2008

Evidence-based or Biased? The Quality of Published Reviews of Evidence-based Practices

Julia H. Littell
Bryn Mawr College, jlittell@brynmawr.edu

Evidence-based or biased?

The quality of published reviews of evidence-based practices

to appear in *Children and Youth Services Review*

Final version submitted 20 April 2007

DO NOT CITE WITHOUT AUTHOR'S PERMISSION

Julia H. Littell

Graduate School of Social Work and Social Research

Bryn Mawr College

jlittell@brynmawr.edu

Running head: Reviews of evidence-based practices

Key words: evidence-based practice, research synthesis, systematic reviews, confirmation bias

Author's note: Portions of this work were supported by grants from the Smith Richardson Foundation, the Swedish Centre for Evidence-based Social Work Practice (IMS), and the Nordic Campbell Center. I thank Burnee' Forsythe for her assistance with the search process and document retrieval. I thank Jim Baumohl, Dennis Gorman, Mark Lipsey, David Weisburd, David B. Wilson, and anonymous reviewers for helpful comments on an earlier version of this paper.

Abstract

Objective: To assess methods used to identify, analyze, and synthesize results of empirical research on intervention effects, and determine whether published reviews are vulnerable to various sources and types of bias.

Methods: Study 1 examined the methods, sources, and conclusions of 37 published reviews of research on effects of a model program. Study 2 compared findings of one published trial with summaries of results of that trial that appeared in published reviews.

Results: Study 1: Published reviews varied in terms of the transparency of inclusion criteria, strategies for locating relevant published and unpublished data, standards used to evaluate evidence, and methods used to synthesize results across studies. Most reviews relied solely on narrative analysis of a convenience sample of published studies. None of the reviews used systematic methods to identify, analyze, and synthesize results. Study 2: When results of a single study were traced from the original report to summaries in published reviews, three patterns emerged: a complex set of results was simplified, non-significant results were ignored, and positive results were over-emphasized. Most reviews used a single positive statement to characterize results of a study that were decidedly mixed. This suggests that reviews were influenced by confirmation bias, the tendency to emphasize evidence that supports a hypothesis and ignore evidence to the contrary.

Conclusions: Published reviews may be vulnerable to biases that scientific methods of research synthesis were designed to address. This raises important questions about the validity of traditional sources of knowledge about "what works," and suggests need for a renewed commitment to using scientific methods to produce valid evidence for practice.

The emphasis on evidence-based practice appears to have renewed interest in "what works" and "what works best for whom" in response to specific conditions, disorders, and psychosocial problems. Policy makers, practitioners, and consumers want to know about the likely benefits, potential harmful effects, and evidentiary status of various interventions (Davies, 2004; Gibbs, 2003). To address these issues, many reviewers have synthesized results of research on the impacts of psychosocial interventions. These reviews appear in numerous books and scholarly journals; concise summaries and lists of "what works" can be found on many government and professional organizations' websites.

In the last decade there were rapid developments in the science of research synthesis, following publication of a seminal handbook on this topic (Cooper & Hedges, 1994). Yet, the practice of research synthesis (as represented by the proliferation of published reviews and lists of evidence-based practices) and the science of research synthesis have not been well-connected (Littell, 2005).

In this article, I trace the development and dissemination of information about the efficacy and effectiveness of one of the most prominent evidence-based practices for youth and families. I examine the extent to which claims about the efficacy of this program are based on scientific methods of research synthesis, and whether they are vulnerable to several sources and types of bias.

Research Synthesis

The synthesis of results of multiple studies is important because single studies, no matter how rigorous, have limited utility and generalizability. Partial replications may refute, modify, support, or extend previous results. Compared to any single study, a careful synthesis of results of multiple studies can produce better estimates of program impacts and assessments of conditions under which treatment impacts may vary (Shadish, Cook, & Campbell, 2002).

Research synthesis has a long history. Most readers are familiar with traditional literature reviews, which rely on narrative summaries of results of multiple studies. Systematic reviews and meta-analyses are becoming more common, but the traditional model prevails in the social sciences despite a growing body of evidence on the inadequacy of narrative reviews.

Sources and Types of Bias in Research Reviews

There are several potential sources and types of bias in research synthesis. These can be divided into three categories: biases that arise in the original studies, in the dissemination of study results, and in the review process itself.

Treatment outcome studies can systematically overestimate or underestimate effects due to design and implementation problems that render the studies vulnerable to threats to internal validity (e.g., selection bias, statistical regression, differential attrition), statistical conclusion validity (e.g., inadequate statistical power, multiple tests and "fishing" for significance), and construct validity (e.g., experimenter expectancies, inadequate implementation of treatment, treatment diffusion; Shadish et al., 2002). "Allegiance effects" may appear when interventions are studied by their advocates (Luborsky, Diguer, Seligman, Rosenthal, Krause, Johnson, et al., 1999); these effects may be due to experimenter expectancies or to high-fidelity implementation (Petrosino & Soydan, 2005).

Confirmation bias (the tendency to emphasize evidence that supports a hypothesis and ignore evidence to the contrary) can arise in the reporting, publication, and dissemination of results of original studies. Investigators may not report outcomes or may report outcomes selectively (Dickersin, 2005). Studies with statistically significant, positive results are more likely to be submitted for publication and more likely to be published than studies with null or negative results (Begg, 1994; Dickersin, 2005). Mahoney (1977) found that peer reviewers were biased against manuscripts that reported results that ran counter to their expectations or

theoretical perspectives. Other sources of bias in dissemination are related to the language, availability, familiarity, and cost of research reports (Rothstein, Sutton & Bornstein, 2005). Selective citation of reports with positive findings may make those results more visible and available than others (Dickersin, 2005).

These biases are likely to affect research synthesis unless reviewers take precautions to avoid them. The review process is most vulnerable to bias when reviewers sample studies selectively (e.g., only including published studies), fail to consider variations in study qualities that may affect the validity of inferences drawn from them, and report results selectively.

The synthesis of multiple results from multiple studies is a complex task that is not easily performed with "cognitive algebra." Since the conclusions of narrative reviews can be influenced by trivial properties of research reports (e.g., Bushman & Wells, 2001), several quantitative approaches to research synthesis have been developed and tested. Perhaps the most common of these is "vote counting" (tallying the number of studies that provide evidence for and against a hypothesis), which relies on tests of significance or directions of effects in the original studies. Carlton and Strawderman (1996) showed that vote counting can lead to the wrong conclusions. Meta-analysis can provide better overall estimates of treatment effects, but these techniques have limitations as well.

Systematic Reviews

Systematic reviews are designed to minimize bias at each step in the review process. Systematic approaches to reviewing research are not new, nor did they originate in the biomedical sciences (Chalmers, Hedges, & Cooper, 2002; Petticrew & Roberts, 2006). However, systematic reviews have received more attention in recent years as advances in the science of research synthesis showed that review methods matter (Cooper & Hedges, 1994), as centers for research synthesis evolved in the U.K. and elsewhere, and as the general public became more

aware of potential pitfalls of haphazard reviews (following, for example, the alleged suppression of negative findings on effects of Vioxx in the U.S. in 2005).

Two international, interdisciplinary collaborations of scholars, practitioners, and policy makers have established guidelines and principles for minimizing bias in the synthesis of research on treatment effects. The Cochrane Collaboration synthesizes results of studies on effects of interventions in health care (see www.cochrane.org) and The Campbell Collaboration synthesizes results of interventions in the fields of social care (education, social welfare, mental health, and crime and justice; www.campbellcollaboration.org). Building on advances in the science of research synthesis (e.g., Cooper & Hedges, 1994; Lipsey & Wilson, 2001; Moher, Cook, Eastwood, Olkin, Rennie, Stroup, et al., 1999), these groups have produced useful background papers and evidence-based guidelines for reviewers (e.g., Becker, Hedges, & Pigott, 2004; Higgins & Green, 2006; Rothstein, Turner, & Lavenberg, 2004; Shadish & Myers, 2004) along with studies of methodological qualities of systematic reviews.

A systematic review follows the basic steps in the research process (Cooper & Hedges, 1994). Systematic reviews are observational studies, in which prior studies are treated as sampling units and units of analysis. The basic steps and principles in conducting a systematic review are as follows.

Transparent intentions and methods. A detailed plan for the review is developed in advance, specifying central objectives and methods. Steps and decisions are carefully documented so that readers can follow and evaluate reviewers' methods (Moher et al., 1999). Conflicts of interest and sponsorship arrangements are disclosed (Higgins & Green, 2006) because these issues can affect reviewers' conclusions (e.g., Jørgensen, Hilden, & Gøtzsche, 2006).

Explicit inclusion/exclusion criteria. Systematic reviews have clear boundaries so they

can be replicated or extended by others. Reviewers specify the study designs, populations, interventions, comparisons, and outcome measures that will be included and excluded. Reasons for exclusion are documented for each excluded study. This limits reviewers' freedom to select studies on the basis of their results, or on some other basis.

Search strategies. Reviewers use a systematic approach and a variety of sources to try to locate *all* potentially-relevant studies. This involves collaboration with information retrieval specialists to generate keyword strings used to search relevant electronic databases. It also involves attempts to locate the "grey literature" (unpublished and hard-to-find studies) to minimize publication bias and the "file drawer problem" (Begg, 1994; Hopewell, McDonald Clarke, & Egger, 2006; Petticrew & Roberts, 2006; Rosenthal, 1994; Rothstein et al., 2004; Rothstein et al., 2005). This can be accomplished through contacts with a snowball sample of key informants (experts on the topic) until data saturation is achieved. Hand searching of the contents of relevant journals is often needed to find eligible studies that are not properly indexed (Hopewell, Clarke, Lefebvre, & Scherer, 2006). The search process and its results are carefully documented.

Inter-rater agreement on all key decisions. Decisions on full-text retrieval, study inclusion/exclusion, and study coding should be made by two independent raters, who compare notes, resolve differences, and document reasons for their decisions (Higgins & Green, 2006).

Systematic extraction of data from original studies. Raters extract data from study reports onto paper or electronic coding forms. These data are then available for use in the analysis and synthesis of results. The data forms provide a bridge between the primary research studies and the research synthesis, and a historical record of reviewers' decisions (Higgins & Green, 2006).

Analysis of study qualities and results. Aspects of methodology that relate to the validity of a study's conclusions are assessed individually, rather than being summed into total study-

quality scores (Shadish & Myers, 2004). Campbell's threats-to-validity approach is a useful framework in this regard (Wortman, 1994). Some reviews focus on assessment of potential sources and types of bias (Higgins & Green, 2006).

Analysis of study results. Study findings are represented as effect sizes (ES) whenever possible. Raters document the data and formulas used for effect size calculations (Becker et al., 2004).

Synthesis of results. Since conclusions of narrative reviews can be influenced by trivial properties of research reports (Bushman & Wells, 2001) and *a priori* assumptions, methods used to combine results across studies should be transparent (Becker et al., 2004; Gambrill, 2006). Quantitative methods (meta-analysis) lend themselves to this purpose. Meta-analysis is used to produce pooled estimates of ES that account for variations in the precision of estimates drawn from different samples (due to variations in sample size and within-sample variance), explore potential moderators of effect size, and examine potential effects of publication bias (Cooper & Hedges, 1994; Lipsey & Wilson, 2001; Rothstein et al., 2005). It is important to note that meta-analyses are not necessarily systematic reviews (e.g., a meta-analysis of a convenience sample of published studies is not a systematic review), and systematic reviews do not always include meta-analysis.

Reporting of results. Moher and colleagues (1999) developed the Quality of Reporting of Meta-analyses (QUOROM) statement to improve reports on systematic reviews and meta-analyses. The statement includes a checklist of items that should be reported and a flow diagram for authors to use to describe how studies were identified, screened, and selected for the review.

Updating reviews. Systematic reviews must be updated regularly so that they remain current and relevant for policy and practice.

Current Practice

What criteria and methods do reviewers actually use to find, assess, and compile evidence of intervention effects? How "systematic" are their reviews? That is, to what extent do they use explicit inclusion and exclusion criteria, well-defined search and retrieval procedures, attempts to avoid publication bias (the "file drawer" problem), clear standards of evidence, and quantitative (or at least transparent) methods of research synthesis?

These issues have been the topic of considerable interest and analysis in health care. Several studies compared Cochrane reviews to other "systematic" reviews and meta-analyses. For example, Jadad and colleagues (Jadad, Moher, Browman, Booker, Sigouin, Fuentes, et al., 2000) analyzed 50 systematic reviews and meta-analyses of asthma treatment and found that most reviews published in peer-reviewed journals had serious methodological flaws that limited their usefulness; Cochrane reviews were more rigorous and better reported than those published in peer-reviewed journals. All industry-funded reviews were judged to have serious flaws. Shea, Moher, Graham, Pham, and Tugwell (2002) found the overall quality of "systematic reviews" was low, but noted that the development of evidence-based criteria for reporting systematic reviews (the QUOROM statement) may help improve their quality.

As mentioned above, there is an extensive body of work on the biases and limitations of traditional, narrative reviews of empirical research (e.g., Bushman & Wells, 2001; Carlton & Strawderman, 1996; Cooper & Hedges, 1994). Yet narrative reviews are typical in the social sciences. Systematic review methods have been discussed in the social sciences for decades, and systematic reviews have appeared more often in recent years (Petticrew & Roberts, 2006). However, to my knowledge, there have been no attempts to evaluate the quality of published reviews of research in the fields of social care.

A Case Study: Multisystemic Therapy

Following standards and procedures established by the Cochrane Collaboration, the

Campbell Collaboration, and the QUOROM statement, my team conducted a systematic review and meta-analysis of research on effects of a prominent, model program called Multisystemic Therapy (Littell, Popa, & Forsythe, 2005). Multisystemic Therapy (MST) was selected as the topic for that review (and the primary example for this paper) because it has an unusually strong research base, including several randomized controlled trials. MST has been cited as an effective, evidence-based treatment model by the U.S. Department of Justice Office of Juvenile Justice and Delinquency Prevention (OJJDP), the Center for Substance Abuse Prevention (2000), the National Institute on Drug Abuse (1999, 2003), National Institute of Mental Health (2001, 2003), and the Surgeon General's office (U.S. Department of Health and Human Services, 1999, 2001). MST is one of the Model Programs identified by the Substance Abuse and Mental Health Services Administration (SAMSHA, 2004) and by the OJJDP-funded Blueprints for Violence prevention (Henggeler, Mihalic, Rone, Thomas, & Timmons-Mitchell, 1998).

MST is a short-term, home- and community-based intervention for families of youth with social, emotional, or behavioral problems. MST uses a "family preservation service delivery model" to address complex psychosocial problems and provide alternatives to out-of-home placement of children and youth. Treatment teams consist of professional therapists (mental health professionals with masters or doctoral degrees) and crisis caseworkers, who are supervised by clinical psychologists or psychiatrists. Therapists have small caseloads and are available to program participants 24 hours a day, 7 days a week. Treatment is individualized to address specific needs of youth and families, and includes work with other social systems including schools and peer groups. "Intervention strategies are integrated from other pragmatic, problem-focused treatment models" (Henggeler & Borduin, 1995. p. 121) and MST follows 9 general principles (see Henggeler, Schoenwald, Borduin, Rowland, & Cunningham, 1998; Henggeler, Schoenwald, Rowland, & Cunningham, 2002).

There are approximately 120 licensed MST programs in more than 30 states in the USA. At last count, there were 18 licensed MST programs in Norway, 7 in Sweden, 5 in Canada, 3 in the Netherlands, 2 in Australia, 2 in England, and single programs in Denmark, Ireland, and New Zealand (MST Services Inc., 2006). In total, there are more than 250 licensed MST teams in North America and Europe, treating 10,000 serious juvenile offenders each year (Henggeler, 2003). Considerable attention has been paid to the dissemination of MST and the fidelity of MST replications (e.g., Henggeler, Schoenwald, Liao, Letourneau, & Edwards, 2002; Schoenwald, Henggeler, Brondino & Rowland, 2000; Schoenwald & Hoagwood, 2001).

The results of a systematic review of research on MST (Littell et al., 2005) were not consistent with the published works of current authorities on the topic (see Henggeler, Schoenwald, Borduin, & Swenson, 2006; Littell, 2005, 2006). While most (but not all) of the primary studies showed that MST had statistically significant effects on at least one outcome measure, these effects were inconsistent across studies; that is, different studies showed effects on different outcome measures. In meta-analysis, there were no significant overall effects (across studies) on any single outcome measure (Littell et al., 2005). It is possible that prior reviews focused on positive effects, not the entire pattern of positive, negative, and null results. This article examines methods used in prior reviews to identify, assess, and synthesize this body of evidence, to determine whether these reviews were vulnerable to confirmation bias.

Study 1: Methods of Published Reviews

This study sought to determine how prior, published reviews of research on effects of MST were conducted. I expected to find few fully-systematic reviews of research in this area, but thought published reviews might become more systematic over time (i.e., reviews published in later years might contain more of the elements of a systematic review mentioned above).

Methods

To be included in this analysis, reviews had to be published after 1996, when at least 10 reports on MST outcome studies were in print. Included reviews had to cite at least 2 original studies of effects of MST (i.e., published or unpublished reports on non-overlapping samples), and provide a summary of results across MST studies.

Most reviews were identified in the Spring of 2003. Using keyword searches of electronic databases (including PsychINFO, MEDLINE, Dissertation Abstracts International, ERIC, CINAHL) and government websites (U.S DHHS, CDC, GPO, NIH, and the UK Home Office) and contacts with experts, my colleagues and I identified 86 potentially-relevant, published reviews of MST outcome research. We retrieved available reviews in order to scan their reference lists to find relevant outcome studies. We read abstracts of all 86 reviews and retrieved full-text reports on 66 (77%). This purposive sample includes reviews published in scholarly books and articles, reviews that are cited more often than those we did not attempt to retrieve.

Of the 66 reviews examined, 7 were published before 1997, 19 relied solely or primarily on other reviews, 1 meta-analysis did not cite its sources, and 3 reviews provided no analysis or synthesis of results for MST per se. The remaining 37 reviews met the inclusion criteria for this study, and are described below.

Results

MST outcome studies have been reviewed in relation to a variety of youth and family problems and policy issues. As shown in Table 1, reviews have focused on effects of MST (and other interventions) on crime, delinquency, antisocial behavior, and/or conduct disorder (14 reviews); substance abuse (3 reviews); other mental health problems among children (9 reviews); and child maltreatment (2 reviews). Several reviews assessed the effects of MST across populations and problems (7 reviews) or effects of a broader array of family-based services (2 reviews).

Purpose and Hypotheses Regarding MST

The purposes of the reviews (as stated in the abstract or introduction) were to

- describe MST (Burns, 2003; Burns, Schoenwald, Burchard, Faw, & Santos, 2000; Schoenwald, Brown, & Henggeler, 2000; Swenson & Henggeler, 2003);
- provide practitioners with information on evidence-based practices (Corcoran, 2003; Henggeler, Mihalic et al., 1998; Schoenwald & Rowland, 2002);
- "discuss the emergent success" of MST (Borduin, 1999), provide an "empirical rationale" for MST (Borduin, Schaeffer, & Ronis, 2003), "present empirical support" for MST (Swenson, Henggeler, Schoenwald, Kaufman, & Randall, 1998);
- review treatment models that are promising (Borduin, Heiblum, Jones, & Grabe, 2000; Kazdin, 1998; Kazdin & Weisz, 1998), empirically-supported (Brestan & Eyeberg, 1998), effective (Chorpita, Yim, Donkervoet, Arensdorf, Amundsen, McGee, et al., 2002; Burns, Hoagwood, & Mrazek, 1999; Hoagwood, Burns, Kiser, Ringeisen, & Schoenwald, 2001; Letourneau, Cunningham, & Henggeler, 2002), or efficacious (Henggeler & Sheidow, 2003);
- review research on treatment effects (Corcoran, 2000; Farrington & Welsh, 2003; Fraser, Nelson, & Rivard, 1997; Henggeler, Schoenwald, Rowland et al., 2002; Miller, Johnson, Sandberg, Stringer-Seibold, & Gfeller-Strouts, 2000; Smith & Stern, 1997; Sudderth, 2000; Tarolla et al., 2002; U.S. DHHS, 1999)
- review well designed studies of treatment effects (Brosnan & Carr, 2000; Cormack & Carr, 2000; Vaughn & Howard, 2004);
- "examine effectiveness" (Curtis, Ronan, & Borduin, 2004) or "determine effects" (Woolfenden, Williams, & Peat, 2003); or
- "find programs that save more money than they cost" (Aos, Phipps, Barnoski, & Lieb, 2001).

Some reviewers acknowledged their debt to previous reviews. For example, "Given the

conclusions of previous authoritative reviewers of the field, this chapter is confined to a consideration of well-designed studies which evaluate the effectiveness" of selected interventions (Brosnan & Carr, 2000, p. 134; see also, Henggeler & Sheidow, 2003). Other reviews began with hypotheses that MST is "promising" (Fonagy & Kurtz, 2002), "supported" (Pushak, 2002), "well-validated" (Schoenwald & Rowland, 2002), or has "favorable outcomes" (Randall & Cunningham, 2003).

Thus, reviews varied in the clarity of the stated purpose, whether the purpose was stated in confirmatory terms (e.g., to find or show evidence of effects), and whether a priori assumptions about the state of the evidence were expressed. When assumptions or hypotheses about the direction and strength of effects were stated, reviewers usually cited previous reviews.

Review Methods

Most (22 or 60%) of the 37 reviews relied solely on narrative synthesis of convenience samples of studies. One review provided a narrative synthesis based on a systematic search for published studies (Brestan & Eyberg, 1998). Five reviews described studies and their results in tables and text. Three reviews provided study-level effect sizes, 5 included quantitative synthesis (meta-analysis), and one included both meta-analysis and cost-benefit analysis.

More detailed information on review methods is shown in Table 2. In this table reviews were organized by publication year, to see whether there were any discernible changes in review methods over time. Contrary to expectations, there were no apparent increases in the use of explicit inclusion criteria, systematic searches, unpublished reports, study assessment methods, or quantitative analysis over time.

Authors' Independence.

Twenty-two (60%) of the 37 reviews in our analysis were authored by people who were not affiliated with MST program developers or MST Services Inc. Hereafter, these are referred to

as "independent" reviews.

Inclusion Criteria, Search Strategies, and Their Results.

Eight (22%) of the 37 reviews used explicit inclusion and/or exclusion criteria. With one exception (Curtis et al., 2004), the reviews that used explicit criteria were authored by independent investigators. Nine reviews (24%) used systematic keyword searches of electronic databases; 8 of these reviews also had explicit inclusion criteria. Nine reviews included references to unpublished MST research reports; only one (Farrington & Welsh, 2003) also had explicit inclusion criteria and/or a systematic search strategy.

As shown in Table 2, the number of MST research reports included in reviews ranged from 1 to 29, representing up to 25 separate studies (non-overlapping samples). Reviews that had more specific foci (e.g., substance abuse outcomes; cf. Cormack & Carr, 2000; Sudderth, 2000; Vaughn & Howard, 2004) tended to cite fewer studies than those that focused on MST research across problems and populations (e.g., Henggeler, Schoenwald, Rowland, et al., 2002).

Independent reviews were somewhat more likely than reviews co-authored by MST developers to use explicit inclusion criteria (31.8% vs. 6.7%) or systematic search strategies (36.4% vs 6.7%), but less likely to include unpublished reports (13.6% vs. 40.0%). Independent reviews tended to include fewer research reports (means 4.9 vs. 13.5) and fewer studies (4 vs. 9.5) than those co-authored by MST developers.

Standards of Evidence

Study design or allocation method. Most reviews distinguished randomized and nonrandomized studies, but variations in study quality within these two design categories were rarely considered, and results of randomized and nonrandomized studies were usually given equal weight in the analysis (a notable exception is the review by Aos et al., 2001, discussed below). Seven reviews (including Aos et al., 2001) limited their included studies to randomized

controlled trials (RCTs) or assessed the method used to allocate participants to treatment groups.

Two reviews altered their initial methodological inclusion criteria. After preliminary analysis showed that only 7.4% of studies met one of their initial criteria (randomization), Fonagy and colleagues (2002) relaxed this criterion. Similarly, Carr and colleagues intended to limit their reviews to RCTs, but if this criterion "yielded a particularly small pool of studies, the criteria were relaxed and less methodologically robust studies were included" (Carr, 2000, p. 6).

Attrition. Only 4 reviews assessed attrition in primary outcome studies (Brosnan & Carr, 2000; Cormack & Carr, 2000; Farrington & Welsh, 2003; Woolfenden et al., 2003), but all 4 underestimated attrition (Littell, 2005, 2006). This is likely due to the practice (articulated by Woolfenden and colleagues) of selecting the most recent report when there were multiple reports per study, instead of tracking attrition over time through different reports. The review by Aos and colleagues intended to address attrition and to limit meta-analyses to studies that provided an intent-to-treat analysis, but did not do so. A full account of attrition was not always provided in published reports of primary studies (Littell, 2005, 2006).

Study quality ratings. Seven reviews rated study quality (all were independent). Brosnan and Carr (2000) used a 25-point scale to rate methodological features of included studies; scores (for MST trials and other studies) ranged from 10 to 18 on this scale. Cormack and Carr (2000) used a similar, 24-item rating scale; scores ranged from 11 to 17 (MST trials were rated 11 and 12). Vaughn and Howard (2004) adapted the Methodological Quality Rating Scale (Miller, Brown, Simpson, Handmaker, Bien, Luckie, et al., 1995) for use in their review. Scores on this scale could range from 0 to 16; the actual range was 8 to 15 (MST trials were rated 10 and 13).

Brestan and Eyberg (1998) recorded information on 4 "minimal criteria of good designs": use of a comparison group, random assignment, use of reliable measures, and report of descriptive statistics. They also recorded information on other methodological criteria. Their

study quality ratings were not reported, however.

The Scientific Methods Scale (SMS) was used by Farrington and Welsh (2003) to define high quality evaluation designs. Farrington and Welsh only included studies that were randomized experiments (level 5) or quasi-experiments with matched control groups (level 4).

Aos and colleagues (2001) used a 5-point rating scale similar to the SMS in their analysis and weighted study-level effect sizes (ES) by study quality. The findings of randomized experiments (level 5) were not discounted (weighting factor = 1.0), findings of quasi-experiments with controls for selection bias (level 4) were weighted .75, findings of quasi-experiments with matched comparison groups (level 3) were weighted .5, and other quasi-experiments (level 2) and single-group designs (level 1) were not included (weighted 0).

Synthesis of Results

Selected outcomes. Several reviews summarized the evidence in tables of "key findings" or selected outcomes. In some reviews, this evidence was organized by outcome domains, and tables showed which studies provided evidence that MST had *favorable* effects on outcomes (e.g., Corcoran, 2003, p. 182). Other reviews organize the evidence by study, highlighting *positive* results from each study (e.g., Burns et al., 2000, pp. 291-292; Henggeler, Schoenwald, Rowland et al., 2002, pp. 207-208). Notably, null results and negative effects were not mentioned in these summaries. A similar approach was used in some narrative syntheses (e.g., Henggeler & Sheidow, 2003; Letourneau et al., 2002). The practice of highlighting positive or favorable outcomes is an example of confirmation bias.

More complete summaries of evidence are provided by Fraser et al. (1997), who use tables to show study-level effect sizes; and by Henggeler et al. (1998, p. 37), who provide a table indicating that 3 of 4 MST trials had null results on at least one outcome measure; all 4 trials also had positive results on at least one outcome measure.

Vote Counting. Several reviews reported the number of studies that showed statistically significant differences in favor of the MST group on one or more outcome measures (e.g., Burns et al., 1999; Miller et al. 2000; US DHHS, 1999). This "vote counting" method does not take sample size or precision into account, and can lead reviewers to miss important effects in underpowered studies and count trivial differences in large studies (Bushman, 1994).

The "Chambless criteria." Several reviews classified MST as a "probably efficacious" treatment (Brestan & Eyberg, 1998; Burns et al., 1999, 2000; Burns, 2003; Chorpita et al., 2002), according to the criteria for empirically supported treatments (ESTs) developed by an American Psychological Association (APA) Division 12 (Clinical Psychology) task force (Chambless, Baker, Baucom, Beutler, Calhoun, Crits-Christoph, et al., 1998).

Quantitative synthesis (meta-analysis). The first meta-analysis of results from MST trials appeared in 2000. Six reviews provided pooled ES estimates across 3 to 6 MST trials. These estimates are not strictly comparable, since they are based on different pooling methods and some are more rigorous than others (Cooper & Hedges, 2004; Lipsey & Wilson, 2001).

Several authors used inverse variance methods to account for differences in the precision of estimates (due to differences in sample size and variability). I converted the direction of effects as needed, so that positive ES always favor MST. Results, reported as weighted standardized mean differences, for MST trials were:

- .31 for recidivism (3 trials, Aos et al., 2001);
- .41 for delinquency (6 trials, including antisocial outcomes for studies that did not provide measures of delinquency; Farrington & Welsh, 2003); and
- .11 for family adaptability, .18 family cohesion, .02 peer adaptability, .02 peer bonding, .15 peer aggression, .03 peer maturity, .50 risk of incarceration, .05 parental mental health, and .50 child behavior (3 trials, Woolfenden et al., 2003).

The pooled estimates reported by Aos and colleagues and by Farrington and Welsh were statistically significant (although Farrington & Welsh noted that statistically significant differences were observed in only 2 of the 6 studies in their analysis). Pooled results for MST trials in the Woolfenden review were not significantly different from zero.

Three other reviews reported pooled ES, but these were not weighted and the pooling methods were not clear. These results include:

- Mean effect sizes of .8 for parent-reported improvement in conduct problems, 1.2 for self-reported improvement in these problems, .7 for improvements in family functioning, and 1.2 for recidivism rates between 2 and 4 years post-treatment (Brosnan & Carr, 2000);
- a mean effect of .5 on the Revised Behavior Problem Checklist (Chorpita et al., 2002); and
- an average effect across all outcomes of .55 (Curtis et al., 2004).

Recall that a fully-systematic review showed that MST did not produce significantly better or worse results than other treatments on any (of 21) outcome measures (Littell et al., 2005).

Reviewers' Conclusions

Although there is considerable variation in the methods used and studies included in these reviews, there is somewhat more consistency in their conclusions. As shown in Table 1, only 3 of 37 reviews mentioned negative or null effects in their conclusions (Farrington & Welsh, 2003; Swenson & Henggeler, 2003; Woolfenden et al., 2003). Nine reviews provided some caveat about the evidence (e.g., results were not classified as "well-established," results appear to depend on fidelity, and findings have not yet been replicated by other research teams). However, most (25) of the reviews seemed to provide unqualified support for MST. These conclusions were not related to whether authors were independent (e.g., negative or null findings were mentioned by 9% of independent reviews and 7% of reviews authored by MST developers). Hence, there was no evidence of allegiance bias in the reviews.

Since all reviews included studies that had mixed results, it is not clear whether or how these results were factored into the reviewers' conclusions. How do reviewers determine whether positive results outweigh negative or null findings, especially when they do not use quantitative methods to pool results across studies? The next study takes a closer look at these issues.

Study 2: From Results to Reports to Reviews

This study examines findings from a single, published MST outcome study, and compares them to published reviews of this study.

Methods

I selected a study that is (to my knowledge) the only completed trial of effects of MST in cases of child maltreatment. Reported by Brunk, Henggeler, and Whelan (1987), this trial included 43 families of abused or neglected children who were randomly assigned to MST or parent training groups (PT).

I categorized the direction of results on the scales and subscales used in the Brunk study, using three categories: favors MST, favors PT, and neutral (no difference between groups, unclear, or missing). I then tallied the number of items in each category. Although such "vote counting" is not ideal, I will show that it is not possible to calculate accurate effect sizes from published results of this study.

Content analysis was used to identify the number and direction (favors MST, favors PT, and neutral) of discrete phrases used by the study's authors to characterize results of the study in the original abstract. The same method was used to analyze summaries of the Brunk study that appeared in the text and (if applicable) in summary tables of published reviews.

Results

The Brunk Study: Findings and Reports

The sole published report on the Brunk study provided data on 33 families who

completed treatment (77% of 43 families in the experiment). Pre- and post-test means were presented for subgroups (abuse or neglect) within treatment conditions (MST or PT), but only for outcome measures with significant changes; standard deviations were not provided. Results were analyzed with 2 X 2 X 2 (pre-post X subgroup X treatment) multivariate analysis of variance (MANOVAs) of 4 groups of outcome measures. Three-way univariate ANOVA was used for individual outcome measures. Child age and parental age were used as covariates in the MANOVAs and ANOVAs. F-values were provided in the text, but only for results that were statistically significant. No follow-up data or intent-to-treat analyses were provided.

Table 3 provides a summary of results provided by Brunk et al. (1987). According to the text and tables of the original report, there were 16 client self-report measures (including 10 subscales of the Family Environment Scale, FES). PT was superior to MST on one measure, results for one measure (the Behavior Problem Checklist, BPC) were not reported (presumably because there were no significant differences between pre- and post-test scores), and there were no significant differences between treatment groups on the remaining 14 measures. There were no significant differences between MST and PT on 3 measures derived from therapist reports. On 11 observational measures, 5 favored MST, 1 favored PT, 2 showed subgroup interaction effects (MST was superior for one subgroup but not the other) with no significant main effects, and 2 showed no significant differences between MST and PT. Thus, for 30 possible tests of main effects, MST was superior to PT on 5 tests, PT was superior on 2 tests, there were no significant differences on 22 tests, and results of 1 test were not reported.

Since authors reported main effects of treatment and treatment effects for 2 subgroups plus 4 multivariate analyses, there were at least 94 possible tests of significance in which effects of MST could have appeared. With a total of 33 cases in 3-way analyses with 2 covariates, these tests had little statistical power. Nevertheless, with alpha set at p=.05, we would expect about 5

(4.7) of 94 tests to be statistically significant purely by chance.

Content analysis of the authors' summary of results in the abstract produced 5 codeable phrases, indicating that there were no significant between-group differences in 3 domains ("parental psychiatric symptomology, reduced overall stress, and… severity of identified problems"), MST was superior in 1 domain ("restructuring parent-child relations"), and PT was superior in another ("reducing identified social problems").

Research Reviews

Of the 37 reviews in the previous analysis, 17 cited the Brunk study, but only 13 provided specific comments on results of that study. The bottom portion of Table 4 shows results of content analysis of the text and tables in these 13 reviews.

Burns et al. (2000) summarized the results of Brunk in this way: "Parents in both groups reported decreases in psychiatric symptomatology and reduced overall stress following treatment. In addition, both groups demonstrated decreases in the severity of the identified problems. The study also included observational measures of parent-child interactions. The outcomes indicated that MST had improved such interactions, implying a decreased risk for maltreatment of children in the MST condition" (p. 293). The reviewers also provide a table describing MST studies; under the column headed "MST outcomes," the entry for the Brunk study reads, "improved parent-child relations" (p. 291). Table 4 shows the coding of these comments: 3 neutral phrases and 2 positive phrases in the text, and 1 positive phrase in the table.

Regarding the Brunk study, Corcoran (2000) stated:

> "Both approaches acted to reduce psychiatric symptoms in parents and parental stress, as well as to alleviate individual and family problems. Each approach also offered unique advantages. Multisystemic therapy was more effective in improving parent-child interactions, helping physically abusive parents manage child behavior, and assisting

> neglectful parents in responding more appropriately to their child's needs. Surprisingly, parent training was more advantageous for improving parents' social lives. The hypothesis is the group setting for parent training reduced isolation and improved parents' support system." (p. 568).

Shown in Table 4, this passage is coded as having 4 neutral phrases, 3 phrases that favor MST, and 3 phrases that favor PT.

Curtis et al. (2004) calculated an average effect size for the Brunk study, presumably across all outcome measures. They report a result of $d = 1.32$ (sd = .65, N = 43; p. 414). This is an enormous effect size (it indicates that, after treatment, the average family in the MST group was functioning better than 90% of cases in the PT group across all outcome measures). Given the results of the Brunk study, this appears to be a mistake. Littell and colleagues (2005) and David Wilson (an expert on effect size calculations) could not derive effect sizes from the Brunk report. Wilson, however, was able to approximate the effect size reported by Curtis et al., but only by: 1) ignoring all non-significant differences, 2) assuming that all significant differences favored MST, and 3) misusing effect size formulas (treating reported F values as if they were from one-way ANOVAs, ignoring variance extracted in the original analysis which used a mixed factorial design with covariates); even then, the d-index he obtained was not statistically significant (David B. Wilson, personal communication, March 2, 2005).

Henggeler and associates (Henggeler, Schoenwald, Rowland, et al., 2002) cited the Brunk study as support for the following statement, "MST has consistently produced improvements in family functioning across outcome studies with juvenile offenders and maltreating families. Several of these studies used observational methods to demonstrate increased positive family interactions and decreased negative interactions" (p. 209). These authors used a single phrase to characterize results of the Brunk study in a table of MST

outcomes: “improved parent-child interactions” (p. 207).

Other reviewers summarized results of the Brunk study as follows (see codes in Table 4):

- “MST was significantly more effective [than PT] at restructuring problematic parent-child relations” (Henggeler, Mihalic, et al., 1998, p. 33).
- “Successful MST outcomes have been observed… for children in maltreating families” (Henggeler & Sheidow, 2003, p. 512).
- “The effects of [MST] have been further demonstrated among …abused or neglected children” (Hoagwood et al., 2001, p. 1183).
- “The [MST] outcome studies have extended to … parents who engage in physical abuse or neglect…. Thus, the model of providing treatment may have broad applicability across problem domains among seriously disturbed children" (Kazdin, 1998, p. 79).
- MST “treatment effects have been replicated … with parents who engage in physical abuse or neglect” (Kazdin & Weisz, 1998, p. 27).
- “MST is considered to be a promising treatment for families with children who are at risk of being abused by their parents” (Pushak, 2002).
- “Randomized trials with… families in which maltreatment occurred (Brunk, Henggeler, & Whelan, 1987) suggested the promise of MST with these populations” (Schoenwald & Rowland, 2002, p. 113).
- “MST was more effective than Parent Training for improving parent-child interactions associated with maltreatment. Abusive parents showed greater progress in controlling their child’s behavior, maltreated children exhibited less passive noncompliance, and neglecting parents became more responsive to their child’s behavior. Parent training was superior to MST [in] decreasing social problems (i.e., social support network)” (Swenson & Henggeler, 2003, pp. 75-76).

- "The effectiveness of MST has been supported in controlled outcome studies with… maltreating families" (Swenson et al., 1998, p. 332).

As shown in Table 4, three patterns emerged as we traced results from the original measures, through the published report, to published reviews of these findings. The first pattern is *overall data reduction*: a complex pattern of results was summarized in increasingly more succinct ways (evident in the column on "number of items"). This reduction is often essential if results are to be conveyed in ways that are meaningful and accessible to diverse audiences.

The second trend is a *reduction in uncertainty*: the proportion of neutral items or statements diminished as the data (i.e., total number of items or statements) were reduced. Put more succinctly, non-significant differences were minimized. This trend becomes troubling when the weight of the evidence – the balance between positive, negative, and neutral items – is not adequately represented. In the original report, the proportion of neutral items dropped from 77% (23) of 30 subscales, to 76% of 29 reported results, to 63% of 19 provided results, to 60% (3) of 5 comments in the abstract. Although the balance was not perfect, even the abstract indicated that there were more between-group similarities than differences in outcomes. However, only 2 reviews even mentioned neutral (null) results; the other 11 reviews appeared to ignore the modal pattern of non-significant results in the Brunk study.

Third, while the original research report retained a balance between positive, neutral, and negative results, this balance was absent in all but 1 of 13 reviews (Corcoran, 2000). Most of the reviews over-emphasized the positive results of MST and minimized or ignored other kinds of evidence. In fact, 11 reviews used a *single* positive number or statement to characterize results of the Brunk study in their text or tables. Thus, there is evidence of *confirmation bias in* reviewers' summaries of the Brunk study.

Limitations

Based on a nonprobability sample of published reviews, the results reported here are not generalizable to other reviews, to reviews of interventions other than MST, or to MST trials other than the Brunk study. However, many of the reviews in this analysis also considered evidence about other interventions and included other MST trials; there is no logical reason to believe that these reviewers would have handled the evidence for MST differently from evidence on other programs. Nor is it sensible to think that reviewers would treat the Brunk study differently from other MST studies.

Discussion

Reviewers use different methods and criteria to identify, analyze, and synthesize empirical evidence. Most of the 37 reviews in this study relied on narrative summaries of convenience samples of published studies. This approach has been shown to be vulnerable to several sources and types of bias. Fewer than one-quarter of the reviews use explicit inclusion criteria, systematic search strategies, unpublished studies, assessment of study allocation methods, assessment of attrition, or quantitative synthesis. Independent reviews were more likely to use some of these strategies, but less likely to include unpublished reports when compared with reviews authored by program developers. Some reviews were partially systematic, but none met established criteria for systematic reviews.

Some reviews did not aim (or claim) to be comprehensive or systematic; nevertheless, they drew conclusions about effects of MST (see Table 1) that are only warranted when based on a comprehensive, systematic review. In some reviews, these claims were based on very few studies.

Reviews tended to confirm prevailing beliefs, even when the data were equivocal. As prior conclusions were repeated, readers may have mistaken this consistency for valid evidence. (Not included in this analysis were 20 published summaries of MST trials that relied primarily or

solely on previous published reviews; e.g., Kazdin, 2000, 2002; Lehman, Goldman, Dixon & Churchill, 2004; US DHHS, 2001; U.S. National Institutes of Health, 2004). Confirmation bias appeared in independent reviews as well as those authored by program developers.

Understanding Confirmation Bias

Confirmation bias is the ubiquitous, often unintentional tendency to seek information that supports a hypothesis, give preferential treatment to evidence that confirms existing beliefs, and dismiss evidence to the contrary (Nickerson, 1998). Initially identified by Francis Bacon (1621/1960) and investigated by Watson (1960, 1968) and others, numerous studies show that people (including scientists) are reluctant to consider evidence that is inconsistent with their predictions (e.g., Fugelsang, Stein, Green, & Dunbar, 2004; Mahoney, 1977). This may be because confirmatory information is easier to process cognitively. That is, it is easier to see how information supports a position than it is to see how the same information might counter that position. Further, information that supports a hypothesis is more likely to be recalled than information to the contrary (Gilovich, 1993).

Confirmation bias may be the source of many myths and self-fulfilling prophesies. It gives us an illusion of consistency, leads us to misinterpret new information, and induces over-confidence in beliefs (Nickerson, 1998, Schrag, 1999). The scientific method is constructed to compensate for this human tendency, so that we must try to disprove our hypotheses. This strategy of falsifying hypotheses is not something that people do naturally (Watson, 1960, 1968; Nickerson, 1998).

Confirmation Bias in Political Context

Policy makers, practitioners, and scholars want to know what works best in response to pressing human and social problems. Most of the reviews in this study were written by scholars and experts in the U.S. at a time when there was pressure to demonstrate the efficacy,

effectiveness, and cost-effectiveness of psychosocial interventions to insure their continued political and financial support. This pressure may have exacerbated the natural tendency to seek information that confirms our hopes and expectations. For example, "As pressure increases for the demonstration of effective treatment for children with mental disorders, it is essential that the field has an understanding of the evidence base. To address this aim, the authors searched the published literature for *effective* interventions for children and adolescents…" (Burns et al., 1999, p. 199; emphasis added). Hence, these authors cited "studies with large effect sizes" in child welfare, but did not the mention larger, more rigorous trials in that field that did not produce large effect sizes.

To their credit, the researchers and reviewers who sought to demonstrate effects of treatment cared about evidence. They sought to improve fields of practice that relied (and still rely) on practices that are largely untested. However, in the search for positive, confirming examples of effective interventions, it seems that valuable information on ineffective or harmful practices was ignored. The focus on positive evidence detracts from a full understanding of the evidence base.

The pressure to find out what works best pits one program against others. In this competitive, market-driven context, the real message of the Brunk study—that different interventions have different effects—was lost. Following Brunk, the choices policy makers face may depend on which approaches or outcomes they *prefer*. For example, is it more important to reduce parents' social problems or improve aspects of parent-child interactions? The Brunk study provided no guidance on which outcomes were "better" or more important than others (there were no *a priori* hypotheses in this regard), but other studies might.

Confirmation Bias and the High-Fidelity Hypothesis

To explain variations in outcomes across MST trials, several authors have pointed to the

finding of Curtis et al. (2004) that appear to indicate that MST performed better in efficacy studies than in studies of effectiveness (Henggeler; 2004; Petrosino & Soydan, 2005). In their analysis of 7 MST trials, Curtis et al. (2004) classified the Brunk study as 1 of 3 efficacy studies in which MST developers exercised more control over the treatment conditions than in the remaining 4 studies. It is unclear why the Simpsonville South Carolina project (also known as the FANS study) was not included in the efficacy category, after Henggeler and colleagues described this study as one of the trials "in which the developers of MST provided ongoing clinical supervision and consultation (i.e., quality assurance was high)" (Henggeler, Schoenwald, Rowland, et al., 2002, p. 211). Since Schoenwald and colleagues observed that these original MST trials "could be considered hybrids of 'efficacy' and 'effectiveness' research" (Schoenwald, Sheidow, Letourneau, & Liao, 2003, p. 234), it appears that post hoc classifications were used in the Curtis et al. study.

Using an implausibly high effect size (d = 1.32) for the Brunk study (discussed above), Curtis and colleagues calculated a pooled effect size of d = .81 for 3 efficacy studies compared with an average ES of d = .26 for 4 studies of effectiveness. For unknown reasons, corrections for small sample bias were only applied to 1 study, and not to the Brunk study (valid N = 33, not 43 as reported by Curtis et al., 2004). Pooled ES were not weighted using inverse variance methods; hence, it appears that the Brunk study contributed as much to the average effect for efficacy studies as results from a much larger study (N=176 cases) with a smaller ES.

These results have been used to suggest that the impact of program developers-as-evaluators on results of controlled trials has more to do with their fidelity than allegiance bias (Henggeler, 2004; Petrosino & Soydan, 2005). However, the calculations by Curtis and colleagues do not provide a sound basis for any conclusions about the efficacy of MST or high-fidelity conditions.

From Efficacy to Effectiveness to Transportability: On What Basis?

Several states, professional organizations, private foundations, and federal agencies have taken the lead in identifying evidence-based practices and encouraging their replication. Now that lists of evidence-based practices have been compiled by experts (often with U.S. government funding), the emphasis in the health and mental health fields has begun to shift from research synthesis to translation of results into directions for policy and practice, from questions about efficacy and effectiveness to concerns about transportability and dissemination. The movement to transport "effective" practices is a high priority for many government agencies (including the U.S. National Institutes of Health). This is based on the twin assumptions that 1) we already know "what works" in response to certain pressing social problems and 2) this knowledge, derived largely from controlled studies, can be applied with success in other samples and settings. However, results reported here raise important questions about the validity of current knowledge about "what works" for certain problems and populations.

The movement to transport "effective" practices may be premature if it is based on evidence of efficacy or effectiveness that has been compiled with haphazard reviews that are vulnerable to publication, selection, and confirmation biases. If knowledge about "what works" is tainted in these ways, we may waste valuable resources trying to transport ineffective practices (albeit ones that have produced *some* positive results in a few controlled trials) and fail to investigate other practices that may be equally or more effective. A closer look at the evidence is warranted.

Implications for Social Science

While practitioners and policy-makers are urged to make better use of scientific evidence, it is ironic that social scientists rarely cumulate evidence scientifically (Chalmers, Hedges & Cooper, 2002, p. 12). To support evidence-based practice and policy, social scientists must make

better use of the science of research synthesis. “If a review purports to be an authoritative summary of what ‘the evidence’ says, then the reader is entitled to demand that this is a comprehensive, objective, and reliable overview, and not a partial review of a convenience sample of the author’s favorite studies” (Petticrew & Roberts, 2006, p. 6).

Advanced training in systematic review methods is needed to prepare the next generation of scholars to produce valid evidence for policy and practice. Systematic methods can minimize bias in the review process. Systematic reviews can incorporate contradictory information--including much of the data that have been lost in traditional, narrative or haphazard reviews—and use it to answer important questions about why intervention effects may vary. These reviews are very labor intensive, hence they are more costly than traditional literature reviews; but systematic reviews may be more cost-effective in the long run if they reduce bias (misinformation) and prevent missteps in the development and dissemination of effective practices.

Reviewers often struggle with decisions about the types of evidence to consider, sometimes lowering the bar (deviating from their original standards) in order to be able to say *something*. This is a slippery slope. Reasonable people will disagree about where to set the bar regarding the qualities of evidence needed to support certain inferences. These decisions should be based on careful consideration of substantive, contextual, and methodological issues. Once the decision is made, it is worrisome when reviewers deviate from their original plan (this is not in accordance with the principles of systematic reviews). The Cochrane Collaboration has taken another approach by publishing “empty” reviews that found no credible evidence on a topic. This may not be very satisfying to reviewers or policy makers, but one advantage of an empty (or hyper-vigilant) review is that it does not lead readers to the wrong conclusions. Empty reviews identify important gaps in current knowledge and provide justification for new studies.

It should be recognized that, when properly-implemented, randomized controlled trials provide the most credible evidence of *effects* of social programs (Glaserman, Levy, & Myers, 2002). Questions about "what works" have dominated the discourse about evidence-based practice and policy, although there are other empirical questions are relevant (Davies, 2004; Gibbs, 2003). Current hierarchies of evidence are inadequate to handle the array of important empirical questions for practice and policy (Petticrew & Roberts, 2003).

The peer-review process must be strengthened to counter various forms of selection bias, including confirmation bias (Mahoney, 1977). This is not an easy task, but some important inroads have been made. The American Psychological Association (APA) journals recently adopted the CONSORT statement (Moher, Schultz & Altman, 2001), which provides clear guidelines for reporting trials. Journals should also adopt the QUOROM statement to increase the quality of reporting on meta-analysis and other research reviews. Ultimately social scientists must join health scientists in endorsing the use of prospective registers of trials to avoid publication bias and outcome selection bias (Dickersin, 2005).

Evidence-based practice requires a long-term commitment to building valid information for practice and policy, and an infrastructure that provides consumers with access to relevant and regularly-updated information. Careful primary research and research synthesis can help build an evidence base for the helping professions. However, on questions about "what works," there have been far too few controlled trials and too many haphazard reviews of these trials to produce enough valid evidence for practice and policy—and valuable information has been lost along the way. More scientifically-sound syntheses of credible empirical studies are needed to provide a valid evidence-base for practice.

References

Aos S., Phipps, P., Barnoski, R., & Lieb, R. (2001). *The comparative costs and benefits of programs to reduce crime* (Version 4.0). Document Number 01-05-1201. Washington State Institute for Public Policy.

Bacon, F. (1621/1960). *Novum organum.* New York: Bobbs-Merrill.

Becker, B. J., Hedges, L., & Pigott, T. D. (2004). *Campbell Collaboration Statistical Analysis Policy Brief*. Retrieved June 12, 2006, from http://www.campbellcollaboration.org/MG/StatsPolicyBrief.pdf

Begg, C. B. (1994). Publication bias. In H. Cooper & L. Hedges (Eds.), *The handbook of research synthesis* (pp. 399-409). New York: Russell Sage Foundation.

Borduin, C. M. (1999). Multisystemic treatment of criminality and violence in adolescents. *Journal of the American Academy of Child and Adolescent Psychiatry, 38*, 242-249.

Borduin, C. M., Heiblum, N., Jones, M. R., Grabe, S. A. (2000). Community-based treatments of serious antisocial behavior in adolescents. In W. E. Martin, & J. L Schwartz-Kulstad (Ed.), *Person-environment psychology and mental health: Assessment and intervention.* (pp. 113-141). Mahwah, NJ: Lawrence Erlbaum Associates.

Borduin, C. M., Schaeffer, C. M., & Ronis, S. T. (2003). Multisystemic treatment of serious antisocial behavior in adolescents. In C. A. Essau (Ed.), *Conduct and oppositional defiant disorders: Epidemiology, risk factors, and treatment* (pp. 299-318). Mahwah, NJ: Lawrence Erlbaum Associates.

Brestan, E. V., & Eyberg, S. M. (1998). Effective psychosocial treatments of conduct-disordered children and adolescents: 29 years, 82 studies, and 5,272 kids. *Journal of Clinical Child Psychology, 27,* 180-189.

Brosnan, R., & Carr, A. (2000). Adolescent conduct problems. In A. Carr (Ed.), Whats works with children and adolescents?: A critical review of psychological interventions with children, adolescents and their families (pp. 131-154). London: Routledge.

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

Burns, B. J. (2003). Children and evidence-based practice. *Psychiatric Clinics of North America, 26,* 955-870.

Burns, B. J., Hoagwood, K., & Mrazek, P. J. (1999). Effective treatment for mental disorders in children and adolescents. *Clinical Child and Family Psychology Review, 2,* 199-244.

Burns, B. J., Schoenwald, S. K., Burchard, J. D., Faw, L., Santos, A. B. (2000). Comprehensive Community-Based Interventions for Youth with Severe Emotional Disorders: Multisystemic Therapy and the Wraparound Process. *Journal of Child and Family Studies, 9,* 283-314.

Bushman, B. J. (1994). Vote-counting procedures in meta-analysis. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 193-213). New York: Russell Sage Foundation.

Bushman, B. J., & Wells, G. L. (2001). Narrative impressions of literature: The availability bias and the corrective properties of meta-analytic approaches. *Personal and Social Psychology Bulletin, 27*, 1123-1130.

Carlton, P. L., & Strawderman, W. E. (1996). Evaluating cumulated research I: The inadequacy of traditional methods. *Biological Psychiatry, 39,* 65-72.

Center for Substance Abuse Prevention (2000). *Strengthening America's families: Model family programs for substance abuse and delinquency prevention.* Department of Health

Education, University of Utah, N215 HPER. Salt Lake City, UT 84112.
Chalmers, I., Hedges, L. V., & Cooper, H. (2002). A brief history of research synthesis. *Evaluation & The Health Professions, 25*(1), 12-37.
Chambless, D. L., Baker, M. J., Baucom, D. H., Beutler, L. E., Calhoun, K. S., Crits-Christoph P., et al. (1998). Update on empirically validated therapies, II. *The Clinical Psychologist, 51,* 3-16.
Chorpita, B. F., Yim, L. M., Donkervoet, J. C., Arensdorf, A., Amundsen, M. J., McGee, C., Serrano, A., Yates, A., Burns, J. A., & Morelli, P. (2002). Toward large-scale implementation of empirically supported treatments for children: A review and observations by the Hawaii empirical basis to services task force. *Clinical Psychology: Science and Practice, 9,* 165-190.
Cooper, H., & Hedges, L. V. (Eds.) (1994). *The handbook of research synthesis*. New York: Russell Sage Foundation.
Corcoran, J. (2000). Family interventions with child physical abuse and neglect: A critical review. *Children and Youth Services Review, 22,* 563-591.
Corcoran, J. (2003). *Clinical applications of evidence-based family interventions.* New York: Oxford University Press.
Cormack, C., & Carr, A. (2000). Drug abuse. In A. Carr (Ed.), What works for children and adolescents? A critical review of psychological interventions with children, adolescents and their families (pp. 155-178). London: Routledge.
Curtis, N. M., Ronan, K. R., & Borduin, C. M. (2004). Multisystemic Treatment: A meta-analysis of outcome studies. *Journal of Family Psychology, 18*, 411-419.
Davies, P. (2004). Evidence-based government... Is it possible? Paper presented at the Fourth Annual Campbell Collaboration Colloquium, Washington, DC, February 19, 2004.
Dickersin, K. (2005). Publication bias: Recognizing the problem, understanding its origins and scope, and preventing harm. In H. R. Rothstein, A. J. Sutton & M. Borenstein (Eds.), *Publication bias in meta-analysis: Prevention, assessment, and adjustments*. Chichester, UK: John Wiley & Sons.
Farrington, D. P., & Welsh, B. C. (2003). Family-based prevention of offending: A meta-analysis. *The Australian and New Zealand Journal of Criminology, 36,* 127-151.
Fonagy, P., & Kurtz, A. (2002). Disturbance of conduct. In M. T. Peter Fonagy, David Cottrell, Jeannette Phillips, Zarrina Kurtz (Ed.), *What works for whom? A critical review of treatments for children and adolescents* (pp. 106-191). New York: Guildford Press.
Fraser, M. W., Nelson, K. E., & Rivard, J.C. (1997). Effectiveness of family preservation services. *Social Work Research, 21,* 138-153.
Fugelsang, J., Stein, C., Green, A., & Dunbar, K. (2004). Theory and data interactions of the scientific mind: Evidence from the molecular and the cognitive laboratory. *Canadian Journal of Experimental Psychology, 58*, 132-141.
Gambrill, E. (2006). The ethics of transparency: Systematic reviews and their rivals. Paper presented at the Sixth annual Campbell Collaboration Colloquium, Los Angeles, CA, February 24, 2006.
Gibbs, L. E. (2003). *Evidence-based practice for the helping professions: A practical guide with integrated multimedia*. Pacific Grove, CA: Brooks/Cole-Thompson Learning.
Gilgun, J. F. (2005). The four cornerstones of evidence-based practice in social work. *Research on Social Work Practice, 15*(1), 52-61.
Gilovich, T. (1993). *How We Know What Isn't So: The Fallibility of Human Reason in Everyday Life.* New York: Free Press.
Glazerman, S., Levy, D. M., & Myers, D. (2002). *Nonexperimental replications of social*

experiments: A systematic review. Princeton, NJ: Mathematica Policy Research, Inc.

Henggeler, S. W. (2003). Multisystemic therapy: An overview. Dissemination, data, and direction. PowerPoint presentation. NASMHPD Research Institute Conference, February 2003.

Henggeler, S. W. (2004). Decreasing effect sizes for effectiveness studies - Implications for the transport of evidence-based treatments: Comment on Curtis, Ronan, and Borduin (2004). *Journal of Family Psychology, 18*, 420-423.

Henggeler, S. W. & Borduin, C. M. (1995). Multisystemic treatment of serious juvenile offenders and their families. In I. M. Schwartz & P. AuClaire (Eds)., *Home-based services for troubled children* (pp. 113-130). Lincoln, NE: University of Nebraska Press.

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

Henngeler, S. W., Schoenwald, S. K., Borduin, C. M., & Swenson, C. C. (2006). Methodological critique and meta-analysis as Trojan horse. *Children and Youth Services Review, 28,* 447-457.

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

Henggeler, S., W., Schoenwald, S. K., Liao, J. G., Letourneau, E. J., & Edwards, D. L. (2002). Transporting efficious treatments to field settings: The link between supervisory practices and therapist fidelity in MST programs. *Journal of Clinical Child and Adolescent Psychology, 31,* 155-167.

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

Henggeler, S. W., & Sheidow, A. J. (2003). Conduct disorder and delinquency. *Journal of Marital and Family Therapy, 29*(4), 505-522.

Higgins J. P. T., & Green, S. (Eds). (2006). Cochrane Handbook for Systematic Reviews of Interventions, Version 4.2.5 (updated September 2006). In: *The Cochrane Library, Issue 4,* 2006. Chichester, UK: John Wiley & Sons, Ltd. Accessed April 12, 2007 at: http://www.cochrane.org/resources/handbook/Handbook4.2.6Sep2006.pdf

Hoagwood, K., Burns, B. J., Kiser, L., Ringeisen, H., & Schoenwald, S. K. (2001). Evidence-based practice in child and adolescent mental health services. *Psychiatric Services, 52,* 1179-1189.

Hopewell, S., Clarke, M., Lefebvre, C., & Scherer, R. (2006). Handsearching versus electronic searching to identify reports of randomized trials. In *The Cochrane Database of Systematic Reviews, 2006, Issue 4*. Chichester, UK: John Wiley & Sons, Ltd.

Hopewell, S., McDonald, S., Clarke, M., & Egger, M. (2006). Grey literature in meta-analyses of randomized trials of health care interventions. In *The Cochrane Database of Systematic Reviews, 2006, Issue 2*. Chichester, UK: John Wiley & Sons, Ltd.

Jadad, A. R., Moher, M., Browman, G. P., Booker, L., Sigouin, C., Fuentes, M., et al. (2000). Systematic reviews and meta-analyses on treatment of asthma: Critical evaluation. *British Medical Journal, 320*, 537-540.

Jensen, P. S., Weersing, R., Hoagwood, K. E., & Goldman, E. (2005). What is the evidence for evidence-based treatments? A hard look at our soft underbelly. *Mental Health Services Research, 7*(1), 53-74.

Jørgensen, A. W., Hilden, J., & Gøtzsche, P. G. (2006). Cochrane reviews compared with industry supported meta-analyses and other meta-analyses of the same drugs: Systematic review. *British Medical Journal, 333*, 782-785.

Kazdin, A. E. (1998). Psychosocial treatments for conduct disorder in children. In P. E. Nathan & J. M. Gorman (Eds.), *A guide to treatments that work* (pp. 65-89). New York: Oxford University Press.

Kazdin, A. E. (2000). Treatments for aggressive and antisocial children. *Child and Adolescent Psychiatric Clinics of North America, 9,* 841-858.

Kazdin, A E. (2002). The state of child and adolescent psychotherapy research. *Child and Adolescent Mental Health, 7,* 53-59.

Kazdin, A. E., & Weisz, J. R. (1998). Identifying and developing empirically supported child and adolescent treatments. *Journal of Consulting and Clinical Psychology, 66,* 19-36.

Lehman, A. F., Goldman, H. H., Dixon, L. B., & Churchill, R. (2004). *Evidence-based mental health treatments and services: Examples to inform public policy*. New York: Milbank Memorial Fund.

Letourneau, E. J., Cunningham, P. B., Henggeler, S. W. (2002). Multisystemic treatment of antisocial behavior in adolescents. In S. G. Hofmann, & M. C., Tompson (Ed.), *Treating chronic and severe mental disorders: A handbook of empirically supported interventions.* (pp. 364-381.). New York: Guilford Press.

Lipman, T. (2000). Power and influence in clinical effectiveness and evidence-based medicine. *Family Practice, 17*(6), 557-563.

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

Littell, J. H. (2006). The case for Multisystemic Therapy: Evidence or orthodoxy? *Children and Youth Services, 28,* 458-472.

Littell, J. H. (2005). Lessons from a systematic review of effects of Multisystemic Therapy. *Children and Youth Services Review, 47,* 445-463.

Littell, J. H., Popa, M., & Forsythe, B. (2005). Multisystemic Therapy for social, emotional, and behavioral problems in youth aged 10-17 (Cochrane Review). In: *The Cochrane Database of Systematic Reviews, Issue 4,* 2005. Chichester, UK: John Wiley & Sons, Ltd.

Luborsky, L., Diguer, L., Seligman, D. A., Rosenthal, R., Krause, E. D., Johnson, S., et al. (1999). The researcher's own therapy allegiances: A 'wild card' in comparisons of treatment efficacy. *Clinical Psychology: Science and Practice, 6*, 95-106.

Mahoney, M. J. (1977). Publication prejudices: An experimental study of confirmatory bias in the peer review system. *Cognitive Therapy and Research, 1*(2), 161-175.

Miller, R. B., Johnson, L. N., Sandberg, J. G., Stringer-Seibold, T. A., & Gfeller-Strouts, L. (2000). An addendum to the 1997 outcome research chart. *American Journal of Family Therapy, 28*(4), 347-354.

Miller, W. R., Brown, J. M., Simpson, T. L., Handmaker, N. S., Bien, T. H., Luckie, L. H., et al. (1995). What works? A methodological analysis of the alcohol treatment outcome literature. In R. K. Hester & W. R. Miller (Eds.), *Handbook of alcoholism treatment approaches: Effective alternatives* (2nd ed., pp. 12-44). Needham Heights, MA: Allyn & Bacon.

Moher, D., Cook, D. J., Eastwood, S., Olkin, I., Rennie, D., Stroup, D. F., et al. (1999). Improving the quality of reports of meta-analyses of randomised controlled trials: The QUOROM statement. *The Lancet, 354*, 1896-1900.

Moher, D., Schultz, K. F., Altman, D. G., for the CONSORT Group (2001). The CONSORT statement: Revised recommendations for improving the quality of reports of parallel-

group randomized trials. *The Lancet, 357,* 1191-94.
Also published in *JAMA, 285,* 1987-91 and *Annals of Internal Medicine, 134,* 657-62.

MST Services, Inc. (2006). Licensed MST programs. Accessed June 11, 2006 at http://www.mstservices.com/text/licensed_agencies.htm

National Institute of Mental Health (2001). *Youth in a difficult world*. NIH Publication No. 01-4587. Accessed June 12, 2006 at http://www.nimh.nih.gov/publicat/youthdif.cfm

National Institute of Mental Health Consortium on Child and Adolescent Research (2003). *Data trends*. http://www.nimh.nih.gov/childhp/datatrends.cfm

National Institute on Drug Abuse (1999). *Principles of drug addiction treatment: A research-based guide.* (NIH Publication 99-4180). Bethesda, MD: Author.

National Institute on Drug Abuse (2003). Effective drug abuse treatment approaches: Multisystemic therapy. NIDA Behavioral Therapies Development Program. http://www.nida.nih.gov/BRDP/Effective/Henggeler.html

Nickerson, R.S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology, 2*, 175-220.

Office of Juvenile Justice and Delinquency Prevention. Model programs guide. Accessed June 11, 2006 at http://www.dsgonline.com/mpg2.5/mpg_index.htm

Petrosino, A., & Soydan, H. (2005). The impact of program developers as evaluators on criminal recidivism: Results from meta-analyses of experimental and quasi-experimental research. *Journal of Experimental Criminology, 1*, 435-450.

Petticrew, M., & Roberts, H. (2003). Evidence, hierarchies, and typologies: Horses for courses. *Journal of Epidemiological Community Health, 57*, 527-529.

Petticrew, M., & Roberts, H. (2006). *Systematic reviews in the social sciences: A practical guide*. Oxford, UK: Blackwell Publishing, Ltd.

Pushak, R. E. (2002). The dearth of empirically supported mental health services for children: Multisystemic Therapy as a promising alternative. *Scientific Review of Mental Health Practice, 1.*

Randall, J., & Cunningham, P. B. (2003). Multisystemic therapy: A treatment for violent substance-abusing and substance-dependent juvenile offenders. *Addictive Behaviors, 28*, 1731-1739.

Rosenthal, M. C. (1994). The fugitive literature. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 85-94). New York: Russell Sage Foundation.

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

Rothstein, H. R., Turner, H. M., & Lavenberg, J. G. (2004). *The Campbell Collaboration Information Retrieval Policy Brief*. Retrieved June 12, 2006, from http://www.campbellcollaboration.org/MG/IRMGPolicyBriefRevised.pdf.

Sackett, D. L., Rosenberg, W. M. C., Gray, J. A. M., Haynes, R. B., & Richardson, W. S. (1996). Evidence based medicine: What it is and what it isn't. *British Medical Journal, 312*, 71-12.

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

Schoenwald, S. K., Henggeler, S. W., Brondino, M. J., & Rowland, M. D. (2000). Multisystemic therapy: Monitoring treatment fidelity. *Family Process, 39,* 83-103.

Schoenwald, S. K., & Hoagwood, K. (2001). Effectiveness, transportability, and dissemination of interventions: What matters when? *Psychiatric Services, 52,* 1190-1197.

Schoenwald, S. K., & Rowland, M. S. (2002). Multisystemic therapy. In B. J. Burns & K.

Hoagwood (Eds), *Community treatment for youth: Evidence-based interventions for severe emotional and behavioral disorders* (pp. 91-116). New York: Oxford University Press.

Schoenwald, S. K., Sheidow, A. J., Letourneau, E. J., & Liao, J. G. (2003). Transportability of multisystemic therapy: Evidence for multilevel influences. *Mental Health Services Research, 5*, 223-239.

Schrag, J. L. (1999). First impressions matter: A model of confirmatory bias. *The Quarterly Journal of Economics, 114*, 37-82.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for general causal inference*. Boston: Houghton Mifflin.

Shadish, W., & Myers, D. (2004). *Campbell Collaboration Research Design Policy Brief.* Retrieved June 12, 2006 from http://www.campbellcollaboration.org/MG/ResDesPolicyBrief.pdf

Shea, B., Moher, D., Graham, I., Pham, B., & Tugwell, P. (2002). Comparison of the quality of Cochrane reviews and systematic reviews published in paper-based journals. *Evaluation & The Health Professions, 25*(116-129).

Smith, C. A., & Stern, S. B. (1997). Delinquency and antisocial behavior: A review of family processes and intervention research. *Social Service Review, 71,* 382-420.

Substance Abuse and Mental Health Service Administration (2004). Multisystemic Therapy: Proven results. Accessed June 11, 2006 at: http://modelprograms.samhsa.gov/pdfs/FactSheets/Mst.pdf.

Sudderth, L. K. (2000). What works in treatment programs for substance-abusing youth. In M. P. Kluger, G. Alexander, & P. A. Curtis (Eds.), *What works in child welfare* (pp. 337-344). Washington, DC: Child Welfare League of America, Inc.

Swenson C. C., & Henggeler, S. W. (2003). Multisystemic therapy (MST) for maltreated children and their families. In B. E. Saunders, L. Berliner, & R. F. Hanson (Eds.), *Child Physical and Sexual Abuse: Guidelines for treatment (Final report: January 15, 2003)* (pp. 75-78). Charleston, SC: National Crime Victims Research and Treatment Center.

Swenson, C. C., Henggeler, S. W., Schoenwald, S. K., Kaufman, K. L., & Randall, J. (1998). Changing the social ecologies of adolescent sexual offenders: Implications of the success of multisystemic therapy in treating serious anti-social behavior in adolescents. *Child Maltreatment, 3*, 330-338.

Tarolla, S. M., Wagner, E. F., Rabinowitz, J., & Tubman, J. G. (2002). Understanding and treating juvenile offenders: A review of current knowledge and future directions. *Aggression and Violent Behavior, 7,* 125-143.

U.S. Department of Health and Human Services. (1999). *Mental health: A report of the Surgeon General*. Rockville, MD: Author.

U.S. Department of Health and Human Services. (2001). *Youth violence: A report of the Surgeon General.* Washington, DC: US DHHS Office of the Surgeon General.

U.S. National Institutes of Health (2004). *State-of-the-science Conference Statement: Preventing violence and related health-risking social behaviors in adolescence.* Retrieved 18 October 2004 from http://consensus.nih.gov/ta/023/youthviolenceDRAFTstatement101504.pdf

Vaughn, M. G., & Howard, M. O. (2004). Adolescent substance abuse treatment: A synthesis of controlled evaluations. *Research on Social Work Practice, 14,* 325-335.

Watson, P.C. (1968). Reasoning about a rule. *Quarterly Journal of Experimental Psychology, 20*, 273-281.

Watson, P.C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly*

Journal of Experimental Psychology, 12, 129-140.

Woolfenden, S. R., Williams, K., & Peat, J. K. (2003). Family and parenting interventions in children and adolescents with conduct disorder and delinquency aged 10-17 (Cochrane Review) In: *The Cochrane Database of Systematic Reviews,* Issue 2, 2003. Chichester, UK: John Wiley & Sons, Ltd.

Woolfenden, S. R., Williams, K., & Peat, J. K. (2002). Family and parenting interventions for conduct disorder and delinquency: A meta-analysis of randomized controlled trials. *Archives of Disease in Childhood, 86,* 251-256.

Wortman, P. M. (1994). Judging research quality. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 97-109). New York: Russell Sage Foundation.

Table 1: Summary of MST research reviews' foci, methods, and conclusions

Author (Date)	Substantive Foci	Purpose or hypotheses	Review Methods	Conclusions Regarding Effects of MST
Aos et al. (2001)	Programs to reduce crime	"find programs that save more money than they cost" (p. 5).	Meta-analysis, cost benefit analysis	MST reduces crime (based on three studies, the average effect size for recidivism is -.31, SE=.1). Net direct costs of MST = $4,743 per participant; net benefits per participant (benefits minus costs) are $31,661 for taxpayers, $131,918 when benefits to crime victims are included; the latter represents a benefit-to-cost ratio of $28.33 for every dollar spent on MST.
Borduin (1999)	Adolescent criminality & violence	"discuss the emergent success" of MST (p. 242).	Narrative	MST "can successfully reduce criminal activity and violent offending in serious juvenile offenders. Of course, extensive validation and replication are needed for event the most promising treatment approaches" (p. 248).
Borduin et al. (2000)	Serious antisocial behavior in adolescents	"review some promising models of treatment" (p. 114)	Narrative	"Considerable evidence shows that MST can decrease rates of criminal activity and incarceration for serious juvenile offenders" (p. 130).
Borduin et al. (2003)	Serious antisocial behavior in adolescents	"address the empirical rationale for… MST, as well as the features… that make it well-suited for treating serious antisocial behavior" (p. 300)	Narrative	MST "can successfully reduce criminal activity and violent offending in serious juvenile offenders. Of course, extensive validation and replication are needed for event the most promising treatment approaches" (pp. 314-315).
Brestan & Eyberg (1998)	Treatments for conduct disorder	"identify empirically supported treatments" (p. 180)	Systematic search, narrative review	MST is "probably efficacious" (p. 185).
Brosnan & Carr (2000)	Adolescent conduct problems	Consider "well-designed studies which evaluate the effectiveness" of several interventions (p. 134).	Systematic search, tables, meta-analysis	MST "was effective in reducing family-based conduct problems and halving community-based recidivism rates. (MST) also improved family functioning" (p. 151).

Author (Date)	Substantive Foci	Purpose or hypotheses	Review Methods	Conclusions Regarding Effects of MST
Burns et al. (1999)	Treatment for mental disorders in children and adolescents	provide "an understanding of the evidence base (by reviewing) published literature for effective interventions" (p. 199)	Narrative	"Efficacy has been established in three randomized clinical trials of MST for delinquents. Each of these trials reported significant findings of behavior change, reduced contact with the justice system, and lower costs." (p. 220). "Multisystemic therapy has a well-established evidence base, including both efficacy and effectiveness studies" (p. 240).
Burns et al. (2000)	MST and wraparound services for youth w/ severe emotional disorders	"describe and contrast" MST and wraparound services (p. 284)	Narrative & tabular	"The evidence base for MST is characterized by considerable controlled research, but little diversity among investigators. The efficacy of MST…was established through three randomized clinical trials with delinquents, and effectiveness through the transfer of MST to other clinical populations ... and to multiple organizational settings …. The research base meets the criteria for a 'probably efficacious' treatment, but it was not classified as 'well-established'" (pp. 309-310).
Burns (2003)	Children's services	"provide a succinct summary of interventions… identify exemplary child initiatives… identify models for narrowing the gap between research and practice" (p. 956)	Narrative	MST results in "fewer arrests, fewer placements, decreased aggressive behavior" (p. 959).
Chorpita et al. (2002)	Treatments for disorders in childhood	"examine the efficacy and effectiveness of child treatments" for certain disorders (p. 165)	Narrative w/ some ES calculations (unspecified formulas)	"The effect size for MST was modest, suggesting that the average treated child scored better than 69% of children's scores before treatment. Also, the robustness of this treatment was rated as moderate, possibly due to the elaborate and highly orchestrated supervision network that appears to account for much of the success of the treatment. Consistent with this observation, no studies to date support MST other than those conducted by its developers. Nevertheless, the support for the effectiveness of MST is rather good, given that it has been tested with some of the most challenging youth and that it is one of the only treatments that has demonstrated superiority to realistic and commonly employed alternative treatments" (pp. 177-9).

Author (Date)	Substantive Foci	Purpose or hypotheses	Review Methods	Conclusions Regarding Effects of MST
Corcoran (2000)	Family interventions for child abuse and neglect	"critically review the research on family treatment for child physical abuse and neglect" (p. 563)	Narrative	"There has been strong empirical support for Multisystemic Therapy with juvenile offenders and their families, a population which has considerable overlap with children who have been neglected and abused" (p. 574).
Corcoran (2003)	Evidence-based family interventions	"familiarize the practitioner with evidence-based approaches for common problems for which families seek treatment" (p. 3)	Narrative & tabular	"The success of the multisystemic model seems to depend on fidelity to the treatment" (p. 181). The costs of MST "are offset by the costs saved in incarceration, institutionalization, and out-of-home placement" (p. 202).
Cormack & Carr (2000)	Treatment for drug abuse in children and adolescents	"review the outcomes of (3 groups of well-designed family-based intervevention studies for adolescent drug abusers) in a rigorous manner" (p. 161)	Systematic search, narrative & quantative analysis	Including studies of similar interventions by Liddle et al. (1995) and Scopetta et al. (1979), "multisystemic family therapy is more effective in the short-term than individual or group-based supportive counseling and parent education in treating adolescent drug abuse. However, it was no more effective than family therapy" (p. 175).
Curtis, Ronan & Borduin (2004)	MST	Examine "the effectiveness of MST" (p. 411)	Systematic search, meta-analysis	"As an empirically established treatment for violent and chronic juvenile offenders, MST appears to be worthy of wider implementation and continued evaluation…. More empirical support is required before MST can be considered an effective treatment of substance abuse in adolescents or an effective community-based alternative to the hospitalization of youths presenting psychiatric emergencies" (p. 417). Average effect size (across all outcome measures in 7 samples) d = .55 (not weighted by sample size or inverse variance).
Farrington & Welsh (2003)	Family-based prevention of offending	review "the effectiveness of family-based crime prevention programs" (p. 127)	Systematic search, meta-analysis	Mean ES for MST = .414 (95% CI = .281 to .548). "Three MST programs reduced delinquency or behaviour problems…while the other three did not… However, it should be noted that only two of the six MST evaluations had significant effects…. The large mean effect size for MST was largely driven by these two evaluations" (p. 143).

Author (Date)	Substantive Foci	Purpose or hypotheses	Review Methods	Conclusions Regarding Effects of MST
Fonagy & Kurtz (2002)	Conduct disturbance	MST "is arguably the most promising intervention for serious juvenile offenders" (p. 161).	Narrative	"MST is the most effective treatment for delinquent adolescents in reducing recidivism and ameliorating individual and family problems. It is substantially more effective than individual treatment, even for quite troubled and disorganized families" (p. 181). "Numerous other approaches have been tried but none of these are as effective as multisystemic therapy" (p. 181).
Fraser et al. (1997)	Family preservation services (including MST)	"review recent studies of family preservation and related family-strengthening programs, estimate the effect sizes…" (p. 138).	Narrative, tabular, & quantitative analysis	ES for MST range from .4 to .93 for prevention of rearrest, 1.01 for prevention of incarceration (p. 143).
Henggeler et al. (1998)	MST	Help people "make an informed judgement about a proven program's appropriateness for their local situation, needs, and available resources" (p. xii)	Narrative & tabular	Findings from 4 RCTs "provide strong evidence that MST can produce short- and long-term reductions in criminal behavior and out-of-home placements in serious juvenile offenders" (p. 38).
Henggeler, Schoenwald, Rowland, et al. (2002)	MST	Provide "a summary of the findings from research evaluations of MST and describe current replications of these findings and extensions of the model…" (p. 205).	Narrative & tabular	"Across studies, consistent clinical- and service-level outcomes have emerged. At the clinical level, in comparison with control groups, MST: improved family relations and functioning, increased school attendance, decreased adolescent psychiatric symptoms, decreased adolescent substance use, decreased long-term rates of rearrest ranging from 25-70%. At the service level and in comparison with control groups, MST has achieved: 97% and 98% rates of treatment completion in recent studies, decreased long-term rates of days in out-of-home placement ranging from 47% to 64%, higher consumer satisfaction, (and) considerable cost savings" (pp. 206-208).

Author (Date)	Substantive Foci	Purpose or hypotheses	Review Methods	Conclusions Regarding Effects of MST
Henggeler & Sheidow (2003)	Conduct disorder and delinquency	"review those family-based treatments of conduct disorder and delinquency that have been identified by federal entities and leading reviewers as efficacious…" (p. 505)	Narrative & tabular	In studies of juvenile offenders and delinquents "outcomes have consistently favored MST in comparison with control conditions. For example, MST treatment effects have included improved family relations and functioning, increased school attendance, decreased adolescent psychiatric symptoms, and reduced substance use. Reductions in rates of recidivism have ranged between 25% to 70% across studies for youth treated with MST compared to treated control groups. Moreover, MST has produced decreased rates of days in out-of-home placement ranging from 47% to 64% compared with usual services. Group differences have been observed as much as 5 years posttreatment" (p. 512). Successful MST outcomes have been observed for youths presenting psychiatric emergencies… and for children in maltreating families" (p. 512).
Hoagwood et al. (2001)	Child and adolescent mental health services	"review the status, strength, and quality of evidence-based practice in child and adolescent mental health services" (p. 1179)	Narrative	Results of MST trials "have been among the strongest found for children's services" (p. 1183). Results include lower rates of recidivism, out-of-home placements and arrest for juvenile offenders; reduced psychiatric hospitalization and improved functioning of youth and their families. Effects of MST "have been further demonstrated among juvenile sex offenders and abused or neglected children" (p. 1183)
Kazdin (1998)	Psychosocial treatments for conduct disorder in children	"reviews research for… psychosocial treatments that have shown considerable promise in the treatment of conduct disorder in children and adolescents" (p. 66).	Narrative	MST "has been shown to be superior in reducing delinquency and emotional and behavioral problems and improving family functioning in comparison to other methods of achieving these desirable goals" (p. 65). "On balance, MST is quite promising given the quality of evidence and consistency in the effects that have been produced…. The outcome studies have extended to youths with different types of problems (e.g., sexual offenses, drug use) and to parents who engage in physical abuse or neglect.... Thus, the model ... may have broad applicability across problem domains among seriously disturbed children" (p. 79).
Kazdin & Weisz (1998)	Treatments for child and adolescent internalizing, externalizing, and other disorders	"illustrate promising treatments" (p. 19)	Narrative	"MST is unique insofar as providing multiple replications across problems, therapists, and settings…. This shows that the treatment and methods of decision making can be extended and that the treatment effects are reliable…. Replications by others not involved with the original development of the program represent the next logical step. On balance, MST is quite promising given the quality of evidence and consistency of the outcomes" (p. 28).

Author (Date)	Substantive Foci	Purpose or hypotheses	Review Methods	Conclusions Regarding Effects of MST
Letourneau et al. (2002)	MST	(In a handbook of empirically supported interventions)	Narrative	"In comparison with control groups, and at a cost of approximately $5,000 per family, MST has consistently demonstrated improved family relations and family functioning, improved school attendance, and decreased adolescent drug use…. 25-70% decreases in long-term rates of rearrest, and 47-64% decreases in long-term rates of days in institutional placements" (p. 377).
Miller et al. (2000)	Marriage and family therapies	Present a complete summary of marriage and family therapy outcome research (p. 347)	Narrative	MST "has demonstrated its effectiveness in treating juvenile delinquency. Four outcome studies…show MST is more effective than standard treatments in reducing arrests, self-reported offenses, and jail time..." (p. 351). Two studies "found MST to be effective in treating substance abuse" (p. 352).
Pushak (2002)	Mental health services for children	MST "is an example of a model program with strong empirical support for effectiveness"	Narrative	"It would be safe to conclude that the total impact MST has on high-risk youth and their families and the decreased financial costs, to say nothing of the decreased psychosocial costs, of antisocial youth behavior to society is not yet matched by other psychotherapy programs."
Randall & Cunningham (2003)	Violence, substance abuse	Describe MST, a treatment "that has produced favorable outcomes…" (p. 1731)	Narrative	"MST has been extensively validated and cited as both an effective treatment for youth with violent behavior and as a promising adolescent substance abuse treatment…. MST can reduce violence and substance use of chronic juvenile offenders" (p. 1736)
Schoenwald, Brown et al. (2000)	MST	Highlight key features of MST …	Narrative	"MST has a strong track record in demonstrating favorable long-term outcomes for youth presenting serious clinical problems and their families" (p. 114).
Schoenwald & Rowland (2002)	MST	Purpose: "to facilitate implementation of evidence-based interventions in communities" (editors, p. 13). "MST is a well-validated treatment model"(p. 113)	Narrative	"The original studies of MST documented significant benefit for multiple target populations under conditions of training and close supervision by the MST developers" (p. 116).

Author (Date)	Substantive Foci	Purpose or hypotheses	Review Methods	Conclusions Regarding Effects of MST
Smith & Stern (1997)	Delinquency and antisocial behavior	"critical review of the current research on… the existing treatment outcome research" (p. 382)	Narrative	"In a series of controlled group studies, (MST) has shown consistent and strong results as an effective intervention for serious antisocial behavior and juvenile delinquency in both urban and rural areas and with families of different cultural backgrounds and socioeconomic status" (p. 405).
Sudderth (2000)	Treatment for substance-abusing youth	review of evaluations of treatment programs	Narrative	MST "has been found to be effective in reducing self-reported alcohol and marijuana use and decreasing the number of days juveniles spent incarcerated… Although MST is more expensive to implement than other approaches, initial results suggest that the long-term benefits of reduced residential placement and incarceration time are worth the investment" (p. 342).
Swenson & Henggeler (2003)	MST for maltreated children and their families	Description of MST	Narrative	Results from 8 RCTs "support the short-and long-term clinical effectiveness of MST as well as its potential to produce significant cost savings and capacity to retain families in treatment" (p. 75). One RCT in cases of child maltreatment showed that MST was more effective than Parent Training in improving parent-child relations, but Parent Training was more effective in decreasing social problems.
Swenson et al. (1998)	MST	"presents empirical support for use of an ecological approach with adolescent sexual offenders…"(p. 330)	Narrative	"Findings from several randomized trials have shown that … (MST) …. is an effective treatment for serious and complex problems presented by youths and their families…" (p. 332)
Tarolla et al. (2002)	Juvenile offenders	"provides an overview of available evidence… pertaining to treatment for juvenile offenders" (p. 125)	Narrative	"MST trials have shown reductions in long-term rates of violent offending, drug-related offending, and other delinquent and criminal activities. Also recent resaerch has documented MST's effectiveness with substance abusers, sex offenders, suicidal youth, maltreating families, and individuals with mental health problems" (p. 132).
US DHHS (1999)	Mental health	examine effectiveness of treatments	Narrative	"The efficacy of MST has been established in three randomized clinical trials for delinquents within the juvenile justice system… Initial results are promising for youth receiving MST instead of psychiatric hospitalizations… The efficacy of MST was demonstrated in real-world settings but only by one group of investigators..."

Author (Date)	Substantive Foci	Purpose or hypotheses	Review Methods	Conclusions Regarding Effects of MST
Vaughn & Howard (2004)	Adolescent substance abuse treatment	"assess outcome findings and methodological characteristics of controlled evaluations… (p. 325),	Systematic search, meta-analysis	MST has "evidence of clinically meaningful effect (ES > .20) [on adolescent substance abuse] with relatively strong designs and less than 1-year follow-up and no replication" and has been "shown to be effective in other studies with reducing adolescent violence and problem behavior (p. 334).
Woolfenden et al (2003)	Family and parenting interventions for conduct disorder and delinquency	"determine if family and parenting interventions improve (outcomes for children, parents, and families)"	Systematic search, meta-analysis	MST (and other interventions) "have beneficial effects in reducing the length of time spent by juvenile delinquents in institutions.... These interventions may also reduce rates of subsequent arrest.... At present there is insufficient evidence that family and parenting interventions reduce the risk of (incarceration) or have a beneficial effect on parenting, parental mental health, family functioning, academic performance, future employment, and peer relations" (p. 7). There is no evidence that interventions such as MST cause harm.

Table 2: Characteristics of research reviews (by publication year)

Pub. Year	Authors	Inde-pend-ent	Explicit inclusion criteria	System-atic search strategy	Include un-published reports	N MST study reports cited [a]	Undup. N of MST studies cited	Assess allocation method [b]	Assess study attrition	Rate study quality	Use ITT analysis	Report study-level ES	Report pooled ES
1997	Fraser et al.	✓			✓	4	2					✓	
	Smith & Stern	✓				5	4						
1998	Brestan & Eyberg	✓	✓	✓		3	3	✓		✓			
	Henggeler et al.				✓	25	21						
	Kazdin	✓				6	5						
	Kazdin & Weisz	✓				6	5						
	Swenson et al.					10	8						
1999	Borduin				✓	9	6						
	Burns et al. [c]	✓				6	6						
	US DHHS	✓				5	5						
2000	Borduin et al.					8	5						
	Brosnan & Carr	✓	✓	✓		7	5	✓	✓	✓		✓	✓
	Burns et al. [c]				✓	16	12						
	Corcoran	✓				7	5						
	Cormack & Carr	✓	✓	✓		1	2	✓	✓	✓		✓	
	Miller et al.	✓				5	4						
	Schoenwald et al.				✓	12	8						
	Sudderth	✓				3	3						
2001	Aos et al.	✓			✓	7	6	✓	[d]	✓	[d]	✓	✓
	Hoagwood et al.					6	5						
2002	Chorpita et al.	✓		✓		3	3						✓
	Fonagy & Kurtz	✓	✓	✓		9	6						
	Henggeler et al.				✓	29	25						
	Letourneau et al.					12	8						
	Pushak	✓				5	5						
	Schoenwald & Rowland					12	8						
	Tarolla et al.	✓				4	3						
2003	Borduin et al.				✓	13	7						
	Burns [c]	✓				2	1						
	Corcoran	✓				8	5						
	Farrington & Welsh	✓	✓	✓	✓	6	6	✓		✓		✓	✓
	Henggeler & Sheidow					14	8						
	Randall & Cunningham					13	6						
	Swenson & Henggeler					12	8						
	Woolfenden et al.	✓	✓	✓		3	3	✓	✓	✓		✓	✓

Pub. Year	Authors	Inde-pend-ent	Explicit inclusion criteria	System-atic search strategy	Include un-published reports	N MST study reports cited [a]	Undup. N of MST studies cited	Assess allocation method [b]	Assess study attrition	Rate study quality	Use ITT analysis	Report study-level ES	Report pooled ES
2004	Curtis et al.		✓	✓		11	7	✓				✓	✓
	Vaughn & Howard	✓	✓	✓		3	3		✓	✓		✓	
	Total (N=37)	22	8	9	9			7	4	7	0	8	6

[a] Includes relevant unpublished reports and personal communication.

[b] Includes reviews limited to RCTs.

[c] Burns collaborated with a MST developer in a review published in 2000, but not in the reviews she published in 1999 and 2003.

[d] Contrary to reviewers' intentions, studies that did not report data on drop-outs were treated as if they had no drop-outs.

Table 3: Results of Brunk, Henggeler, & Whelan (1987)

Measure	Domain(s)	Abbreviation	Main effects MST vs PT
Self-reports			
Symptom Checklist-90 (Global Severity Index)	Parent psychiatric symptoms	SCL-90 GSI	NS
Behavior Problem Checklist (total score)	Parent perceptions of child behavior problems	BPC	Not reported
Family Environment Scale (90 items, 10 subscales)	Relationships, personal growth, system maintenance	FES	NS
			NS
			NS
			NS
			NS
			NS
			NS
			NS
			NS
			NS
Family Inventory of Life Events (71 items)	Parental stress	FILE	NS
Treatment Outcome Questionnaire	Parent perceptions of treatment needs and changes in needs	TOQ I-C	NS
		TOQ F-C	NS
		TOQ SS-C	PT > MST
Therapist reports			
Treatment Outcome Questionnaire	Therapist perceptions of treatment needs and changes in needs	TOQ I-T	NS
		TOQ F-T	NS
		TOQ SS-T	NS
Observational measures of parent-child interactions			
Parental effectiveness	Attention	NO-VAT-O	NS (Ng: MST > PT, Ab: PT > MST)
		CT-NAT-O	MST > PT
		NO-NAT-O	NS
	Action	O-VAC-TC	PT > MST
		CT-VAC-TC	MST > PT
		CT-NAC-TC	MST > PT
Child passive noncompliance		O-VAC-O	MST > PT
		CT-VAC-CT	NS (Ab: MST > PT)
Parental unresponsiveness		O-VAT-O	MST > PT
		O-NAT-O	NS
		O-NAT-TC	NS

NS = no significant difference, > = superior, Ng = neglect group only, Ab = abuse group only.
For TOQ: I = individual, F = family, SS = social system, C = client report, T = therapist report.
For observational measures: NO = not oriented, O = oriented, VAT = verbal attention, NAT = nonverbal attention, VAC = verbal action, NAC = nonverbal action, CT = contact, TC = task completed.

Table 4: Summary of results of Brunk, Henggeler & Whelan (1987) as described in the original published report and thirteen published reviews

Source	Type of information	Number of items	Distribution of Results		
			Favors PT	Neutral (no sig. diff. between groups, unclear, or missing)	Favors MST
Original research report					
Brunk et al. (1987)	Data collected (subscales)	30	2	23	5
	Results reported (subscales)	29	2	22	5
	Data provided (subscales)	19	2	12	5
	Abstract (phrases)	5	1	3	1
Research reviews					
Burns et al. (2000)	Text (phrases)	5		3	2
	Table (phrases)	1			1
Corcoran (2000)	Text (phrases)	10	3	4	3
Curtis et al. (2004)	Table (effect size)	1			1
Henggeler et al. (1998)	Text (phrases)	1			1
Henggeler et al. (2002)	Text (phrases)	3			3
	Table (phrases)	1			1
Henggeler & Sheidow (2003)	Text (phrases)	1			1
Hoagwood et al. (2001)	Text (phrases)	1			1
Kazdin (1998)	Text (phrases)	1			1
Kazdin & Weisz (1998)	Text (phrases)	1			1
Pushak (2002)	Text (phrases)	1			1
Schoenwald & Rowland (2002)	Text (phrases)	1			1
Swenson & Henggeler (2003)	Text (phrases)	5	1		4
Swenson et al. (1998)	Text (phrases)	1			1