of

past criminal behaviour in order to qualify for MST. Almost one third (30%) had previously served custody sentences, an average of 47 days.

A clinical profile was drawn from psychometric testing completed by the youth, caregivers and teachers. According to parent ratings, 84% of youth were over the clinical cutoff for conduct problems and half were over the cutoff for depression. Youth self-reports were only slightly lower, with 61% placing themselves over the cutoff for conduct and 48% for depression. One third of parents placed themselves over the cutoff for depression and poor family functioning. Teachers rated the youths as low on academic competence and social skills, placing almost all of them at or below the tenth percentile.

Anticipated Project Outcomes

As noted above, the key research question was this: Will MST be more effective in reducing criminal behaviour in serious young offenders than the services already available to them in southern Ontario? The indicators of success, chosen based upon the research questions of stakeholders, focussed on criminal convictions and rates and length of custody sentences. The research questions of interest were these:

- are recipients of MST less likely to be convicted of offences than youth who did not receive MST?
- do recipients of MST who offend do so after a longer period?
- are recipients of MST less likely to be sentenced to youth custody?
- do recipients of MST spend less time in custody?
- are recipients of MST re-convicted of fewer offences than youths who did not receive MST?
- are those who drop out of MST more likely to offend than program completers?

- do recipients of MST commit less serious offences?
- will the cost of the MST intervention be recouped by savings to the correctional system?

Psychological testing was administered at intake and discharge to measure family functioning, caregiver depression, and several facets of youth functioning including social skills, behaviour problems and pro-criminal attitudes.

Criminal behaviour is measured with the CPIC system of the Royal Canadian Mounted Police so there are no missing data and the youth can be tracked wherever they may move across the country.[9] Follow-up tracking continues until 2004.

Interim Results

All interim findings can be found in the 2002 report titled *Effective Interventions for Young Offenders: Interim Results of a Four-Year Randomized Study of Multisystemic Therapy in Ontario, Canada*. [10] The present document has follow-up data until January of 2002, so 407 of the 409 youth have been tracked at least six months since the termination of treatment or, in the case of the control group, since six months after intake into the project. The majority (89%) have been tracked at least one year and 59% have been tracked for two years. Only 28% are included in the three-year data.

Features of offending were measured in a number of ways, both over time and cross-sectionally at six months and at one, two and three years. In brief, the MST group and the usual services group are not distinguishable statistically on any outcome measure. The MST group has slightly better outcomes on about half of indicators while the usual

services group has slightly better outcomes on about half. No differences were statistically significant.

Caution should be exercised in making conclusions based on statistical significance, specifically because of low power (see “A Note on Power”) and

generally because statistical significance is a poor arbiter of treatment effect (see “The Insignificance of Significance”). A large treatment effect would have been obvious and there is certainly no evidence of that. But one cannot dismiss the remote possibility of a treatment effect so small that it is confused with error variance.

This inevitably brings up the issue of practical significance. If a treatment effect is too small to be detected with a sample of 400, how practical is it? Preference should be given to statistics that convey practical rather than statistical significance. The Number Needed to Treat (NNT) statistic and cost-efficiency data are two techniques recommended below. However, it is not possible at this point to identify a treatment effect, either statistically or practically. The two groups entered the study as equal (because of the random assignment) and they are (so far) performing exactly the same, in the aggregate.

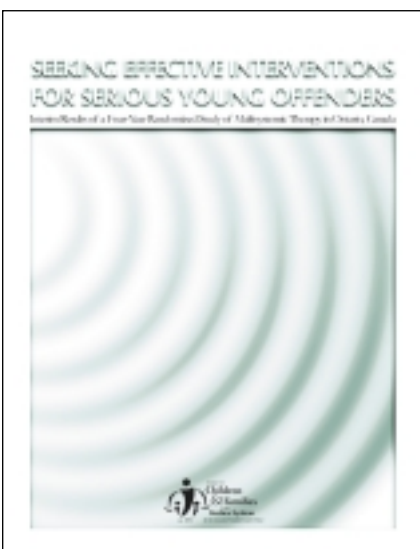
A distribution of the number of offences of conviction is illustrated in Figure 2. Considering prosecutions (i.e., discrete entries on CPIC) the pattern is similar (Figure 3). Survival curves for the samples available in the four follow-up periods are presented in Figures 4 to 7.[15]

Charting the time before a first conviction, both groups averaged about seven months (221 days MST, 229 days usual services). Time to first custody admission was about eight months on average (246 days MST, 263 days usual services).

While the overall incarceration rate, at this point, is slightly lower for MST (38.3%) than usual services (40.4%), there are some interesting patterns. A slightly higher proportion of the MST group has served at least one open custody sentence (32.1% vs. 27.3%). Conversely, the usual services group has a high proportion of closed custody sentences (26.8%) compared with MST (17.2%, $\chi^2=4.9$, $df=1$, $p < .02$).

MST recipients sentenced to custody received longer sentences on average, for both open (101 days vs. 92) and secure custody (125 and 111). These figures will be the total of several discrete sentences if the youth offended more than once. Averaging the costs across the entire sample, using per diem rates of \$255 and \$345 respectively for open and secure custody, the average custody cost per MST recipient was \$15,479 compared with \$16,655 for usual services.

Focussing the analysis only on cases with custody sentences, the average cost for MST so far is \$68,826. For the usual services, the average cost is \$61,768. These figures will rise as the follow-up continues.



The study and follow-up data up to September 2001 are discussed fully in this document.

A Note on Power....

Power is the probability of concluding there is a treatment effect (rejecting the null hypothesis) when there is a true difference between groups in the population, the lower right-hand quadrant of this table.[11] In other words, it is the likelihood a statistical test will yield significance when it should do so. Power is foremost an issue of type II error. When power is low, observed differences might be seen as sampling error when there may, in fact, be a treatment effect.

		THE POPULATION OF ALL SERIOUS YOUNG OFFENDERS	
		Treatment has no Effect	Treatment has an Effect
Study Conclusion	No Treatment Effect	Correct	Type II Error
	Treatment Effect	Type I Error	Correct

Power should never be lower than 0.5 (a 50/50 chance of type II error if there is a treatment effect) and ideally should be 0.95 (a 5% chance of type II error). In practice, 0.8 is an acceptable balance between risk of error and the resources required to study an enormous sample.[12] The consequences of low power are twofold:

1. there is a risk of type II error, concluding that observed between-group differences are sampling error; and
2. relying on tests of statistical significance would be unwise and statistics that convey magnitude of effect would be preferable

Several factors affect power,[13] but sample size is the most obvious. In this field, it is widely reported that the average effective program will reduce criminal behaviour ten percentage points. That being true, a randomized study should have a sample of about 800 to achieve power of 0.8.[14] On the other hand, because MST claimed such a large treatment effect, the most conservative findings published prior to the start of this study suggested that, for 0.8 power, a sample of 100 would be sufficient for replication.

The findings from the Ontario study are characterized by extremely small between-group differences. Using the variable with the largest difference in favour of MST, a retroactive power analysis suggests a sample of 600 (for 0.5 power) for a significance criterion of .05 and a two-tailed test. A sample of 800 would be safer at 0.8 power. A replication of the MST study in Canada would have to be conducted with a large a sample.

Figure 2
Offences of Conviction in Follow-up, Two Groups
(n=407)

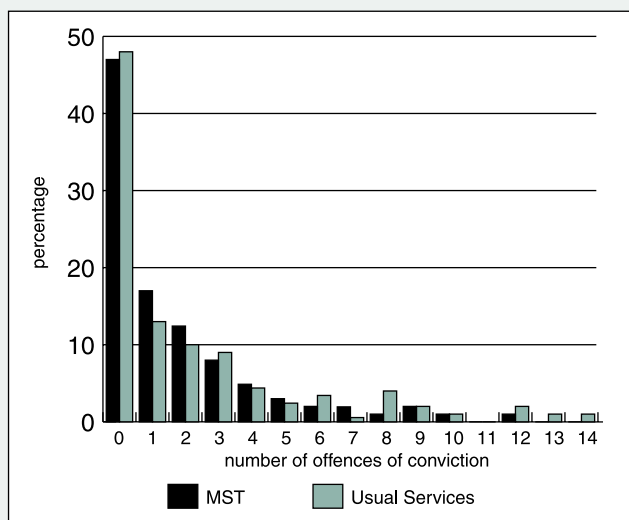
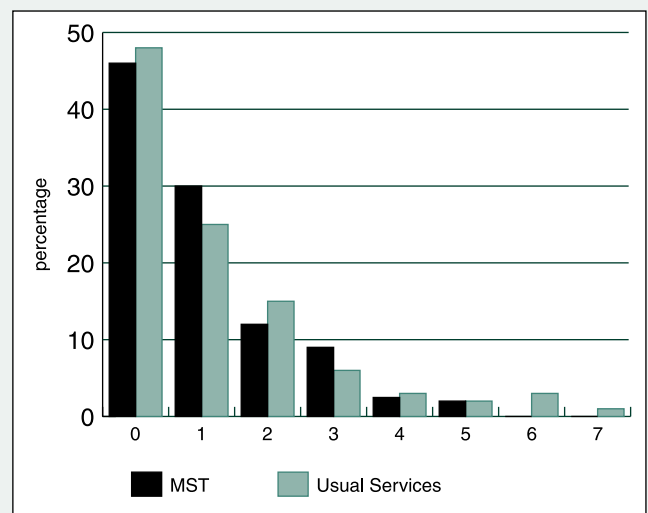


Figure 3
Prosecutions in Follow-up, Two Groups
(n=407)



The Insignificance of “Significance”

Researchers want to answer three basic questions: (a) Is an observed effect real or should it be attributed to chance? (b) If the effect is real, how large is it? and (c) Is the effect large enough to be useful? The first question concerning whether chance is a viable explanation for an observed effect is usually addressed with a null hypothesis significance test [telling] us the probability of obtaining the effect or a more extreme effect if the null hypothesis is true. A significance test does not tell us how large the effect is or whether the effect is important or useful. Unfortunately, all too often the primary focus of research is on rejecting a null hypothesis and obtaining a small p value. The focus should be on what the data tell us about the phenomenon under investigation. This is not a new idea. Critics of significance testing have been saying it for years.[16]

“Significance” is an unfortunate label, inflating the importance of this much misunderstood figure. As many observers have noted, far too much emphasis has been put on “statistical significance” as a threshold for accepting a finding as useful.[17] Null hypothesis significance testing simply evaluates the probability of obtaining the sample outcome because of sampling error. It has little practical value and is affected by many factors, such as type and size of sample, not fully appreciated by all consumers of research. Many programs have been proclaimed successes because pre and post aggregate differences were probably not zero (and don’t even get started on the problems of comparing means[18]). The underlying logic of these statistics is to make inferences from a sample to a general population. A result can be both significant and useless or, worse, completely misleading.

Garbage In, Garbage Out:

Most program evaluations in corrections do not use control or comparison groups, have short or no-post discharge follow-up, do not use behaviour as an outcome measure, and may not use representative samples. Before any statistic is judged, consumers must assess the quality of the process that created that number.

Assumptions of the Test:

The valid use of these tests assumes that the sample has been derived in a way that does not create a bias (e.g., probability sampling), variables are organized at the appropriate level of measurement, and that assumptions of the test have not been violated.

The Null Hypothesis:

So when everything has been done appropriately, what does “significance” mean? In research parlance, it is rejecting the null hypothesis, the hypothesis of no difference or the effect size of 0. Assuming no assumptions have been violated and every aspect of the study conforms to principles of good research, the difference in the groups to which you are generalizing

the findings is probably not zero. As differences increase in magnitude, you can be increasingly confident that they are not the result of chance. In other words, it is highly unlikely (less than a five percent chance), that sampling error was the cause of the difference between the two groups, sometimes termed “proof by contradiction.” A non-significant finding means that we cannot rule out random chance (sampling error) as the reason for the differences, usually because the difference was small.

The Impact of Sample Size:

The larger a probability sample, the closer it approximates the population to which the results will be generalized. Accordingly, the differences can be smaller to reject the null hypothesis. Smaller and smaller differences are needed to reach the level of significance as the sample gets bigger and bigger. One danger in using significance as the determination of importance is that small differences with no practical value may be given undue weight if the sample is large (high power). In reality, a better intervention should be sought. On the other hand, a small sample (low power) will not likely achieve significant findings. One could dismiss a potentially helpful intervention that warrants further study.

P-value:

Another common mis-apprehension is that a finding with a probability value (p) of .001 is stronger than a finding with a p-value of .01. As the value gets smaller, the strength of the association does not grow. But you can be more and more confident the true difference is not zero. For example, a p-value of .04 means that you are 96% confident that the difference is not sampling error. A p-value of .004 simply means you can be 99.6% confident. Ideally, a confidence interval will then be reported as an estimate of the range of values where the true value (i.e., the population value) probably lies.

As an accepted standard, a p-value must be .05 or smaller. So, 5% of significant results will be sampling error. For example, there is one significant correlation

between total Risk/Need Assessment score and one indicator of outcome. The correlation coefficient is .116 and the probability value is .03. In other words, the correlation coefficient is significant at the .03 level so we are 97% confident that we can reject the null hypothesis. However, because it is the only significant correlation among 14 indicators of outcome and the nine inter-correlated RNA scores (i.e., 63 correlations), it is wise to consider the significance of the significance with caution.

Explained Variance:

Another reason to use this number with caution is that the explained variance is so low (i.e., less than 2%) that knowing the value of one variable does not help very much in predicting the value of the other variable. This correlation is significant but not useful. While statistical significance may well be the first step in evaluating outcome differences, research consumers and funders really want to know about practical significance. For this information, they must look elsewhere, to statistics that convey information on the magnitude of association.

Figure 4
Six-Month Survival Data for 407 Youth

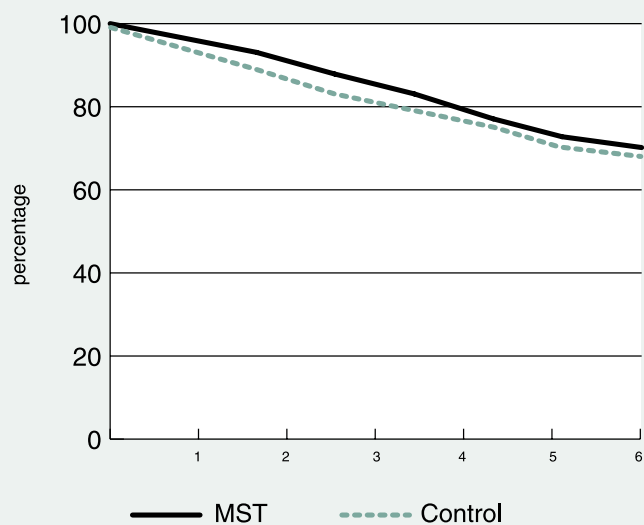


Figure 5
One-Year Survival Data for 363 Youth

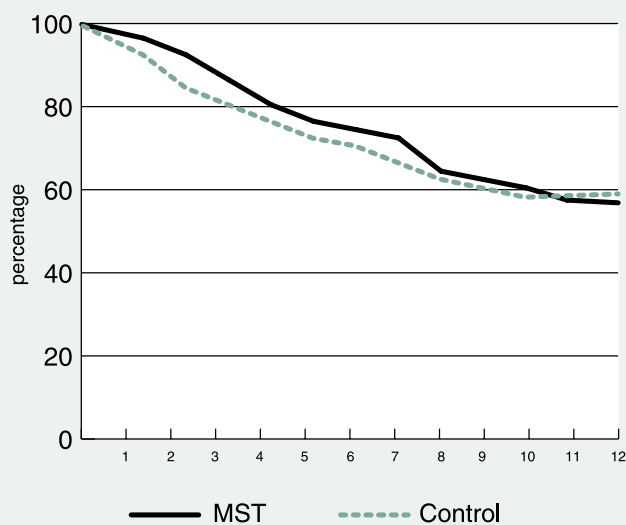


Figure 6
Two-Year Survival Data for 239 Youth

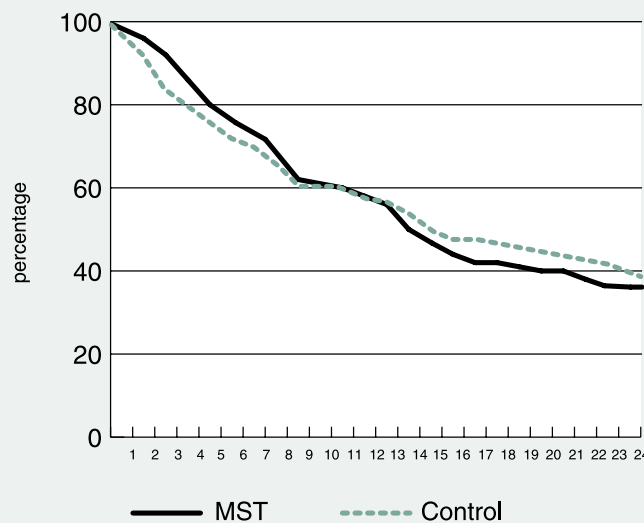
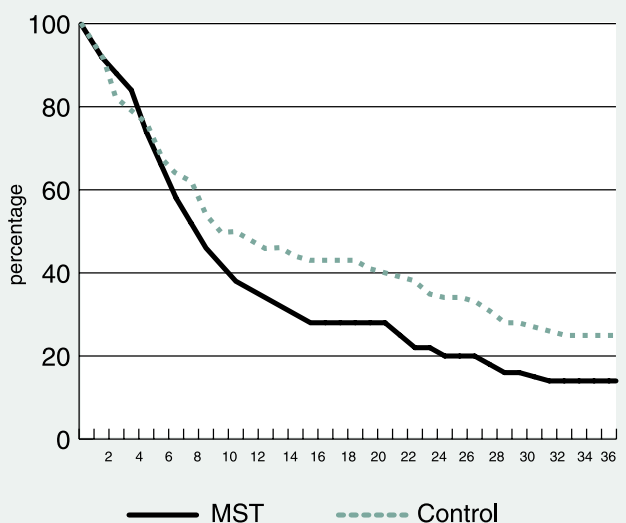


Figure 7
Three-Year Survival Data for 115 Youth



10 Ways to Make the Wrong Conclusion

When program evaluations fail to use classical experimental research designs and instead adopt some arbitrary rule of thumb (whether implicit or explicit) to evaluate program success, the results are quite likely to lead to mistaken conclusions - conclusions that may have serious consequences. The importance of employing experimental group-control group research designs in research evaluations of treatment programs cannot be stressed highly enough.[19]

'How we evaluate' appears to be significantly related to 'what works' in correctional treatment.[20]

In Canada we are in the nascent phase of conducting research that can inform program refinements and wisely direct scarce funding dollars to the best interventions. Randomized field trials can identify interventions which have no effect, so efforts can be re-focused to improving interventions. Randomization has shown how some interventions have an effect opposite to that intended. Boot camps is one example,[21] as is Scared Straight and allied programs.[22]

Twenty years ago, the U.S. Federal Judicial Center (1981) suggested it would be unethical to conduct research on offenders with designs less rigorous than randomization, because of the danger of reaching erroneous conclusions.[23] Indeed, the most interesting lesson learned from this study was the number of ways one might have drawn the wrong conclusion without the randomized design. In other words, the conclusion reached can be related more to how the research was conducted than to the effect of the intervention.[24]

1. Assume the U.S. results would replicate in Canada

- ✗ The first potential mistake in Ontario would have been to assume that the U.S. studies proved MST would be effective in Canada.
- ✓ Because the control group has the same outcomes as the MST recipients, it is unsafe to conclude that the two American studies are sufficient evidence to justify the wide-spread adoption of MST in Canada.

Why? One can only speculate, but the hypotheses cluster into four categories: context, youth, implementation, and research methodology. The usual services available in Ontario are different than those delivered to the youths in South Carolina and Missouri. Local differences in charging and sentencing practices are also likely.

The youth in our sample were not as economically disadvantaged and had less extensive prior records. Even though the MST training and supervision were conducted by MST Services Inc., perhaps the MST in Ontario was different than the MST tested under well resourced conditions in South Carolina and Missouri. And, finally, the outcome measure was conviction rather than arrest, and we tracked recidivism at the national rather than local level.

Many jurisdictions, including 23 American states and several European countries, are implementing MST without first running a randomized trial with their own usual services, their own target clients, and their own therapists.

2. Assume improvements in pre/post testing are because of the treatment

- ✗ Many evaluations take the form of psychological testing administered at intake and again at discharge. Had this been the approach used to evaluate MST, we would have concluded that it was a great success. MST recipients, by self and parent report, improved in the aggregate on all measures, even those that were not explicit treatment targets.
- ✓ The usual service group also improved in all areas measured by psychological testing even though the majority were not targeted by the interventions. In only four of 15 variables was the improvement more pronounced for the MST group (family adaptability, caregiver depression, parent report of child's externalizing symptoms, and youth report of internalizing symptoms).

Why? Improvements in test scores may be regression toward the mean (see Table 2) brought about by the fact that families completed the pre-testing at a time of crisis, when they were willing to consider referral to an intensive intervention. It is also possible that parents amplified the nature of their children's problems in order to qualify for MST, an intervention reserved for the highest risk cases. Whatever the reason, the pre-tests scores were so elevated that we re-scored many tests to verify accuracy of the results. In many cases, scores were in the high end of the clinical range. By the time the post-tests were administered, either both groups had stabilized or the tests were completed more validly, so the scores moved toward the average, which mimicked improvement. In other words, the MST intervention had not caused the improvements because they would have occurred anyway or they were illusory.

Another potential explanation is mortality, because there was only a 62% response rate for the post-tests. While no significant differences were found between responders and non-responders, it is possible that those who failed to respond to the post-tests (or could not be located) were different in some important way. MST drop outs, for example, were not administered discharge testing. Youths in custody at discharge generally did not complete the post-tests. Mortality was not an issue for the recidivism outcomes, however, because 100% of the sample are being tracked.

The control group controls for many other rival plausible explanations of change that might be mis-interpreted as treatment effect: maturation, testing effects, selection, or history. Participants may lie on post-tests to “fake good,” a serious problem in correctional research when release decisions are based on in-program improvements.

3. Assume improvements in pre/post testing will translate to reduced recidivism

- ✗ Because test scores improved so much, it might have been concluded that these improvements signal a reduced probability of criminal behaviour. In many evaluations, this assumption is made.
- ✓ The four areas in which the MST group improved more than the control group did not translate into a lower likelihood of conviction. Moreover, despite dramatic improvements on the psychometric scores, the overall rate of conviction in this sample was extremely high (see Figure 7 above).

Why? The testing scores were not predictive of conviction. Neither intake nor discharge scores on the psychometric tests were correlated with likelihood or extent of conviction in the follow-up period.

4. Use a one-group design

- ✗ At six months, 27.7% of the MST group had been convicted of an offence, a figure that rises to 44.4% at one year, 64.7% at two years, and 85.4% after three years. Are these numbers better or worse than expected? Observers are left to make their own conclusions.
- ✓ There were no differences on any outcome measure using the usual services group as a basis of comparison. For example, in the parallel data to those just listed, 30.8% of the usual services group had been convicted in the first six months of the follow-up, 43.2% after one year, 62.6% at two years, and 73.3% at three years.

Without a basis of comparison, even the most meticulously compiled outcome data have no context for interpretation. For example, using average rates of recidivism in Ontario would not have worked because the members of this sample were purposively selected (and therefore not average) and not all were previously involved with the justice system.

Table 2 Rival Plausible Explanations for Improvements Observed in One-Group Evaluations	
Threat	Alternative Explanations for Apparent Program-Related Changes
Maturation	program participants change over time on their own anyway
Testing	the answers on post-tests will be affected by the fact that project participants already did the same tests before
Regression to the Mean	extreme scores will drift toward the average as time passes
Selection Bias	the people selected (or self-selected) into the program or who finish the program are those most likely to be successful anyway
Mortality	the group at the end is fundamentally different than the group at the beginning because some people have dropped out (e.g., the most problematic individuals have gone to jail so the post-test scores look better because they are gone)
History	intervening events such as changes in laws, policies or program context affect outcomes

5. Compare program completers with drop outs

- ✗ Those who left MST prematurely[25] were more like to be convicted at least once in the follow-up period ($\chi^2 = 6.6$, $df = 1$, $p < .01$) and were more likely to be sentenced to a period in custody ($\chi^2 = 4.3$, $df = 1$, $p < .04$), compared with those who completed MST. The assumption could be that drop outs are not exposed to a high enough dosage of treatment to have reaped any benefit, so they are more likely to fail.
- ✓ The 19% in the MST group who left the program prematurely differed from MST completers on the only variables correlated with likelihood of conviction during

follow-up: age at first conviction ($t = 2.2$, $df = 145$, $p < .03$), number of prior prosecutions ($t = -2.9$, $df = 47$, $p < .006$) and number of offences of conviction in prior record ($t = -2.8$, $df = 46$, $p < .007$). Therefore, it could be predicted that drop outs are more likely to be convicted regardless of their MST dosage.

Drop out is a self-selection bias. Program drop outs probably do poorly for the very reasons they drop out of treatment: lack of commitment to or readiness for change, instability in their lives, incarceration, or needs that exceed the program and must be met elsewhere. To exclude the drop outs would bias the results toward MST, create the perception that therapists could manipulate the outcome of the study, and erode the benefits of randomization by creating unequal groups. Drop outs must be included in the analysis to avoid attrition bias (also called mortality, see Table 2) and because we cannot identify a drop-out group from within usual services (which are varied). In addition, methods of engagement and retention in treatment are key components of MST training so this is a valid part of the test of MST.

It had been hypothesized that drop outs, getting a partial dose of MST, would have better outcomes than the usual services group. This prediction was based on the Missouri study (see Figure 1 above). However, the usual services group had better outcomes than the MST drop outs.

Ultimately, without being able to distinguish the usual service recipients who also dropped out, it is unwise to make conclusions based on drop outs. Moreover, the drop outs had similar outcomes to MST completers and usual services recipients in two sites, and did better than both groups in another site.

6. Use one indicator of outcome

- ✗ With a standardized mean difference effect size of 0.32, MST is effective compared with other interventions in youth justice.[26]
- ✓ Using the same variable, it is also possible to report that MST led to an average of 24% fewer offences per youth who offended. However, using this or any one figure to indicate the success of MST would be a mistake for many reasons.

Research undertaken to *prove* something works is quite different from studying *if* something works. Stakeholders wanted to know if MST works in Ontario. If the goal was to prove MST works in Ontario, those figures alone could be reported here. But researchers should answer the questions

of stakeholder groups, not *be* a stakeholder group.

The effect size of 0.32 and the percentage difference of 24 are calculated from the data in Figure 2, but dropping cases without a conviction. Stakeholders might look at Figure 2 and see little meaningful difference between the two groups. Statistics summarize data by reducing a lot of figures to a single number. Precisely because of this power, statistics can also be used to obfuscate or over simplify. Consumers need to examine all the factors that went into crafting a number and researchers must provide sufficient information to permit sound judgements.

For example, examine the effect size of 0.32. Because the MST group has a higher overall rate of conviction at this point (54.1% compared with 52.5%), eliminating non-offenders biases the results in favour of MST. Including non-offenders in the analysis, the effect size drops to 0.17. Eliminating administrative offences, the effect size is 0.25 without non-offenders and 0.10 with non-offenders. So the effect size will vary by question. If the question is, "among youth who offend, does MST reduce the number of offences of conviction?", the effect is 0.32. If the question is, "among all youth who received service, did MST reduce the number of convictions for criminal offences?", the effect is 0.10.

Also, many factors influence number of offences of conviction (besides the number of offences committed) including whether the youth had legal representation, the type of offences, and local charging and plea bargaining practices. A better indicator of the extent of offending might be the number of prosecutions. The effect size for this variable, excluding non-offenders, is 0.19. Adding non-offenders reduces the effect size again. Add a few more months of follow-up to the data set and all these numbers will change.

So which number is correct? The answer is this: no one number can summarize the effect of MST, nor should it. Different stakeholders have different perspectives that should be reflected in study design. In addition, some good results are to be expected by chance. Remember from above that 100 correlations would include five significant results just by chance. The weight of 95 negative findings should warn that the five positive findings could well be spurious. Put another way, when there is no treatment effect, 5% of tests will be significant. Researchers should report all analyses so consumers can judge if the reported improvements may be spurious.

Because multiple indicators of outcome were used here, many effect sizes can be generated, the largest of which is

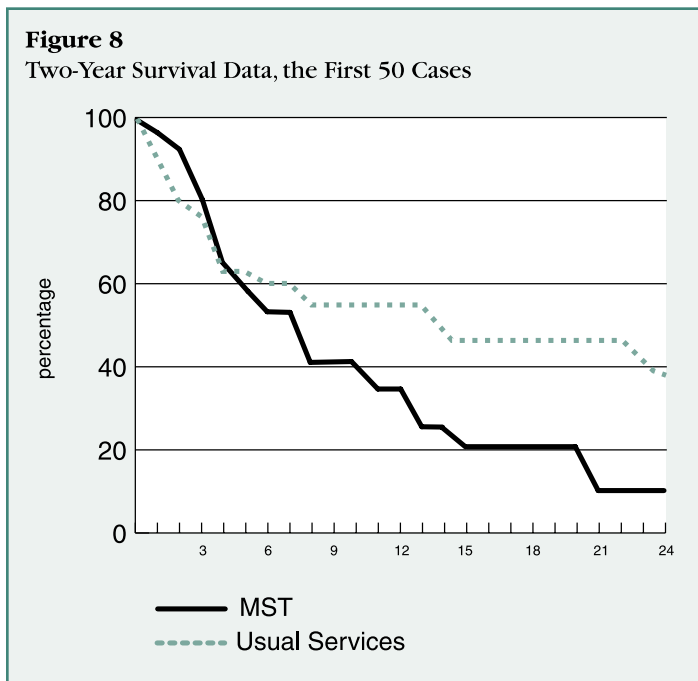
0.33. Most hover around zero. Some of the effect sizes show the superior effect of the usual services because that group had marginally better outcomes on many variables.[27] On the other hand, the MST group evidenced slightly better outcomes on most other variables.[28] All the differences were small and some measures were exactly the same.

When outcome is reduced to one number, stakeholders are left to trust the integrity of the researchers. What unreported analyses did not support hypotheses? What would the conclusion be with different outcome measures? How many studies were not made public because they did not support effectiveness? In correctional research today, political agendas and the profit motive associated with commercial products suggest that it is prudent to be sceptical.

When profit is at stake, is it safe to assume that research is value neutral? The potential for conflict of interest should be considered when selecting evaluators and when interpreting results.

7. Use a small sample

- ✗ Using only the first 50 cases, 25 MST and 25 control, the obvious conclusion is that usual services have a lower rate of conviction (see Figure 8).
- ✓ Using 239 cases over the same time period, the results are illustrated above in Figure 6. The two lines are much closer together and the conclusion is that there is no difference.



Why? Two possible reasons. First, the MST may have been different in these early days. Therapists were getting familiar with the MST method and fidelity to the model might have been low. Indeed, there had been a pilot phase for this reason when 40 cases were treated prior to the start of the research project. The pilot cases were not tracked. In a related vein, the closer monitoring of an intensive in-home intervention may have brought to light offences that would otherwise have escaped official attention. Only when proficiency with the method increased was there an absolute reduction in criminal behaviour.

A more likely explanation is simply random fluctuation. When the sample is small, random assignment is less successful at evening out the differences between groups. For example, among the first 50 cases, by random chance there were more boys in the usual services group (84%) than in the MST group (76%). Four of the 25 youth in the usual services group identified themselves as Aboriginal, compared with only one of the MST recipients. In the MST group, in contrast, the youth were more likely to have prior criminal convictions at referral and half of them had been in custody (compared with only one third of the usual services group). In fact, the MST group had on average twice the number of prior prosecutions. The higher concentration of youths with prior records may explain the difference between the two groups.

In summary, therefore, randomization may not be an effective way of making a control group when the sample is so small. The efforts required to use randomization would have been wasted. With small samples, matching might be a better way of crafting a comparison group, as long as all variables influencing outcome are known in advance. However, a study with such a small sample would not be recommended.

8. Use a short follow-up

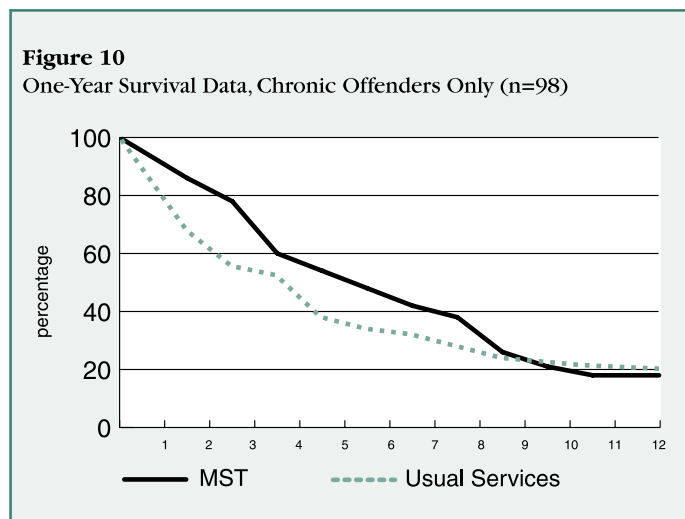
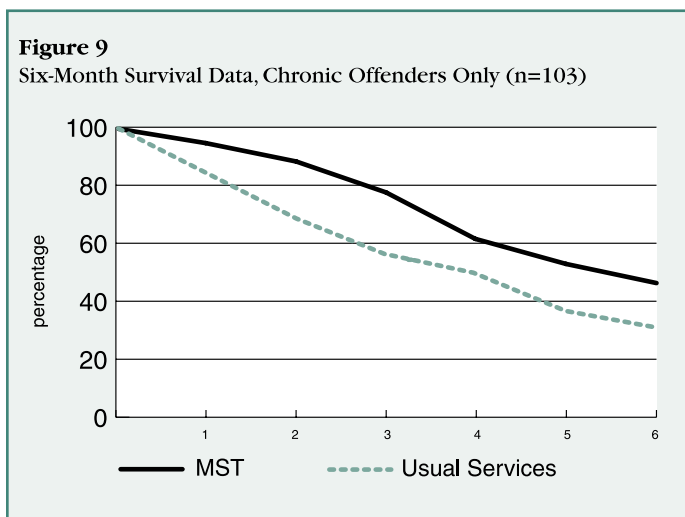
- ✗ After six months of follow-up, 71% of the youth had not been convicted. Youth qualified for the project only if they had high likelihoods of future criminal behaviour, so it might be concluded that both interventions were successful.
- ✓ Tracking the youth for 36 months reveals a very different picture with an overall rate of conviction of 79%. It appears that no intervention was very successful if crime prevention is the target outcome.

Another danger with short follow-ups is that a program could make temporary differences that even out over time. Witness Figure 9 and 10 where only chronic offenders are

included in the analysis (i.e., only those with four or more prosecutions before, during and after the intervention). The intent was to see if MST was more effective for youths with more entrenched criminal behaviour. After examining this hypothesis from many perspectives, it was rejected. However, a study with a six-month follow-up might have come to the wrong conclusion.

Another reason for having a long follow-up, when recidivism is the outcome, is that the court process is typically lengthy and there can be a significant delay between arrest and conviction. There can be a further delay in having the conviction registered on the CPIC system.

A follow-up in the Canadian youth justice system must be at least one year. Refer back to Figure 7 to see the different conclusions would have been drawn in a study with a six-month and a three-year follow-up. Note that the three-year sample of 115 is relatively small, but the between-group differences evident in the first 50 cases have evened out.



9. Study the program in only one place

- ✗ If the study had been conducted in only one of the four sites, the conclusion might be that MST increased the likelihood of criminal conviction.
- ✓ There were fairly dramatic differences in outcome across the sites.

In no site did the MST group perform significantly better than the usual services group, but a great deal was learned from the variation in outcome observed across the four areas. Indeed, this study was (quite wisely) conducted in four places specifically to test the effectiveness of MST in different types of communities under different conditions of implementation.

The communities were selected because they were large and small, urban and rural. As it turned out, there was also differences in the profile of youth referred, especially in terms of prior record, age, assessed risk, and socio-economic status. In one community, almost all youth assigned to usual services were referred to other treatment programs. In the other areas, probation supervision was the most typical usual service but a range of programs were available for referral, rich in the urban areas and sparse in the less-populated areas.

Examining the patterns of outcomes, the ability of MST to reduce criminal behaviour may depend upon the characteristics of the youth referred, the nature of usual services, and the myriad factors that affect implementation. A great deal was learned about the conditions of implementation necessary at the team, agency and community level. Implementing MST was extremely challenging and required a tremendous commitment by the individual therapists, supervisors and agency administrators. These topics are discussed in the full report.

10. Use only one study to indicate a program's effectiveness

It is a key tenet of the scientific method, especially as applied to the social sciences, that confidence in a conclusion strengthens with successive tests. One or two studies will never be enough to accept or reject a hypothesis. This study, even with its multi-site nature, is only one study. Before observers make decisions about the use of MST based on these data, it is important to consider how these results might generalize to their areas. In the full report, the methodology, samples, and findings are described in detail and anyone interested in adopting or rejecting MST should refer to that document.

Likewise, they should examine the particulars of the American studies, especially the characteristics of their

Table 3
Evolution of Questions and Methodologies in Program Evaluation

Purpose	Research Question	Methodologies	Result
NEED	Do we have a program gap in our community?	Needs analysis, stake-holder & community consultation	Decision to persue or abandon program development
	What program do we deliver to fill that gap?	Literature review, consultation with others	Decision to adopt a specific program strategy
PROCESS	Can we implement that program here?	Observation of implement-ation and challenges faced	Conclusion that program is or is not feasible in this community
	Are we meeting the needs of the client group and stakeholders?	Consumer and stake-holder surveys or interviews	Feedback to modify the program (target group, referral stream, method,etc.)
OUTCOME	Do members of the client group make gains in the desired areas?	pre-and post testing or observation, follow-up	Data that documents gains in target areas
	Do members of the target group make more gains than they would have anyway without the program?	Experimental design with control group for comparison	Data that documents effectiveness of the program
EFFICIENCY	Does the program make as many or more gains at less cost than other programs?	Experimental design with control group for comparison	Cost-efficiency analysis

samples, the nature of usual services, and the resourcing of MST delivery. Will those results generalize to your jurisdiction?

MST might work better in some areas than others and there might be communities where it could be useful. In addition, the resources to support implementation are important. However, before a major long-term financial commitment is made, adoption of MST outside the teams already trained should be conducted only in the context of a randomized study, with local youths and local therapists compared with available services.

Lessons Learned

Precisely because this study represents one step forward in the search for effective interventions in Canada, it is prudent to capitalize on what has been learned from this experience.

A Randomized Study was Appropriate for MST...

Program evaluation can take many forms and answer questions of need, process, outcome, or efficiency (see Table 3). Evaluation involves four basic questions, only the last two of which require randomized studies:

- what is the nature and severity of the problem?

- what programs have been deployed to resolve the problem and how well have they been implemented?
- what is the relative effectiveness of alternative programs?
- what is the cost-effectiveness ratio for the alternative approaches?[29]

In other words, experimental methodologies are not necessary for all evaluations, nor should they be used for all evaluations. It is important to match the methodology to a program’s stage of development. For many programs – perhaps most – an experimental evaluation would be premature.

MST was a technique with a strong theoretical underpinning, developed over many years, implemented successfully in several jurisdictions, and tested to verify it could modify desired outcomes. It had a treatment manual and a standardized training program. Moreover, the nature of the research questions under consideration – focussing on outcome, efficiency, and replication in Ontario – dictated the use of a control group.

...But Not Popular

Random assignment was necessary but not popular, to put it mildly. Some referral agents felt it constituted denial of treatment to select qualifying cases and then treat only

half. The concept of a control group was initially confused by some with a placebo, where half the subjects get nothing. In retrospect, the control group received interventions of equal effectiveness but, when the study began, MST was seen by many as a superior service.

To address criticisms of the random assignment, a document was created called Frequently Asked Questions About the Clinical Trial of Multisystemic Therapy in Ontario. Many meetings and presentations were also held to answer questions and allay concerns. The messages were these:

- we do not know if MST is better than the services already available
- we believe the services now available are of high quality
- the families themselves have consented to the random assignment
- the opportunity to try MST only exists because of the research study
- this is the only way to find out if MST works here

Rigorous attention was paid to the ethical tenets of research with human subjects. Everyone knew they had a 50/50 chance at an innovative treatment that may or may not have been more successful than usual services. They also had to understand how much MST would change their lives during the course of treatment, meaning how much work it would mean for the entire family.

Referrals to MST were probably lower than would have been the case without the random assignment. While the teams had the theoretical capacity to service 400 families (meaning a total research sample of 800), about 200 families began MST over the 40-month period referrals were accepted.

Randomization can break down, as famously occurred in an early experiment involving the Scared Straight program in New Jersey.[30] In the Ontario study, the integrity of the assignment process was maintained for the entire four years and we have no reservation in saying that the two groups were the same.

Opposition to randomization is one of the most frequently cited reasons it is not used more often in the justice system. Ethical and legal issues also impinge on the study of some measures such as sentencing options.[31] Even once a randomized study gets off the ground, there is not always clear sailing. Potential problems include:

- gaining and maintaining cooperation from agencies involved in the experiment

- problems in sample size due to overestimates in the size of the eligible population
- resistance and/or hostility to experiments by those within the affected agencies
- hostility engendered outside the affected agencies, and
- high staff turnover [among researchers] caused largely by such hostility.[32]

Weisburd's Principles for Amenity to Randomized Experiments

- 1 There are generally fewer ethical barriers to experimentation when interventions involve the addition of resources to agencies or communities, assuming the control group will continue to receive traditional services**
- 2 There are generally fewer objections to experiments that test sanctions that are more lenient than existing criminal justice penalties**
- 3 Experiments with lower public visibility will generally be easier to implement**
- 4 In cases where treatment cannot be given to all eligible subjects, there is likely to be less resistance to random allocation**
- 5 Randomized experiments are likely to be easier to develop if the subjects of intervention represent less serious threat to community safety**
- 6 Experimentation will be more difficult to implement when experimenters try to limit the discretion of criminal justice agents who traditionally act with significant autonomy and authority**
- 7 It will be easier to develop randomized experiments in systems in which there is a high degree of hierarchical control**
- 8 When treatments are relatively complex, involving multiple actions in the part of criminal justice agents, experiments can become prohibitively cumbersome and expensive and accordingly less feasible to develop**

Source: D. Weisburd (2000). *Randomized Experiments in Criminal Justice Policy: Prospects and Problems*. *Crime and Delinquency*, 46(2): 181-193.

David Weisburd has eight principles to guide researchers in selecting which situations are more or less amenable to randomized experiments.[33]

In great measure to quell the opposition to randomization, alternate strategies have been developed to mimic the benefits, including wait-list controls and assigning participants to one or two other treatments.[34] Randomization may be conducted at several levels of analysis. For example, five of ten schools can be randomly assigned to implement a program. When carefully done, statistical control can also control for many variables.[35]

Sound Research Methods are Crucial

Examining the 10 ways that poor methodologies can lead to the wrong conclusion, it is clear that sound research methods are crucial if the goal is to determine which interventions are effective.

- **match methodology to a program's development**
Outcome evaluations should only be attempted once a program has a track record of successful implementation, measures of fidelity in place, and sufficient funding. When appropriate to the stage of development of a program, use only experimental designs for outcome and efficiency evaluations.
- **fund the evaluation at a level commensurate with expectations**
A review of crime prevention programs sponsored by the U.S. Department of Justice determined that few evaluations met the threshold test for scientific rigour. In other words, few studies were able to advance the debate about "what works," often because of insufficient funding for the research.[36] A key recommendation of that panel was to fund high-quality evaluations of a few promising programs rather than fund non-scientific evaluations of many programs. It was noted that a high-quality evaluation can cost as much or more than the program itself.
- **evaluators should be independent and have no profit motive**
Those who undertake evaluations should be independent of the developers of a program and the agencies at which it is being delivered. More importantly, there should be no potential for conflict of interest such as that which might arise when evaluators are in a position to profit from the findings.
- **use a sample with ample statistical power**
Because studies in this field typically show small effects and modest differences between treatment and control groups, large samples are needed to generate enough statistical power. Studies with less than 0.5 power may only mislead and confuse.[37] That being true, a randomized study in this field should have a sample of at least 600 and ideally 800. In addition, a large sample is important if randomization is being used, to reach the point where it can even out differences between the groups.
- **consider generalizability**
Focus on sample size and power should not divert attention from other equally important features of the sample including representativeness. Is this group appropriate for the intervention? Are the members of the

10 Ways to Make the Wrong Conclusion

- 1 **Assume the U.S. results would be replicated in Canada**
- 2 **Assume improvements in pre/post testing are because of the treatment**
- 3 **Assume improvements in pre/post testing will translate to reduced recidivism**
- 4 **Use a one-group design**
- 5 **Compare program completers with drop outs**
- 6 **Use one indicator of outcome**
- 7 **Use a small sample**
- 8 **Use a short follow-up**
- 9 **Research the program in only one place**
- 10 **Use only one study to indicate a program's effectiveness**

sample the same in every way to those to whom the findings will be generalized? Conversely, the results should not be generalized to groups unlike the study sample.

- **have a basis of comparison, ideally a control group**
Outcomes with no basis of comparison necessitate speculation in interpretation. There is little point to conducting an outcome-focused evaluation where the results will be speculative. A control group is the best basis of comparison.
- **measure target behaviour**
Self-reports of knowledge, attitude, or symptoms should not be the principal outcome measure in crime prevention research. If the goal of a program is to reduce crime, measure crime as the outcome. While official statistics on offending will be underestimates of criminal behaviour, randomization will even out this bias.
- **use multiple indicators of outcome**
It is important that more than one indicator of outcome be used, if only because different stakeholders have different priorities. In addition, it is rarely possible to summarize the results of a program with one variable. Researchers should report findings related to all outcomes, including those that are contrary to hypotheses.
- **follow the participants for at least one year after the program ends**
In the criminal justice field, a follow-up of one year should be the absolute minimum. As this study

demonstrates, two and three year follow-ups add useful information to the conclusions.

- **measure program fidelity**

The generalization of evaluation findings depends upon the quality of program implementation. If results are promising, replication is possible only when detailed information is available. If results are not promising, one must know if poor implementation is the reason.

- **apply the same method in more than one place**

Two factors recommend the testing of a program in more than once place. The first is volume. To achieve a large sample such as 600, pooling data from several sites can be the only practical solution. Second, when the same intervention is delivered in several places, one can learn how implementation context affects the quality of service delivery. Even randomization cannot overcome the weakness of single-site evaluations and multi-site evaluations offer useful information that meta-analyses and replications cannot provide.[38]

Random Assignment did not Control for all Threats to Internal Validity

Had an uncontrolled design been used to test MST in Canada, as already described, the wrong conclusion could have been made. Changes observed in non-controlled studies cannot be unambiguously linked to the program in the face of so many rival explanations for any observed improvements. Some of these rival possible explanations for change, also called threats to internal validity, were listed above in Table 2.

That said, random assignment does not *ipso facto* guarantee a perfect study. Other problems can occur such as inability to gather valid outcome data.[39] This was not a problem for the MST study, but there were drawbacks associated with the fact that people who made key decisions about the youths (e.g., arrest, breach, or sentence) were not blind to group assignment. Youths may have been treated differently depending upon their group assignment.

Some threats to internal validity could not be controlled (see Table 4), such as compensatory rivalry where those assigned to the control group make a conscious effort to improve their behaviour. Compensation would have occurred here if, for example, a probation officer treated members of the control group more leniently when they committed a breach (or came down harder on a MST recipient).

Of concern at some sites was the issue of “diffusion,” or how the youth in the control group might indirectly benefit

from MST principles. In Ottawa, for example, many of the control group members received the services of the Community Support Team (CST) delivered at the same agency as the MST. Moreover, there was one clinical supervisor for both programs and MST principles were diffused into her work with CST. Normally, this diffusion would be a welcome spin off but it is problematic in a randomized trial.

Another potential uncontrolled confound was the possibility that MST increased the likelihood of conviction, especially for offences against the administration of justice. This could happen with improved parental monitoring and communication with the probation officer. One of the few differences apparent between the two groups is that the MST youth were more likely to be convicted of administrative offences such as breach of disposition in the absence of a criminal offence that involved a victim. For example, a youth might have been breached for a technical violation of probation conditions. While MST did not increase the criminal behaviour, it may have increased the extent to which that behaviour was responded to with criminal charges.

Role Clarification

In the initial period of project start-up, there was some confusion about the role of the evaluators vis-à-vis the delivery of MST in the four communities. The role confusion initially came about in great measure because the project had been championed by Dr. Alan Leschied of the London Family Court Clinic. His charismatic and passionate enthusiasm about the potential of MST was crucial to having the myriad pieces of a complicated funding and operational puzzle fall into place.

Another factor that contributed to role confusion in the early days was the multi-site nature of the project. As the provider agencies gained familiarity with the MST approach and worked through a myriad of start-up issues, it was not uncommon to refer questions to the evaluation team. Indeed, there was a need for evaluators to be involved. Too much distance would have made the evaluation unfeasible. And without the cooperation of the program deliverers, the evaluation would simply not have been possible.

Operational decisions properly lay with the agency. However, it was appropriate for the evaluators to monitor the process of training and implementation and to oversee compliance with the research protocol, including the application of inclusionary and exclusionary criteria. Maintaining consistency across sites was also important. In addition, the evaluators had ultimate responsibility for

matters of ethics and Dr. Leschied, as a clinical psychologist and the Principal Investigator, was available to respond to complaints by participants or clinical crises that could not be dealt with locally (neither of which occurred).

In retrospect, as a lesson for multi-site studies, there should be clearly spelled out roles and expectations for evaluators vis-à-vis program providers that permit reasonable access to the required information (e.g., client files) but maintain a distance from operational decisions. For multi-site projects where the same intervention is tested, a central oversight body should be identified at the outset to provide overall guidance on operational issues.

In the end, there never was any such oversight body but the supervisors of the four teams forged close and supportive relationships and consulted each other as questions arose. MST Services Inc. also performed an oversight function by monitoring the delivery of MST.

Securing Appropriate Referrals: The Importance of Buy-in

The sample size (409) was lower than anticipated (800). This is a problem for several reasons. Statistical power was already noted. Cost-efficiency of the intervention was affected by a doubling of per-case cost, because certain costs are fixed regardless of case load. Finally, we are not able to test the effectiveness of MST teams operating at full capacity.

As noted above, lower than expected referrals is common in randomized studies. Two factors are typically posited: over-estimation of eligible cases and lack of support for the study. In this study, one site may have had too few high-risk youth to be a viable. The eligibility “bar” was set high at the outset to forestall the seeping of low-risk youth into the study. Fortunately, teams were able to resist taking inappropriate cases merely to “keep the numbers up.”

A more salient factor was the on-going difficulty of securing enough referrals to keep the teams at full capacity. Of the pool of potential referral agents, a few made referrals while many did not. Not all qualms were related to the randomization. Many features of the intervention itself attracted criticism. At the outset, community presentations were made to introduce both MST and particulars of the study. Over the entire four years, the MST teams spent considerable time securing and maintaining the cooperation of referral agents. Without these efforts, the sample would have been considerably smaller and less suited to MST.

Ultimately, lack of buy-in can compromise the entire study. In Florida, prosecutors used the courts to unsuccessfully

Table 4

Threats to Internal Validity not Controlled for with Randomization

Compensatory Rivalry

- * the control group members try really hard to do better

Diffusion

- * features of the treatment leach into the program delivered to the control group

Compensation

- * those who are treating the control group “compensate” by providing a more intensive service than they normally would or by treating the participants differently

contest random assignment in a study of batterers’ treatment that the local judges supported.[40] In that study, failure to gain buy-in of all key players resulted not only in litigation but in hostility against researchers and low morale.[41]

Cooperation must be sought at all levels, from senior management down to the front line. Because of staff turnover, expect this to be an on-going process. The time required for these tasks must not be under estimated. A thorough survey of the concerns of all parties will assist in responding to specific issues. Also, the research questions of stakeholders should be accommodated in the study design, to ensure they actually do have a stake in seeing the study done well. Perhaps most importantly, the alternate intervention must be one which is acceptable to all parties and that meets client needs. Denial of a service or provision of a lesser quality intervention should be avoided.

Good Research Takes Time

Five years past the start of this study, the conclusions here are still interim. Final results will not be available until 2004. Policy makers and funders wanted feedback on effectiveness quite a bit more quickly. The lesson is that the delay in getting useful information must be clearly communicated at the outset.

Cases were accepted into the research from the summer/fall of 1997 until the end of 2000. At the end of the first year, no outcome data were available because only six research cases had been concluded. Yet many necessary activities had been conducted. Teams had been formed, trained, and finished a pilot period with 40

referrals. A research protocol had been developed and harmonized with local contingencies and an ethical protocol was in place. Supervisors and therapists were trained on the various research tasks that unfortunately fell to them in the absence of funding for on-site researchers. Many presentations had been conducted in the four communities to inform potential referrals sources about the types of cases that are amenable to MST (and to defend the use of random assignment).[42]

Stakeholders were anxious to get findings, especially those related to cost-efficiency. The plan of having the initial CPIC check at one year was revised to six months. Moreover, the data set was structured to provide feedback at six, 12, 24 and 36 months, with analysis of the waves of cases available at these points. No case is included in the follow-up analysis until six months had elapsed from discharge. No case is included in the 12, 24 or 36 month analysis unless they have spent at least that period at risk to offend.

Balancing Scope of Project and Resources

There was a staggering array of possible research questions that could have been addressed with this study. The lesson learned is the wisdom of matching project scope to the resources available. The study as designed was scalable in that there was room to add pieces should additional funding have been secured. However, priority was given to the questions posed by funders. In doing so, the integrity of the project was maintained and the data answered the key research questions designed at the outset.

As it was, the MST teams had to absorb a great deal of the research burden. A study of this scope should have a research assistant at each site.

Treatment Fidelity

When implementing any “off-the-shelf” program with empirical evidence of efficacy, developing and maintaining fidelity to the original model is crucial. To this end, a contract was signed with MST Services Inc. of South Carolina to conduct training and supervision of the four Ontario teams.

Their involvement took the form of a one-week orientation session, weekly review of case summaries by the MST consultant in South Carolina, weekly conference calls with the consultant for case-specific feedback and supervision, and quarterly booster training sessions for all four teams conducted by the MST consultant. In the first year, the cost of this service (\$91,000) was paid by the Ministry of Community and Social Services. As new therapists joined the teams, agencies absorbed the cost of orientation

training, \$750 (USD) plus travel and accommodation.

When the project began, it was assumed that involvement of MST Services Inc. would end after one year. In the interim, new data were released about the role of fidelity in achieving outcomes like those in the Simpsonville and Missouri study. Using an instrument called the Therapist Adherence Measure (TAM),[43] developed at the Family Services Research Center at the Medical University of South Carolina, “modest support” was found for an association between TAM scores and subsequent arrest.[44] These data recommended a high degree of fidelity to MST to replicate the outcomes of Simpsonville and Missouri. It was further concluded that on-going involvement by consultants at MST Services Inc. was necessary to maintain a high level of adherence.

In the second year, because the exchange rate changed, about \$115,000 were paid to MST Services Inc. At that point, the possibility that these costs would continue in perpetuity prompted efforts to move toward independence from MST Services Inc. This was the true test of MST: could it be implemented in Ontario with success? In the last years of the study, the Ministry paid various costs associated with booster training, supervision, site licensing, and the position of an Ontario-based System Supervisor who took on supervisory responsibility for the four teams.

The TAM instrument was used but no correlation was found between TAM score and case outcome. However, assuming that case outcome itself is a valid indicator of service quality, it does seem apparent that levels of success varied over time. There is a visible difference in outcomes over three discrete time periods (Figure 11) that is not so apparent in the usual services group (Figure 12).

The first 50 MST cases, distributed proportionally across the sites, have the highest rate of conviction. The assumption could be that this reflects the learning curve with MST or, as described above, the impact of increased monitoring. The middle 50 cases have the best outcomes. The last 50 cases, fall closer to the initial group. Focussing the analysis on the middle 100 cases, the between-group differences are neither statistically nor practically significant, so any treatment effect during that time was extremely weak. However, these findings highlight the importance of monitoring implementation in some quantitative way.

Therapist Attrition and Burnout

One of the most challenging features of MST on the implementation side was the high degree of therapist attrition. Only one team remained intact for the entire four

years. In consequence, 25 different people delivered MST and average cases per therapist was eight. We learned from MST service providers in the United States that staff turnover was common there as well, with this intensive and demanding intervention.

The degree of turn over was much higher than ever expected at the outset and may have comprised fidelity to the model. In retrospect, because turnover is such a common feature of MST work, having more than one team in each agency would have been desirable. A larger group can better absorb new members without affecting the effectiveness of the overall team. As with many of these post hoc observations, it is probably a suggestion that would not have been possible given the available funding. Suggesting that the teams be larger would probably have compromised the ability to secure the funding.

Useful Description of Data

To attempt a summary of what has been found here, there are no statistically significant differences between the two groups. Moreover, the usual services group has better outcomes on some indicators and the MST group has better outcomes on others. Using the American MST studies as a guide, a sample of 100 is sufficient to discern a treatment effect. Finding no differences in the Ontario sample suggests the American results will not replicate in Canada.

However, precisely because of the small differences, it is retroactively determined that a sample of 409 may not have enough power to validate the use of hypothesis testing so there is a risk of making a Type II error. More research would be needed before the use of MST in Canada can be recommended. This would take the form of a randomized study with 600 to 800 youth, probably at several locations.

This is not a very helpful conclusion for funders and policy makers.

A key problem with statistical significance is that it imposes a yes/no cutoff on a continuous reality. This is also a source of its appeal. The expectation of stakeholders is for a red or green light to the wider use of MST in Canada. An amber light is not welcome after so much time and effort.

Upon reflection, if an effect is so small it can be confused with sampling error, how practical is it? Practical significance is subjective and depends upon the nature of the questions posed by stakeholders.

Suggested replacements for significance tests include measures of the magnitude of between-group differences collectively called effect sizes. For continuous variables, the standardized mean difference is the most popular. For binary outcomes, standardized measures of effect include risk differences and risk ratios. The percentage difference is also commonly reported but there are some caveats.

All these figures can be difficult to interpret and the practical value of any non-zero treatment effect is difficult to determine. Often, readers of research studies are left to puzzle through the meaning of a confusing array of numbers. Some guidance on interpretation is provided here.

Are there better ways to convey practical significance? The “number needed to treat” and outcomes that monetize findings into cost-efficiency statements are probably the most useful. Ultimately, if existing services provide the same outcomes at less cost, it is difficult to argue in favour of adopting MST. Researchers should report findings in terms that are meaningful for stakeholders, so they can draw their own conclusions.

Figure 11
Six-Month Survival Data for Three Time Periods, MST Only

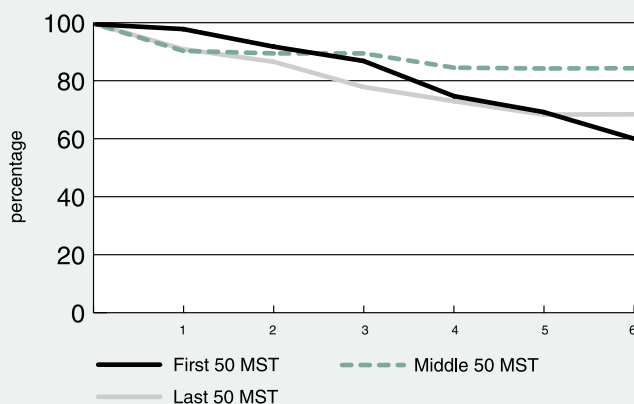
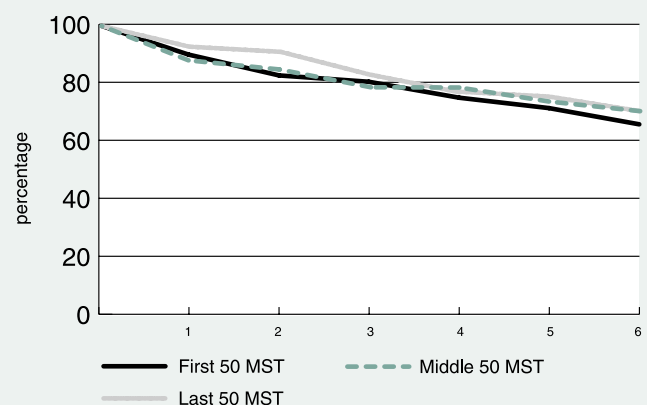


Figure 12
Six-Month Survival Data for Three Time Periods, Usual Services Only



Cohen's *d* for Dummies

Cohen's *d*, or a standardized mean difference, may look like a correlation co-efficient (*r*) but it should not be interpreted in the same way. For example, an effect size of 0.5 is equivalent to a correlation of 0.24.[45] In addition, this effect size has a wider range of values than *r* and can reach to infinity.

Selected Interpretations of Cohen's *d*

Effect Size	Percentile Ranking	Pearson's <i>r</i>	Probability of a Correct Guess
0.0	50	.000	.50
0.2	58	.100	.54
0.5	69	.243	.60
0.8	79	.371	.66
2.0	98	.707	.84

The larger the absolute value of the effect size, the greater the effect. But how big is big? Cohen suggested a convention: 0.2 is small, 0.5 is medium, and anything over 0.8 is large. A 2.0 would be a whopper. Effect sizes can also be thought of as the percentile standing of the average treated youth relative to the average control group member. A 0.0 indicates the mean of the treatment group is at the 50th percentile of the distribution of the control group. With a 0.5, the mean of the treated group is at the 69th percentile of the untreated group. Put another way, an effect size of 0.5 means that, knowing a person's outcome, you could accurately predict whether they were in the control or treatment group 60% of the time. With a 0.0 effect size, you would have a 50/50 chance of being correct.

Another way to interpret effect sizes is to compare them to the results of other studies. Relative to other types of interventions, those aimed at juvenile delinquency seem to have small effect sizes. One widely reported meta-analysis determined the average to be 0.17 compared with, say, positive reinforcement in the classroom (1.17).[46] This is why large samples are needed to detect treatment effect, as already noted.

While significance tests of the null hypothesis should not be used as the arbitrator of success, they can help interpret the practical significance of effect sizes.[47] In the same way that a statistically significant outcome may not be practically meaningful, a practically meaningful outcome (e.g., an effect size) may have occurred by chance. A difference which is not statistically significant and which has a small effect size means that the results are probably not practically significant and unless future studies indicate otherwise, it is prudent to assume that the two groups are not different (see Table 5).

Standardized Mean Difference

When outcomes are measured as a continuous set of numbers, as in numbers of conviction, Cohen's *d* is the most frequently observed statistic. In essence, it is the difference between two means divided by either the common or pooled standard deviation of the two groups. This process reduces the difference between means to standard units to permit comparisons across studies. That, and the fact that it is not distorted by sample size, is the reason it is used so often in meta-analyses.

Standing on its own, it can be difficult to interpret the practical implications (see "Cohen's *d* for Dummies.") and a test of significance may assist in making conclusions (see Table 5).

Risk Ratio

A risk ratio of zero denotes a large difference between two groups while a value close to one shows there is little difference. A value greater than one means the control group did better. Risk ratios for two binary measures of outcome can be found in Tables 6 and 7.

Percentage Difference

The percentage difference can be a straight forward way to quantify the difference between two numbers. Take the incarceration rate of 38.3% for the MST group and 40.4% for usual services. That is a percentage difference of 5.2%. In other words, 5.2% fewer members of the MST group have been sentenced to custody so far.

This statistic can also be misleading. A percentage difference should never be reported independent of the two proportions being compared. For example, there is a 25% difference between 60 and 80. There is also a 25% difference between 3 and 4.

Risk Difference

Sometimes called absolute risk reduction, this is the difference in the proportion of participants in two groups in whom an event (e.g., conviction) is observed. A risk difference of zero indicates no difference between comparison groups. A risk difference less than zero indicates that the intervention was effective in reducing the risk of that outcome.

Number Needed to Treat

In randomized studies, the number needed to treat (NNT) statistic[48] is becoming popular, in great measure because it is a straight forward way to present the practical implications of between-group differences and also because it is amenable to cost efficiency comparisons. Developed for use in clinical trials of medical interventions, the NNT is the number of people who must

Table 5

Fan's Guidelines for Combining Significant Test Outcome with Effect-Size Measure

		EFFECT SIZE		
		← SMALL (.20)	MEDIUM (.50)	LARGE (.80) →
NO SIGNIFICANCE		<ol style="list-style-type: none"> 1. It appears that there is neither statistical nor practical effect 2. Unless future research indicates otherwise, null hypothesis is favoured both statistically and practically 	<ol style="list-style-type: none"> 1. Sample effect looks promising but some caution is warranted in interpreting the effect size by itself because medium effect size could have been the result of chance, even if it may look practically meaningful 2. If one is concerned about Type II error (there is true effect but one fails to find it), look closer at the power of the test because if the sample size is small, one may not have the statistical power to detect potentially meaningful effect 	<ol style="list-style-type: none"> 1. One has some evidence that meaningful effect exists, but a little caution is still warranted about this effect size because large effect size could have occurred by chance when sample size is small 2. If one is concerned about Type II error, look critically at the lack of power of the statistical test 3. Tentatively favour the practical significance of the effect, while keeping an open mind for further research findings
		<ol style="list-style-type: none"> 1. Statistical significance is not accompanied by practical significance and could have been the result of statistical power 2. Considerable caution is warranted in interpreting the statistically significant findings, and they should not be interpreted to mean something practically meaningful 	<ol style="list-style-type: none"> 1. It is very unlikely that the observed effect is due to statistical chance 2. The magnitude of the effect is practically meaningful in many areas of social and behavioural science 3. Conclude that the effect is meaningful both statistically and practically 	<ol style="list-style-type: none"> 1. There is a high degree of certainty that the observed effect is not due to chance statistically, and that the magnitude of the effect is also practically meaningful 2. Conclude with confidence that the effect is meaningful both statistically and practically

Source: X.T. Fan (2001). *Statistical Significance and Effect Size in Education Research: Two Sides of a Coin*. *Journal of Educational Research*, 94 (5): 275-282 at 282.

be treated with an experimental intervention to save the life of one patient compared with the control intervention.[49]

The NNT can be used when the outcome is any adverse incident, such as a criminal conviction or a custody sentence, that can be expressed as a binary variable. The lower the NNT, the stronger the treatment effect compared with the other intervention. A negative NNT means the experimental treatment was followed by a greater proportion of the adverse outcome than was observed in the control group.

Again, the rate of incarceration so far is 38.3% for the MST group and 40.4% for usual services. The difference of 5.2% is not statistically significant. It is a difference, but is it meaningful? In this example, the NNT currently stands at 48. In other words, it is necessary to treat 48 youth with MST to prevent the custody sentence of one youth, compared to what would have happened if the youths received the usual services instead. Cost savings can be calculated if the cost of both interventions and the outcome are known. Similar numbers can be calculated for other binary outcomes.

Table 6

Convictions for Any Offence at Four Time Periods, MST and Usual Services Groups

	MST	Usual Services	Percentage Difference	Risk Ratio	Risk Difference	NNT
Six Months (n=407)	27.7%	30.8%	10.1%	.899	-.031	32
One Year (n=363)	44.4%	43.2%	(2.7)*	1.03*	.012*	negative*
Two Years (n=239)	64.7%	62.6%	(3.2)*	1.03*	.021*	negative*
Three Years (n=115)	85.4%	73.3%	(14.2)*	1.17*	.121*	negative*

* the usual services group has a better outcome

Table 7

Convictions Excluding Administrative Offences Only at Four Time Periods, MST and Usual Services Groups

	MST	Usual Services	Percentage Difference	Risk Ratio	Risk Difference	NNT
Six Months (n=407)	23.4%	26.3%	11.0%	.890	-0.03	34
One Year (n=363)	38.0%	37.5%	(1.3)*	1.01	0.01*	negative*
Two Years (n=239)	55.2%	56.1%	1.6%	.984	-0.01	111
Three Years (n=115)	74.5%	68.3%	(8.3)*	1.09	0.06*	negative*

* the usual services group has a better outcome

Cost Comparison

One of the great advantages of a control group is that a cost-efficiency analysis is possible. For many stakeholders, particularly those who pay for interventions, this is the ultimate test of MST. Can the costs of MST[50] be recouped by savings to the criminal justice system? Since the follow-up began, \$6.5 million were spent on direct costs associated with custody sentences for these 409 youth. As reported above, the cost of the average custody sentence is, so far, over \$60,000. An intervention that reduces crime can save a great deal of money.

A cost-efficiency analysis by an economist is planned by the National Crime Prevention Centre for the MST data.

Factors that May Limit Generalization

If results are to indicate the viability of MST in Ontario, it was important that the sample be typical of youth who might receive MST under non-research conditions. The samples at the four sites are described in the full report. Readers interested in how these findings might generalize to their jurisdictions should consider the referral profile.

Two factors may have affected the nature of the sample relative to the type of youths who would receive MST without the research component.

- **self-selection bias**

The sample may be biased because youths had to agree to the study (but not necessarily the intervention). Parents had to agree to both the intervention and the study. Because of the random assignment, it was crucial that all participants provided informed consent and were free to decline involvement. In most cases where MST was declined, it was parents who resisted the intense, family-focussed and in-home nature of MST.

- **variation in normal referrals**

It is also possible that a different type of youth was referred relative to the youths who would be referred without the research component. Some potential referral agents opted not to make referrals during the study.

What We Would do Differently?

Given what we now know, as a post hoc analysis, how might the study have been done differently?

1. Revised Eligibility Criterion

As with the original MST studies, perhaps there should have been evidence of serious or chronic offending to qualify for MST. Psychological testing revealed extremely high levels of mental health problems, but not all of the youth had prior convictions.

Entry to the project was determined with the Risk/Need Assessment score, a figure which, as it turned out, was not correlated with subsequent conviction in this sample. This instrument has not been validated on the general population and also may have trouble predicting criminal behaviour of a sample with a restricted range of scores.

In retrospect, length of prior record should have been the key inclusionary criterion for MST. Past criminal convictions were the best predictor of future criminal convictions in this sample. Almost one fifth of the sample (18%) had no prior convictions and have had no convictions since referral, a rate too high for a study of serious offenders.

The consequence of this decision, however, would have been low referrals and difficulties securing funding in at least one and perhaps two of the sites. It had also been the intention to test MST as an early intervention technique for younger youth who were deemed at risk of later criminal behaviour. Because of their age, it may be necessary to follow them longer than three years to see this benefit.

2. Broader Range of Outcome Measures

Perhaps the treatment effect of MST was manifested in domains not measured: school completion, employability, or need for residential treatment. Recidivism was the outcome measure because it was in the specific domains of criminal offending and incarceration that MST promised to make improvements. It was also the interest of stakeholders and funders. Resource constraints precluded the gathering of follow-up information about other factors as part of this study, but it is entirely possible to do so in the future.

3. Monitoring Implementation

Why were the results of American MST studies not replicated here? Even the most sophisticated experimental design does not answer that question and we are back in the realm of speculation.

Among the possible reasons for failing to find a treatment effect is the potential that the implementation was compromised. It had been assumed that the oversight of MST Services Inc. would ensure fidelity to the model. In essence, the evaluation was testing both the effectiveness of MST and the effectiveness of MST Services Inc. in disseminating their model.

MST Services Inc. recommends that consultation be a permanent feature of MST delivery but, after two years, the on-going cost threatened the continuation of the project. They worked with the Ministry of Community and Social Services to develop the capacity of the teams to conduct their own supervision for the last year.

Was fidelity compromised? We do not really know. The TAM instrument had been employed since shortly after project start-up. However, it did not show the expected pattern of increasing fidelity and scores were not correlated with conviction in the follow-up. A better measure of fidelity to the model was needed.

In retrospect, it would have been prudent to implement the service delivery component of the project about one year in advance of the initiation of the research component. Another benefit of this approach is to enable comparison of the referral profile under normal operating conditions to that after randomization began.

4. Component Analysis

MST comes as a package and its developers discourage any modifications to the method. It is touted to work equally well with members of all ethnic groups, both males and females, and with all presenting issues once the exclusionary criteria have been applied (e.g., acute psychosis).

Little data were collected on the specific types of interventions used (e.g., marital therapy, parent skills training, etc.). Rather, the intent was to evaluate the ability of the MST method to craft the most appropriate intervention for each case. In the same way, the package of usual services was tested. It was expected that youth would receive the most appropriate package of interventions available in their communities.

In retrospect, after failing to identify a treatment effect, it would have been helpful to examine which specific components of MST were associated with the best outcomes. Is it the in-home component, the small caseload, the focus on the family? Having this information would help define the most profitable direction for service delivery priorities.

5. Collect Consumer Feedback

Through anecdotal reports, the families appeared to derive great benefit from the intensive and in-home nature of MST. In retrospect, it would have been informative to survey the members of both groups to determine their opinions.

The Next Steps

The full interim report will be disseminated on the Internet and follow-up tracking of cases will continue. We are still at least one year away from having definitive answers about the long-term effectiveness of MST in reducing youthful offending in Ontario and readers are cautioned to interpret the results presented here in that context.

We will also seek funding to compare the two groups using non-criminal justice indicators such as admission to residential treatment, school completion, social assistance rates, and employment. It would also be possible to extend the follow-up of these youth to learn more about the criminal trajectories of youth. Finally, the development of a research agenda for youth justice would be timely.

Conclusions

Randomized field trials such as this study are becoming common and will be a key part of future research efforts to inform crime prevention efforts.[51] Experimental evaluations are not appropriate or necessary for all programs, depending upon their stage of development. Needs analysis, program audits, and process evaluations are all necessary in developing a program to the point where outcomes can be evaluated. Even then, there should be good evidence of a program's to affect desired changes before the expense and effort of an experimental evaluation is worthwhile.

This was the case for the MST project, where the key research question was this: Would MST be followed by lower levels of criminal conviction among serious young offenders than the services already available in southern Ontario? The developers of MST had verified its efficacy in two well-resourced randomized studies in the United States. Still at issue was its effectiveness in the field. It was also necessary to conduct a randomized study independent of the developers and to learn if results from the two MST studies would replicate to Canada.

To summarize the interim findings, the usual services group had better outcomes on some variables and the MST group had better measures on others, but in no case did the

differences reach the level of statistical significance. Measures of practical significance are not promising, in great measure because the control group had better outcomes on many indicators. The follow-up will continue until 2004 so the results could well change as more data are collected. However, at this point, it is not possible to recommend the adoption of MST in Canada.

Lessons learned included how easy it would be to make the wrong conclusion had a less rigorous methodology been used. This suggests that high-quality research can and should become the norm in the crime prevention field. Most program evaluations in corrections do not use control or comparison groups, have short or no-post discharge follow-up, do not use behaviour as an outcome measure, and may not use representative samples. The consequence has been that the potential of many programs to change behaviour may well have been overstated.

The lessons learned here should be helpful for the next step forward in finding programs to promote community safety.

Selected Resources

For more information about the MST in Ontario project, see:

A.W. Leschied and A. Cunningham (2002). *Seeking Effective Interventions for Young Offenders: Interim Results of a Four-Year Randomized Study of Multisystemic Therapy in Ontario, Canada*. London: Centre for Children and Families in the Justice System.

This and all other reports from the MST project are available for download at www.lfcc.on.ca

For more information on MST, see:

S.W. Henggeler (May, 1997). *Treating Serious Anti-Social Behavior in Youth: The MST Approach*. OJJDP Juvenile Justice Bulletin. [available at www.ncjrs.org]

S.W. Henggeler, S.F. Mihalic, L. Rone, C. Thomas and J. Timmons-Mitchell, J. (1998). *Blueprints for Violence Prevention, Book Six: Multisystemic Therapy*. Boulder, CO: Center for the Study and Prevention of Violence.

S.W. Henggeler, S. K. Schoenwald, C.M. Borduin, M.D. Rowland, and P.B. Cunningham (1998). *Multisystemic Treatment of Antisocial Behavior in Children and Adolescents*. New York: Guilford.

www.mstservices.com

Endnotes

- [1] C.A. Visher & D. Weisburd (1997). Identifying What Works: Recent Trends in Crime Prevention Strategies. *Crime, Law & Social Change*, 28(3-4): 223-242.
- [2] S.W. Henggeler, G.B. Melton, L.A. Smith, S.K. Schoenwald & J.H. Hanley (1993). Family Preservation Using Multisystemic Treatment: Long-term Follow-up to a Clinical Trial with Serious Juvenile Offenders. *Journal of Child & Family Studies*, 2: 283-293; and, S.W. Henggeler, G.B. Milton & L.A. Smith (1992). Family Preservation Using Multisystemic Therapy: An Effective Alternative to Incarcerating Serious Juvenile Offenders. *Journal of Consulting & Clinical Psychology*, 60: 953-961.
- [3] C.M. Borduin, B.J. Mann, L.T. Cone, S.W. Henggeler, B.R. Fucci, D.M. Blaske & R.A. Williams (1995). Multisystemic Treatment of Serious Juvenile Offenders: Long-term Prevention of Criminality and Violence. *Journal of Consulting & Clinical Psychology*, 63: 569-578.
- [4] S.W. Henggeler, S.G. Pickrel, M.J. Brondino & J.L. Crouch (1996). Eliminating (Almost) Treatment Dropout of Substance Abusing or Dependent Delinquents Through Home-Based Multisystemic Therapy. *American Journal of Psychiatry*, 153: 427-428; and, A.W. Henggeler, S.G. Pickrel & M.J. Brondino (1997). Multisystemic Treatment of Substance Abusing and Dependent Youth: Outcomes for Drug Use, Criminality and Out-of-home Placement, unpublished manuscript.
- [5] S.W. Henggeler, G.B. Melton, M.J. Brondino, D.G. Scherer & J.H. Hanley (1997). Multisystemic Therapy with Violent and Chronic Juvenile Offenders and Their Families: The Role of Treatment Fidelity in Successful Dissemination. *Journal of Consulting & Clinical Psychology*, 60: 953-961.
- [6] C.M. Borduin, S.W. Henggeler, D.M. Blaske and R. Stein (1990). Multisystemic Treatment of Adolescent Sexual Offenders. *International Journal of Offender Therapy & Comparative Criminology*, 34: 105-113.
- [7] S.W. Henggeler, S.F. Mihalic, L. Rone, C. Thomas & J. Timmons-Mitchell (1998). *Blueprints for Violence Prevention, Book Six: Multisystemic Therapy*. Boulder CO: Center for the Study and Prevention of Violence.
- [8] The Campbell Collaboration maintains a data base called C2-SPECTR, with references to over 10,000 articles on randomized and "possibly randomized" trials in the fields of social work and welfare, education, and criminal justice (see <http://campbell.gse.upenn.edu>). See A. Petrosino, R.F. Boruch, H. Soydan, L. Duggan & J. Sanchez-Meca (2001). Meeting the Challenges of Evidence-based Policy: The Campbell Collaboration. *Annals of the American Academy of Political & Social Sciences*, 578: 14-34. The Campbell Crime and Justice Coordinating Group is housed at the Australian Institute of Criminology. They oversee the preparation of systematic reviews relevant to prevention, treatment or control of crime (www.aic.gov.au/campbellcj/). See D.P. Farrington (2001). The Campbell Collaboration Crime and Justice Group. *Annals*
- [9] Access to CPIC records was obtained with consent of the families and with a court order. A CR search of CPIC system will show only convictions and the date of conviction. Accordingly, some of the offences registered in the follow-up period will have occurred during the intervention (or even prior to referral). This source of bias will impact both groups equally.
- [10] This document is available for download at www.lfcc.on.ca The present document (from April 2002) contains four more months of follow-up data than does the full report, tracking the youth
- [11] A good introduction is K.R. Murphy & B. Myers (1998). *Statistical Power Analysis: A Simple and General Model for Traditional and Modern Hypothesis Tests*. Mahwah NJ: Lawrence Erlbaum Associates.
- [12] J. Cohen (1992). A Power Primer. *Psychological Bulletin*, 112(1): 155-159.
- [13] D. Weisburd with Anthony Petrosino & Gail Mason (1993). Design Sensitivity in Criminal Justice Experiments: Reassessing the Relationship Between Sample Size and Statistical Power. In Michael Tonry & Norval Morris (eds.), *Crime & Justice: A Review of Research*, Vol. 17. Chicago: University of Chicago Press, at 337-379.
- [14] D. Weisburd & F.S. Taxman (2000). Developing a Multicenter Randomized Trial in Criminology: The Case of HIDTA. *Journal of Quantitative Criminology*, 16(3): 315-340.

- [15] In this approach, 100% of the youth begin at “time zero” with no convictions, in the top left corner of the graph. Should a youth be in custody at discharge, the tracking begins when that custody sentence is over. The youth who are convicted in the first month cause the line to drop. Those convicted during the next month are added to this group and cause another drop. In other words, the position of the line at six months represents the cumulative total of all youths who have been convicted at least once since time zero. A treatment effect for MST would be demonstrated by a line higher than that for the usual services group, meaning fewer convictions.
- [16] R.E. Kirk (2001). Promoting Good Statistical Practices: Some Suggestions. *Educational & Psychological Measurement*, 61(2): 213-218 at 213.
- [17] A good collection of this literature can be found in L.L. Harlow, S.A. Mulaik & J.H. Steiger, eds. (1997). *What if There Were no Significance Tests?* Mahwah NJ: Lawrence Erlbaum Associates Inc.
- [18] M.D. Maltz (1994). Deviating from the Mean: The Declining Significance of Significance. *Journal of Research in Crime & Delinquency*, 31(4): 434-463.
- [19] F.W. Dunford (2000). Determining Program Success: The Importance of Employing Experimental Designs. *Crime & Delinquency*, 46(3): 425-434.
- [20] J. Lattimer (2001). A Meta-analytic Examination of Youth Delinquency, Family Treatment and Recidivism. *Canadian Journal of Criminology*, 43(2): 237-253 at 238.
- [21] D. Layton MacKenzie, D.B. Wilson & S.B. Kider (2001). Effects of Correctional Boot Camps on Offending. *Annals of the American Academy of Political & Social Science*, 578: 126-143.
- [22] A. Petrosino, C. Turpin-Petrosino & J.O. Finckenauer (2000). Well-meaning Programs can have Harmful Effects!” Lessons for Experiments of Programs such as Scared Straight. *Crime & Delinquency*, 46(3): 354-379.
- [23] Cited in L.W. Sherman (2000). Reducing Incarceration Rates: The Promise of Experimental Criminology. *Crime & Delinquency*, 46(3): 299-314.
- [24] D.B. Wilson & M.W. Lipsey (2001). The Role of Method in Treatment Effectiveness: Evidence from Meta-analysis. *Psychological Methods*, 6(4): 413-429.
- [25] Reasons for a case being categorized as “drop out” included withdrawal of consent to MST by the family, youth going into custody, family and/or youth moving from the jurisdiction, and the youth’s whereabouts being unknown. In addition, one youth was taken into the care of the Children’s Aid Society and one youth was placed in a residential program because of his psychiatric needs. One case was closed because the level of violence in the family was inconsistent with a home-based intervention because of risk to the therapist. The drop-out rate varied among the sites from 26.5% to 14%.
- [26] For a good elementary description for how to calculate effect sizes, see Washington State Institute for Public Policy (2001). *The Comparative Costs and Benefits of Programs to Reduce Crime*. Olympia WA: WSIPP, Evergreen State College.
- [27] Including average number of days to first conviction, average number of days to first custody admission, average number of days of open custody sentences, rate of conviction at two years post, rate of conviction at three years post, rate of conviction overall (excluding offences against the administration of justice), rate of conviction (all offences), rate of adult convictions, and rate of sentencing to secure custody.
- [28] Including average number of days in secure custody, average number of offences of conviction, average number of prosecutions, average number of offences against the administration of justice, rate of recidivism at six months post, rate of sentencing to custody, average length of adult prison sentence, and rate of open custody sentences.
- [29] R. Boruch, B. Snyder & D. DeMoya (2000). The Importance of Randomized Field Trials. *Crime & Delinquency*, 46(2): 156-180.
- [30] J. Finckenauer (1982). *Scared Straight and the Panacea Phenomenon*. Englewood Cliffs, NJ: Prentice Hall.
- [31] C. Graebisch (2000). Legal Issues of Randomized Experiments on Sanctioning. *Crime & Delinquency*, 46(2): 271-282.

- [32] L. Feder, A. Jolin & W. Feyerherm (2000). Lessons From Two Randomized Experiments in Criminal Justice Settings. *Crime & Delinquency*, 46(3): 380-400.
- [33] D. Weisburd (2000). Randomized Experiments in Criminal Justice Policy: Prospects and Problems. *Crime & Delinquency*, 46(2): 181-193.
- [34] For example, G.L. Staines, K. McKendrick, T. Perlis, S. Sacks & G. DeLeon (1999). Sequential Assignment and Treatment-as-usual: Alternatives to Standard Experimental Designs in Field Studies of Treatment Efficacy. *Evaluation Review*, 23(1): 47-76; H.A. Liddle, G.A. Dakof, K. Parker, G.S. Diamond, K. Barrett & M. Tejada (2001). Multidimensional Family Therapy for Adolescent Drug Abuse: Results of a Randomized Clinical Trial. *American Journal of Drug & Alcohol Abuse*, 27(4): 651-688.
- [35] M.W. Lipsey & D.S. Corday (2000). Evaluation Methods of Social Intervention. *Annual Review of Psychology*, 51: 345-375.
- [36] L.W. Sherman et al. (1997). *Preventing Crime: What Works, What Doesn't, What's Promising*. Washington, DC: Office of Justice Programs, U.S. Department of Justice.
- [37] Murphy & Myers, supra note 11.
- [38] Weisburd & Taxman, supra note 14.
- [39] E.W. Gondolf (2001). Limitations of Experimental Evaluation of Batterer Programs. *Trauma, Violence & Abuse*, 2(1): 79-88.
- [40] J.F. Short Jr., M.A. Zahn & David P. Farrington (2000). Experimental Research in Criminal Justice Settings: Is There a Role for Scholarly Societies? *Crime & Delinquency*, 46(3): 295-298.
- [41] L. Feder, A. Jolin & W. Feyerherm (2000). Lessons from Two Randomized Experiments in Criminal Justice Settings. *Crime and Delinquency*, 46(3): 380-400.
- [42] These and other tasks undertaken in the developmental and pilot phases of the project are detailed in London Family Court Clinic (1998). *Clinical Trials of Multisystemic Therapy with High-Risk Phase I Young Offenders: Year-end Report 1997/98*. London: London Family Court Clinic.
- [43] TAMs were completed by the principal caregiver at several points during treatment and again at case closure. The caregiver was to make reference to the previous two or three sessions when making their ratings. This instrument yields six sub-scale scores: overall adherence, non-productive settings, therapist/family problem-solving effort, therapist attempts to change interaction, lack of direction, and family-therapist consensus. Three of the sub-scales have been validated by the authors of the instrument.
- [44] Schoenwald, S.K., S.W. Henggeler, M.J. Brondino & M.D. Rowland (2000). Multisystemic Therapy: Monitoring Treatment Fidelity. *Family Process*, 39(1): 83-103.
- [45] This discussion is drawn from J. Cohen (1988). *Statistical Power Analysis for the Behavioral Sciences, 2nd Edition*. Lawrence Erlbaum & Associates.
- [46] M.W. Lipsey & D.B. Wilson (1993). The Efficacy of Psychological, Educational, and Behavioral Treatment. *American Psychologist*, 48: 1181-1209.
- [47] X.T. Fan (2001). Statistical Significance and Effect Size in Education Research: Two Sides of a Coin. *Journal of Educational Research*, 94 (5): 275-282.
- [48] $NNT = 1 / (\text{proportion of failures in control group} - \text{proportion of failures in experimental group})$.
- [49] H.T Reynolds (1977). *Analysis of Nominal Data*. Beverley Hills: Sage Publications; R.J. Cook & D.L. Sackett (1995). The Number Needed to Treat: A Clinically Useful Measure of Treatment Effectiveness. *British Medical Journal*, 310: 452-454.
- [50] For a community-based intervention, MST is costly, mostly because of the low caseload carried by workers and the high cost of on-going training and consultation. Other costs include mileage, cellular telephones, parking and other expenses associated with a program delivered outside the office. Costs associated with the supervision include weekly long-distance telephone charges for case consultations with MST Services Inc. in South Carolina and travel of team members to quarterly booster training sessions.
- [51] D.P. Farrington (1999). A Criminological Research Agenda for the Next Millennium. *International Journal of Offender Therapy & Comparative Criminology*, 43(2): 154-167.

Interested in other publications from the Centre? Want a copy of the full MST report?

- Please send a complete publication list
- Please send an annual report
- Please send ordering information for the full 2002 MST report, *Seeking Effective Interventions for Serious Young Offenders*

Name: _____

Organization: _____

Address: _____

City: _____ Province or State: _____

Postal/Zip Code: _____ Country: _____

e-mail address: _____



Centre for Children and Families in the Justice System

London Family Court Clinic

200-254 Pall Mall St.

London ON

CANADA N6A 5P6

e-mail: publications@lfcc.on.ca

*The Centre is a non-profit organization. Donations are welcome.
Revenue Canada Charitable Registration 12991 5153 RR001*

PRAXIS: Research from the Centre for Children
& Families in the Justice System
One Step Forward:
Lessons Learned from a Randomized Study of
Multisystemic Therapy in Canada
Alison Cunningham, Director of Research and Planning