Benchmarking the effectiveness of community services for youth with anxiety disorders.

Carolyn Houlding

Department of Psychology

Lakehead University, Thunder Bay, Ontario

Submitted April 2013

Supervisor: Fred Schmidt, PhD,
Co-supervisor: John Jamieson, PhD
Internal examiners: Charles Netley, PhD; Michel Bédard, PhD
External Examiner: John Hunsley, PhD

Author Note

A dissertation submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Clinical Psychology

Dedication

To my children and husband.

"Times are bad. Children no longer obey their parents, and everyone is writing a book."
Marcus Tullius Cicero 106 – 43 BC

Acknowledgments

Thanks to the youth, families, therapists and researchers involved in the published trials and national evaluation study. Thanks for giving your time and information to help others find a better way. Thanks to my supervisor, Dr. Fred Schmidt who has walked with me this long journey and who has been patient, available, knowledgeable and insightful. Thanks to my co-supervisor Dr. John Jamieson who allowed independence and provided encouragement and timely, astute advice. Thanks to Bruce Weaver and Dr. Takyuma Minami for patience, statistics consultation and late night emails; truly, part of life's blessings. Thanks also to Dr. Michael Borenstein, for advice regarding statistical analysis. I greatly appreciate the time my external examiners Dr. Michel Bédard and Dr. John Hunsley invested in the review of the dissertation, as well as their challenging, thoughtful comments. I also appreciate the input of my colleagues in the Department of Psychology at Lakehead University, including Dr. Charles Netley and Dr. Michael Stones, who were always available for hallway musings. Thanks to Dr. Russell Carleton for assistance with the national evaluation database, Liane Kandler for assistance with coding and Barb for assistance with data input. Thanks to fellow (present and past) graduate students Dr. David Armstrong and Derick Cyr for conversations that were catalysts for new perspective as well as ongoing support and friendship.

To Nick and Carrie, Maria and Peder, Anne and Bryan and other good friends who helped plug the gaps. Thanks to my parents, brother (Andrew) and sister (Bronwyn), Lois and Audrey and extended family for sustained (and sustaining) cheerleading. Thanks to my husband for standing by my side, providing back up and support. This process has been a team effort. Lastly, thanks to my awesome children (Elizabeth, David, Abbie) who never complained and continued to obey (though someone was writing a book).

Abstract

The present study applied a benchmarking strategy to evaluate the outcomes of youth (6 - 15 years) with anxiety disorders treated at 'Systems of Care' children's mental health services (SOC CMHS). There were three stages of analysis. The first used meta-analytic technique to aggregate results of 17 randomised controlled trials of treatments of anxiety in youth. From these studies, benchmarks were established for two different outcome criteria: pre-post effect sizes and the proportion of youth evidencing 'clinically significant improvement'. Two subsets from the SOC CMHS data were considered. The first was comprised of youth who were selected on the basis of a combination of Child Behavior Checklist profile and DSM diagnosis or presenting problem. The second was comprised of youth selected primarily on the basis of clinician-generated DSM diagnosis. Neither subset attained levels of improvement commensurate with treatment efficacy benchmarks. Only one subset (selected partly on the basis of Child Behavior Checklist profile) achieved results significantly better than natural remission and this was only for one natural history benchmark (pre-post effect size). The third stage of analysis examined factors associated with reliable improvement and treatment response. Results indicated that the relatively poor response of youth from the SOC CMHS agencies could not be explained by the socio-psychological characteristics of this group. Avenues for future research are suggested, including extension of benchmarking strategies in children's mental health and improved understanding of predictors of treatment response.

Table of Contents

Dedication i

Acknowledgments ii

Abstract iii

Table of Contents iv

Introduction 1

Empirically Supported Treatments and Evidence-Based Practice 2

Case Complexity 7

Treatment Implementation, Supervision, Training and Therapist Characteristics. 12

Summary 16

Benchmarking 17

Comprehensive Community Mental Health Services for Children and Their Families Program Systems of Care (SOC CMHS) 33

Study Aims 36

Hypotheses 37

Method 40

Base Longitudinal SOC CMHS Database 40

SOC CMHS Subsets 43

Measures 54

Data Collection 61

Ethics Clearance 62

Selection of Clinical Trials for Treating Anxiety Disorders 63

Identification of Clinical Trials for Inclusion 65

Data Analysis 75

Pre-Analysis Preparations 84
Hypothesis Testing 104
Results 108
Overview of Results Section 108
Hypothesis 1: Evaluating the SOC CMHS Data Against Pre-Post Effect Size Benchmarks 109
Hypothesis 2: Evaluating SOC CMHS data against 'clinically significant improvement' benchmarks. 112
Hypotheses 1 and 2: Evaluating SOC CMHS 'completer' data against benchmarks. 112
Hypotheses 3, 4 and 5: Factors Associated with Treatment Response 114
Moderators of Effect Size Within Clinical Trials 116
Discussion 117
Treatment Efficacy and Natural History Benchmarks for Pre-Post Effect Sizes 118
Data Reduction 124
Hypothesis 1: Comparison of SOC CMHS Subsets with Treatment Efficacy and Natural History Benchmarks. 126
Hypothesis 2: Comparison of SOC CMHS subsets and 'Clinically Significant Improvement' Benchmarks. 128
Hypothesis 3, 4 and 5: Factors Associated with Treatment Response 132
Implications of Results 134
Limitations of Study 137
Implications for Future Research 140
Summary and Conclusions 142

References 146

Appendix A 199

List of Tables

Table 1 Profile of youth from base longitudinal SOC CMHS and SOC CMHS subsets…….. 52

Table 2 Demographic, treatment and recruitment characteristics of clinical trials……..... 70

Table 3 Study treatment effect size estimates, treatment efficacy benchmark and effect size estimates for SOC CMHS subsets ……..... 88

Table 4 Study wait list control effect size estimates, natural history benchmark and effect size estimates for SOC CMHS subsets……... 95

Table 5 Benchmark and SOC CMHS subset effect size estimates…….. 99

Table 6 'Clinically Significant Improvement' (CSI) across treatment groups of clinical trials and 'Clinically Significant Improvement' treatment benchmark ……..……... 102

Table 7 'Clinically Significant Improvement' (CSI) for wait list control groups of clinical trials and 'Clinically Significant Improvement' natural history benchmark…….. 103

Table 8 'Clinically Significant Improvement' in benchmarks and SOC CMHS subsets……... 104

Table 9 Effect size estimates of SOC CMHS subsets tested against critical values of treatment efficacy and natural history benchmarks …… 111

List of Figures

Figure 1 Data reduction from base longitudinal age matched SOC CMHS to SOC CMHS subsets……..... 51

Figure 2 Study and summary effect size estimate(s) for treatment groups of clinical trials with Shortt et al., (2011) excluded…… 91

Figure 3 Funnel plot of standard error by effect size estimates for treatment groups of clinical trials, with effect size estimates of 'missing' studies imputed …… 93

Figure 4 Study and summary effect size estimate(s) for wait list control groups of clinical trials …… 96

Figure 5 Treatment efficacy benchmark, natural history benchmark and effect size estimates for SOC_{CBCL} and SOC_{diag} subsets…….. 100

Introduction

It is estimated that as many as 2.5-5% of youth meet criteria for anxiety disorders (Breton et al., 1999; Costello, Mustillo, Erkanli, Keeler, & Angold, 2003; Ford, Goodman, & Meltzer, 2003; Lewinsohn, Hops, Roberts, Seeley, & Andrews, 1993; Lewinsohn, Zinbarg, Seeley, Lewinsohn, & Sack, 1997), and reviews suggest they are some of the most common mental health problems in children and adolescents (Craske, 1997). A large proportion of children who are diagnosed with one anxiety disorder can be diagnosed with an additional anxiety disorder (40-60%) (Silverman & Ollendick, 2005). Further, anxious children are more likely to be diagnosed with depression (Angold, Costello, & Erkanli, 1999; Brady & Kendall, 1992, Costello et al., 2003; Ford et al., 2003). Youth who suffer from anxiety disorders are at risk of a constellation of other difficulties including academic and interpersonal problems (Essau, Conradt, & Petermann, 2000; Kubik, Lytle, Birnbaum, Murray & Perry, 2003; Strauss, Frame, & Forehand, 1987) with the disorders having a moderate to high impact on functioning (Ezpeleta, Keeler, Alaatin, Costello, & Angold, 2001).

Not only do anxiety disorders have concerning implications for the short-term, a child who develops these disorders at a young age can struggle with them into adolescence (Bittner et al., 2007) and adulthood (Achenbach, Howel, McConaughy, & Stranger, 1995; Caspi, Elder, & Bem, 1988; Pine, Cohen, Gurley, Brook, & Ma, 1998; Gregory et al., 2007). Anxiety has been described as a 'gateway' disorder (Kendall, Seppatini, & Cummings, 2012), with symptoms of anxiety in childhood placing the child at increased risk for depression (Costello et al., 2003; Pine et al., 1998; Roza, Hofstra, van der Ende, &

Verhulst, 2003), substance misuse (Kendall, Safford, Flannery-Schroeder, & Webb, 2004) and externalizing disorders (Bittner et al., 2007; Costello et al., 2003) in adulthood.

Thus, addressing this group of disorders in children is important for at least two major reasons. First, as mentioned, they are among the most common mental health concerns in youth and thus use of effective interventions may have considerable impact on the efficiency and helpfulness of treatment agencies (Hodges & Wotring, 2004). Second, the course of these disorders can be chronic and (as mentioned) place the youth at increased risk for poor outcomes in adolescence and adulthood. Intervening when clients are children may prevent the development of other problems, including impairment secondary to the primary disorder.

Empirically Supported Treatments and Evidence-Based Practice

Given the seriousness and pervasiveness of anxiety and other co-morbid mental health concerns in children, the question arises as to how these disorders might best be treated. Research suggests that outcomes of youth receiving the treatment usually provided in community mental health settings ('usual care'; UC) are disappointing when compared to therapies with established empirical support (Kazdin, Esveldt-Dawson, French, & Unis, 1987; Mufson et al., 2004; Taylor, Schmidt, Pepler, & Hodgins, 1998; Weiss, Catron, & Harris, 2000; Weisz, Weiss, & Donenberg, 1992). Interventions delivered in UC tend to be eclectic and reflect preferences of the therapists delivering them rather than interventions informed by research. Some results suggest that, consistent with the adult literature (Addis et al., 2004; Linehan, Armstrong, Suarez, Allmon, & Heard, 1991), outcomes for youth receiving UC may be no better than natural remission, i.e. the passage of time alone (Weersing & Weisz, 2002) or may be quite limited (Ollendick & King, 2004). Outcomes of

clients receiving non-specific therapy are generally lower than those found with research supported treatments (Beidel, Turner, & Morris, 2000; Hudson et al., 2009; Muris, Meesters, & Gobel, 2002) and can improve when therapists change from UC to research supported treatments (Cukrowicz et al., 2005). In fact, in contrast to interventions with empirical support, meta-analyses (including studies of predominantly externalising disorders), show that the average effect size of community treatment is modest and may be near to zero (Weiss et al., 1999; Weisz, Jensen-Doss, & Hawley, 2006).

Research-based interventions, including more recent studies of treatment of anxiety, do not always outperform nonspecific therapy or UC, however (Barrington, Prior, Richardson, & Allen, 2005; Kendall, Hudson, Gosch, Flannery-Schroeder, & Suveg, 2008; Last, Hansen, & Franco, 1998; Silverman et al., 1999b; Southam-Gerow et al., 2010; Weisz et al., 2009; Weisz et al., 2012). This lack of difference in outcomes may be because these comparisons are typically under-powered (c.f. Kazdin & Bass, 1989). Alternatively, it may be because efforts at disseminating research-supported treatments mean that UC is increasingly likely to incorporate elements of these research-based treatments and thus be more effective (Kendall et al., 2008; Southam-Gerow et al., 2010). Further, while youth receiving UC may achieve comparable outcomes to those receiving research-supported treatments, they may take significantly longer to reach these outcomes, or make use of more resources to do so (Southam-Gerow et al., 2010).

Interest in identifying and evaluating psychological therapies supported by research evidence first crystallised in the 1990s with the seminal work of the Task Force on Promotion and Dissemination of Psychological Procedures (1995). The work of the task force coincided with, and was complemented by, the broader ‘evidence-based’ movement,

which included interest in identifying and classifying the amount of empirical support for particular treatments for youth and other populations. The Task Force developed a system for classifying treatments according to the quality and quantity of empirical support for their efficacy and this system (among others) remains in use today. Clinical practices (such as treatments) were classified within four levels, ranging from 'well established' to 'experimental'. For instance, 'well established' treatments are those interventions evaluated using randomised controlled designs, conducted by investigators independent of the treatment developers, that yield outcomes superior to placebo or alternative treatments (Chambless & Hollon, 1998). 'Empirically supported treatment' (EST) is a term applied to describe interventions with empirical support for their efficacy.

Recent reviews of treatments to address anxiety disorders in youth have identified interventions that have a strong evidence base for their use (Chorpita, Daleiden et al., 2011; In-Albon & Schneider, 2006; Reynolds, Wilson, Austin & Hooper, 2012; Silverman, Ortiz et al., 2008; Silverman, Pina, & Viswesvaran, 2008). Most of these interventions are based on cognitive behavioural models of intervention and are delivered in group or individual format. Many (for instance, 'Coping Koala'; Barrett, 1995), are based on one of the earliest manualised therapies for treatment of anxiety in children 'Coping Cat' (Kendall, 1994), a sixteen session individual therapy typically delivered to 7- 13 year olds.

Within the Task Force's classification system, research findings from studies with greatest internal validity – that is, randomised controlled trials (RCTs) - are given the most weight. Privileging results of RCTs over other research designs has generated controversy, in part related to the degree to which results of these treatment trials can be replicated in 'real world' settings. This is relevant to the present study, which will involve evaluating

outcomes of community agencies treating youth with anxiety disorders against results of published treatment trials. For this comparison to be valid, it should first be established that ESTs for youth with anxiety disorders are effective in community settings. Addressing this issue requires an understanding of two of the main types of trials used to evaluate clinical interventions.

Most interventions are initially evaluated in the context of 'efficacy' trials. Prototypically, 'efficacy' studies use designs where conditions are optimal for success and there are tight controls on many aspects of treatment delivery. In practice, this means that efficacy trials are conducted in research settings (rather than in agencies where the prime task is to deliver healthcare), have rigorous inclusion and exclusion criteria to ensure all participants have the disorder of interest, and use samples who are often recruited specifically for the study (rather than through clinical referral) (Nathan & Gorman, 2002; Weisz, Doss, & Hawley, 2005). Further, the intervention sites often specialise in the particular disorder, tend to use research therapists (often graduate students) supervised by programme developers, are more likely to train therapists to a prescribed level of competency before the treatments begin, and then employ ongoing supervision and monitoring of intervention delivery (Hunsley, 2007). The size of clinician caseloads would typically be smaller and less diverse than those of clinicians in community agencies (Spielmans, Gatlins, & McFall, 2010). This level of client homogeneity (Borkovec & Costonguay, 1998; Chambless & Hollon, 1998), therapist training and expertise within the organisational context that treatment is delivered is assumed to maximise the likelihood of treatment success, and may compromise generalisability of results from efficacy trials.

The conditions of 'efficacy' studies are often contrasted with those of 'effectiveness'

studies. Prototypically, 'effectiveness' studies are designed to address whether interventions are effective in the 'real world'. That is, they are designed to maximize external validity while maintaining adequate internal validity (Hunsley & Lee, 2007). Effectiveness studies are thus generally conducted in community settings, and typically make use of therapists who work in these settings (rather than those having or being supervised by those with specialty expertise). Samples may be more heterogeneous and more likely to access the service through referral rather than in response to advertising. The degree to which a study is characterised as an 'efficacy' rather than an 'effectiveness' trial lays on a continuum that emphasizes internal validity on one end and external on the other (Depp & Lebowitz, 2007).

Hunsley and Lee (2007) conducted a review comparing results of effectiveness trials against efficacy for internalizing and externalizing disorders in adults and youth. Results of the review were telling. Outcomes and participant retention in effectiveness and efficacy trials were comparable. Further, they found that the levels of therapist supervision and exclusion criteria in the studies were comparable to conditions in regular community agencies. The small number of available studies of treatments for youth meant that they included studies where some recruiting was conducted through advertising (rather than referral). Nonetheless, the review suggested that implementation of ESTs in the community (i.e. effectiveness trials) can achieve similar outcomes to 'ideal' conditions (i.e. efficacy trials). With respect to the treatment of anxiety disorders in youth, they found that most effectiveness trials had results comparable or superior to the efficacy benchmark trial.

Thus, evidence indicates that when community agencies implement empirically supported treatments, they obtain similar outcomes to those within treatment efficacy trials.

It is possible that differences in conditions at community agencies and those within efficacy trials might moderate outcomes in community settings. Research regarding potential moderators of outcome in treatment will therefore be appraised. This review will commence with consideration of the complexity of cases within efficacy trials as well as the potential impact of complexity on outcome.

Case Complexity

Case complexity encompasses factors such as the severity of the youth's psychopathology, comorbidity in clinical presentation, and adversity in socio-demographic context (including poverty and parental psychopathology). Westen, Novotny and Thompson-Brenner (2004) articulate the common concern that typical exclusion criteria mean that efficacy trials often include clients with less complex clinical profiles. The implication of this concern is that clients within efficacy trials are easier to treat than clients in the 'real world'. This concern is problematic, given the high rates of comorbidity found in clinical populations, including youth with anxiety disorders seen at treatment facilities (Southam-Gerow, Weisz, & Kendall, 2003). Reviews of efficacy and effectiveness studies for a variety of populations and presenting problems suggest beliefs regarding selectiveness in recruiting clinical trial populations are not necessarily well founded. For instance, some studies suggest that not only are clients in clinical trials comparable to community samples in terms of complexity and severity, they may actually have more severe symptoms than clients treated in the community (Carroll, Nich, McLellan, McKay, & Rounsaville, 1999; Jacobson & Christensen, 1996; Oei & Boschen, 2009).

Hunsley and Lee's (2007) study included an examination of exclusion criteria within

efficacy trials in youth and adults. They found most efficacy trials excluded participants (1) with conditions that were more urgent or took precedence for treatment (e.g. substance abuse) (2) who were participating in concurrent psychotherapy and (3) with mental retardation. They concluded that these common criteria seemed clinically appropriate and therefore would not affect the generalizability of results of efficacy trials to community settings. Consistent with Hunsley and Lee (2007), authors of a recent meta-analysis of treatment of mental health disorders in youth (including anxiety disorders), concluded that inclusion and exclusion criteria for RCTs appeared clinically sound; clients were not excluded due to comorbidity unless it would impact treatment plans (i.e. unless the additional diagnosis would alter treatment priority or be a substantial moderator of outcome – for instance, a developmental disability) (Chorpita, Daleiden et al., 2011). Further, these conclusions are consistent with Weisz, Hawley and Doss (2004) whose review of treatment for youth with anxiety disorders, depressive disorders, Attention Deficit Hyperactivity Disorder (ADHD), and Conduct Disorder (CD) revealed that almost half of treatment studies had no exclusion criteria related to comorbidity. Stirman, DeRubeis, Crits-Christoph and Brody (2003), found that 80% of adult clients within their treatment agency data-base would be eligible for inclusion in published randomised clinical trials, also suggesting that community populations are not substantially different or more difficult to treat than samples within efficacy trials. Altogether, these studies suggest that RCTs usually do not systematically exclude individuals with more complex clinical presentations. It is possible, however, that while complex individuals may not be explicitly excluded, those recruited for involvement in clinical trials tend to have a less complex profile than is typical of clients in community mental health settings.

When examining the characteristics of youth with anxiety disorders in a community clinic, Southam-Gerow et al. (2003) found that there was a higher proportion of youth from single parent families and with comorbid disruptive disorders than is typical of participants of clinical trials. This finding was consistent with Baker-Ericzen, Hurlburt, Brookman-Frazee, Jenkins and Hough (2010) who examined the clinical and social profile of parents from a large community mental health data-base and compared them to the samples within ESTs for parent training. They found that the demographic and symptom profile of youth in the community data set and in the RCT samples were similar, although family and child contextual characteristics were more complex in the community data set.

The forgoing review suggests that samples within clinical trials do not systematically exclude complex clients and that clients in community settings often have symptoms or presentations that are no more severe than those in clinical trials and are sometimes less so. There is some evidence that community samples include individuals with more complex profiles than samples from efficacy trials - for instance, more comorbid disorders or more adverse social contexts.

Research has also examined whether these and other variables influence treatment response. Compton et al. (2004) reviewed treatments for internalising disorders in youth and noted that the most frequent finding of studies that look at these socio-demographic factors is that they are not related to outcome. This conclusion is consistent with a more recent meta-analysis for treatment of anxiety disorders in youth, which also found no significant relationship between age, gender, ethnicity and outcomes (Silverman, Pina et al., 2008). With some exceptions (Barrett et al., 1996; Ginsburg et al., 2011; Pina, Silverman, Fuentis, Kurtines, & Weems, 2003; Reynolds et al., 2012) numbers of studies

of treatments have found that age (Alfano, 2012; Kendall, 1994; Kendall, Brady, & Verduin, 2001; Rapee, Schniering, & Hudson, 2006; Treadwell, Flannery-Schroeder, & Kendall, 1995), gender (Kendall, 1994; Kendall et al., 2001; Southam-Gerow, Kendall, & Weersing, 2001; Rapee, Abbott, & Lyneham, 2006; Treadwell et al., 1995) and ethnicity (Kendall, 1994; Kendall et al., 2001; Southam-Gerow et al., 2001; Treadwell et al., 1995) are not significantly related to outcome in treatment of children with anxiety disorders.

Comorbidity appears to have an inconsistent relationship to outcomes. A number of studies have found no significant relationship between comorbid conditions and outcomes (Flannery-Shroeder & Kendall, 2005; Kendall et al., 2001; Öst, Reuterskiöld & Costa, 2010; Rapee, 2003; Rapee et al., 2006) whereas others have found youth with depressive (Berman, Weems, Silverman, & Kurtines, 2000; Southam-Gerow, et al., 2001, Storch et al., 2008), externalising (Storch et al., 2008) or non-anxiety comorbidity (Liber, Widenfelt, Leeden, Goedhart, Utens, & Treffers, 2010) had worse treatment outcomes than those without. Doss and Weisz (2006) found that comorbidity accounted for very little (1%) variance in outcome of youth in treatment trials and Ollendick, Jarrett, Grills-Taquechel, Hovey, & Wolff (2008) concluded that most studies do not find a significant relationship between comorbidity and outcome. Overall, while comorbidity may complicate the delivery of ESTs (Southam-Gerow et al., 2003; Southam-Gerow, Chorpita, Miller, & Gleacher, 2008), these mixed findings are consistent with Hunsley's (2007) conclusion that comorbidity does not have a consistently negative effect on outcomes of therapy.

The impact of initial symptom severity on outcome has also been examined. Some research has found that adults with more severe symptoms at baseline are likely to improve more than those with less severe symptoms (Garfield, 1986; Oei & Boschen, 2009;

Minami, Wampold, Serlin, Kircher & Brown, 2007), and recent studies of children with anxiety have also found that youth with more severe initial symptoms improved more than others (Kley, Heinrichs, Bender, & Tuschen-Caffien, 2012; Liber et al., 2008). Consistent with Oei and Boschen (2009), Liber, et al. (2008) found that while more severe pre-treatment symptoms were associated with greater improvement, fewer of these youth recovered (i.e. lost their diagnosis) by the end of treatment. Southam-Gerow et al. (2001) also found that more severe symptoms at baseline were associated with lower rates of 'recovery' from anxiety diagnosis. Altogether, the most consistent findings appear to be that individuals with more severe initial symptoms improve more than those with less severe symptoms at pre-treatment baseline. However, they recover (i.e. lose their diagnosis) less often. This may be because higher scores are associated with greater regression to the mean (and hence greater improvement); but that a larger magnitude of improvement is required before reaching 'normal' range of functioning.

Other research has also considered the ability of various indices of social disadvantage to predict outcome. Weisz, Donenberg, Han and Weiss (1995) found that studies including cases that were rated (by observers) as more complex had lower effect sizes than studies with cases that where rated as more straightforward. Two meta-analyses examining predictors of outcome in parenting programmes (Dumas & Wahler, 1983; Reyno & McGrath, 2006) found social disadvantage was associated with negative treatment outcomes. Gardner et al. (2009) found a limited number of family risk factors predicted outcome in a parenting intervention. In their study, contrary to expectations, low maternal education predicted greater improvement in child problem behavior, although being a single parent predicted less improvement on some outcome measures. Authors

speculated that the counter-intuitive improvements related to extra efforts to engage parents from deprived backgrounds. In studies of treatment response for children with anxiety, Southam-Gerow et al. (2001) found income and family composition (single versus dual parent household) were not significantly related to outcomes. There have been some mixed findings regarding the impact of parent psychopathology on outcomes for youth with mental health needs (Gardner et al., 2009), although in general, parent pathology (particularly maternal depression) seems related to negative outcomes (Berman et al., 2000; Crawford and Manassis, 2001; Liber et al., 2010; Southam - Gerow et al., 2001). Legerstee et al. (2008) found that maternal anxiety actually predicted better outcomes in adolescents (but not children) with anxiety and Liber et al. (2010) found maternal warmth predicted worse outcomes. Cobham, Dadds and Spence (1998) found that children whose parents had anxiety disorders improved more if the parents participated in anxiety management treatment themselves.

Treatment Implementation, Supervision, Training and Therapist Characteristics.

Other factors may affect the generalizability of results of clinical trials to the community. Efficacy trials make use of interventions that are of a standard length (generally 10-16 sessions for cognitive behavioural therapy protocols). This standardisation of length compares to the conditions within community agencies (UC) (Hurlbert, Garland, Nguyen, & Brookman-Frazee, 2010).

RCTs typically include monitoring of adherence to the content of a treatment manual within their treatment protocol. In a meta-analysis comparing outcomes from lab and clinic based interventions, Weisz et al. (1995) found that monitoring treatment fidelity (i.e. adherence to the intervention) was associated with better outcomes in treatment of youth.

Research in other populations found weak adherence to procedures was associated with worse outcomes (Schoenwald & Hoagwood, 2001). Weisz et al. (2012) found that treatments making use of manuals outperformed UC, although this superior performance was not clearly related to manual use alone. There is little data regarding the prevalence of treatment manual use or adherence monitoring within UC, although the cost or perceived compromises to therapist autonomy (among other factors) might mean that they are not widely or consistently used (c.f. Zayas, Drake, & Johnson-Reid, 2011).

Training is another key consideration that influences the generalizability of results of ESTs to community agencies (Han & Weiss, 2005; Lochman et al., 2009; Merrill, Tolbert, & Wade, 2003). Higher 'doses' of training might be associated with more competent delivery of treatments (Lochman et al., 2009). ESTs tend to involve more intense therapist training than 'training as usual', which typically includes provision of a treatment manual, a one or two day workshop and little or no follow up (Beidas, Barmish, & Kendall, 2009; Beidas & Kendall, 2010; Weisz, Doss et al., 2005; Weisz, Sandler, Durlak & Anton, 2005). This model of training may be insufficient to impact clinician behaviour and thereby client outcome (c.f. Grimshaw et al., 2001; Herschell, 2010; Lochman et al., 2009; Sholomskas, et al., 2005). In contrast, Beidas and Kendall (2010) found that training that incorporated active learning strategies such as coaching and feedback was associated with improved therapist competence and client functioning.

Supervision can be conceptualised as part of the training process. The nature of supervision in clinical trials is likely to be quite different in content and impact than in community agencies (Accurso, Taylor, & Garland, 2011). Supervision within UC may include a large proportion of time spent in attending to administrative issues with relatively

little time or intensity invested in practising or improving therapist competency in core components of ESTs (Garland, Plemmons, & Koontz, 2006). In contrast, some studies have found that supervision that includes review of video tape and other forms of adherence to components of ESTs can enhance therapist competence and ultimately client outcomes (Callahan, Almstrom, Swift, Borja, & Heath, 2009; Ng, 2005), including in youth therapy (Schoenwald, Sheidow, & Chapman, 2009; Schoenwald, Sheidow, & Letourneau, 2004). For instance, the most substantial improvement in competency in medical practice was found where supervision consisted of review of actual practice (Kilminster & Jolly, 2000).

These results suggest optimal training involves both didactic and competency-based components. Sufficient quality training and supervision is required to implement ESTs with competence, and training and supervision delivered within UC do not usually reflect that typically utilised in efficacy trials. Presumably, results of efficacy trials cannot be validly generalized to community settings if they require unrealistic demands in terms of training or supervision. So, the question remains whether implementing training and supervision to this level of competence is feasible within community settings.

The length of supervision in EST trials and community practices appear comparable (approximately one hour every week or two) (Accurso et al., 2011; Schoenwald et al., 2008), including within RCTs for treatment of youth with anxiety disorders (Bögels & Sigueland, 2006; King, Heyne, & Ollendick, 2005; Muris, Merckelbach, Holdrinet, & Sijsenaar, 1998; Nauta, Scholing, Emmelkamp, & Minderaa, 2003; Silverman et al., 1999b; Wood, Piancentini, Southam-Gerow, Chu & Sigman, 2006). Hunsley and Lee (2007) compared effectiveness and efficacy trials and found that the level of training offered was similar across settings. Weisz et al. (2009) considered whether community practitioners

could be trained to deliver an EST to adolescents with depression in a time frame realistic within community settings. They found minimal training in a specific EST (6 hours and additional ongoing supervision) in community practice was sufficient to establish reasonable competence in outcomes in therapists with a mean of four – five years training in mental health. In a recent treatment review, most interventions for anxiety disorders in youth were rated as 'reasonably trainable' (Chorpita, Daleiden et al., 2011). Lastly, models for training or supervision have been developed for use within the constraints of community resources, including education of clients and cross-agency training (Carroll, Martino & Rounsaville, 2010; Southam-Gerow, Rodriguez, Chorpita, & Daleiden, 2012; Weisz & Gray, 2008).

Altogether, these findings suggest that it may be possible to improve the quality and impact of supervision provided within UC with available resources, if focus shifts to skill development and maintenance rather than remaining on administrative or other tasks. That is, whether or not the supervision delivered in UC is similar to RCTs, skill based supervision appears to be associated with improved therapist competency and outcomes, and could be instituted within the resources available. Any differences between UC and ESTs should not compromise generalizability of results of trials evaluating ESTs to community settings.

In sum, although there have been concerns regarding the generalizability of results of ESTs in the community because of factors such as the complexity of cases, training and supervision requirements, these do not appear well founded. There are inconsistent differences between populations treated within community and RCTs. Further, while the content of training, supervision and adherence monitoring are likely to differ between

treatment trials and community agencies, comparable results have been achieved within resources typically available in community settings.

Summary

Overall, it appears that the results of efficacy trials can be generalised to community settings (Addis & Waltz, 2002; Persons, Bostrom, & Bertagbolli, 1999; Sanderson, Raue, & Wetzler, 1998; Tuschen-Caffier, Pook, & Frank, 2001; Wade, Treat, & Stuart, 1998). While there is some inconsistency with recent treatment trials for anxiety disorders, reviews generally suggest that ESTs outperform UC. If effectiveness trials confirm the findings of efficacy, and ESTs typically achieve better results than UC, the question remains why more agencies do not routinely implement evidence-based practices. Despite the existence of ESTs, the type of treatment an individual receives for any mental health disorder can be 'alarmingly arbitrary' (Parry, Roth, & Kerr, 2005). In a review of community practice, Martino, Ball, Nich, Frankfoter and Carroll (2008) found that clinicians over-reported the extent to which UC included components of ESTs and that the rate of EST use was actually very low. This may account for the inferior results often found in UC and is consistent with findings that UC that includes components of ESTs can achieve comparable results (Kendall et al., 2008; Southam-Gerow et al., 2001).

It appears that factors that cannot be changed regarding community practice (e.g. demographic and symptom profile of clients) are not consistent moderators of outcome and those that are modifiable (training; content of supervision; use of manualised treatments) may not be common practice in UC but when systematically altered, are associated with improved client outcomes. Further, it appears that results of treatments deemed efficacious in research settings are consistently replicated in effectiveness studies in community

settings and that these interventions typically achieve superior results to those of UC. Focus will now shift to how the results of clinical trials for ESTs can be used to evaluate outcomes in communities using 'benchmarking' as a strategy. Results of children's mental health agencies utilising the Systems of Care model for mental health practice will be considered.

Benchmarking

The technique of 'benchmarking' was originally developed as a quality assurance strategy in the context of the Japanese manufacturing industry (Yoshikawa, Innes, Mitchel, & Tanaka, 1993) and has since been introduced into other countries and disciplines, including health care in North America and the United Kingdom (Bullivant, 1996; Lorence & Jameson, 2002). Benchmarking is usually conducted within the context of evaluation and organisational behaviour management, including attempts to identify and/or implement best practices (Francis & Holloway, 2007). Ultimately, the process involves setting performance targets that are ambitious yet realistic, and can facilitate the identification of effective practices that might then be generalised to other settings or groups (Weersing, 2005).

There have been some criticisms of the process of benchmarking. While research suggests that differences between research and community settings are not as marked as once thought, most RCTs are still conducted within universities under different circumstances than are typical in publicly funded mental health settings. Further, small changes in pre-treatment means and standard deviations can have relatively large impacts on the magnitude of pre- post effect sizes, reducing confidence in a benchmark based on effect sizes (Lueger & Barkham, 2010). Lastly, use of a benchmarking strategy is

inferential - any differences (or lack of differences) between benchmarks and outcomes of comparator groups can only be interpreted with caution, since the reason for these differences cannot be established by benchmarking (Weersing, 2005). Despite these concerns, benchmarking is one of the more helpful strategies available to bridge the gap between research and clinical practice; and can be used to contextualise outcomes achieved in community treatment agencies (Minami et al., 2007).

Within health care, benchmarking processes have been applied to two main kinds of service characteristics: the processes associated with client care (e.g. waiting times, procedures used) and the outcomes of the services provided (e.g. symptoms, client functioning) (Trosa & Williams, 1996). Quality assurance efforts have largely focused on processes, which may be more easily influenced by clinicians and administrators than outcomes (Sperry, Brill, Howard, & Grissom, 1996). Other work, however, has targeted the arguably more challenging and relevant target of improving actual clinical outcomes (Hamerlynck, 2005).

Benchmarking and mental health. Benchmarking allows comparisons against a standard without needing an experimental control or comparison group and therefore may be an efficient strategy to use when evaluating community treatment centres (Merrill et al., 2003; Minami et al. 2009). Benchmarking strategies have been used to evaluate outcomes of interventions within mental health contexts. The process generally follows four steps; (1) defining the problem, population and treatment model, (2) choosing or creating a 'gold standard' benchmark (often from the research literature), (3) measuring this outcome in the population(s) of interest and (4) comparing performance of these population(s) against the gold standard benchmark (Weersing & Weisz, 2002).

Studies utilising a benchmarking strategy in mental health have generally made use of one of two major approaches. In the first approach, outcomes of efficacy trials are used to establish an outcome benchmark and results of effectiveness trials are judged against this. Typically, these studies involve evaluating whether a particular empirically supported treatment (e.g. cognitive behaviour therapy) achieves results comparable to efficacy trials when implemented in the context of a community agency. This type of benchmarking study can therefore be considered a kind of effectiveness trial (Weersing, 2005). The second approach uses results from published clinical trials to establish two kinds of benchmarks – treatment efficacy and natural history. The 'treatment efficacy benchmark' reflects outcomes expected following receipt of an empirically supported treatment whereas the 'natural history' benchmark reflects outcomes expected following the passage of time alone. The impact of treatment from community agencies is then evaluated against these benchmark standards (e.g. Weersing & Weisz, 2002).

There are several examples in the adult mental health literature where the first type of efficacy/ effectiveness benchmarking strategy has been used. For instance, studies have examined the generalizability of interventions for different diagnoses including depression (Merrill et al., 2003; Persons et al., 1999), Social Phobia (McEvoy, Nathan, Rapee, & Campbell, 2012); Panic Disorder (Stuart, Treat, & Wade, 2000; Wade et al., 1998), Obsessive– Compulsive Disorder (Franklin, Abramowitz, Kozak, Levitt, & Foa, 2000), and with a diagnostically heterogeneous group (McEvoy & Nathan, 2007). Further, the studies have taken place in various types of clinics including outpatient clinics (Martinsen, Olsen, Tönset, Nyland, & Aarre, 1998), community mental health centres (Wade et al., 1998),

public mental health units (Garcĩa-Palacios, Hoffman, Carlin, Furness, & Botella, 2002) and private clinics (e.g. Gaston, Abbott, Rapee, & Neary, 2006).

Some studies found that treatments achieved comparable results in the community to what was achieved in efficacy trials (Persons et al., 1999; Warren & Thomas, 2001) while others found clients improved, although not as much as within efficacy trials (Oei & Boschen, 2009).

This type of benchmarking strategy has also been applied to evaluating treatments for youth. Curtis, Ronan, Heiblum and Crellin (2009) considered the effectiveness of multi-systemic therapy (MST) in the treatment of juvenile offenders seen in a community mental health agency in New Zealand. Benchmark standards for expected outcome were established by aggregating results of three efficacy studies of MST for the treatment of juvenile offenders. They found that when MST was implemented in the community, results were comparable and even superior on some dimensions to the benchmark standard.

Shirk, Kaplinski and Gudmundsen (2009) compared results of youth with depression receiving services in school-based services, to those of youth treated within efficacy trials. They found youth achieved similar and sometimes superior results to published trials. Dobson, Hopkins, Fata, Scherrer, and Allan (2010) compared results from their youth mental health agency to those of a single randomized control trial for the prevention of depression in at-risk adolescents (Clarke, Hawkins, Murphy, & Sheeber, 1995). They found that their intervention produced greater changes than those of the RCT. While results of this study are encouraging, generating a benchmark standard based on a single treatment study is somewhat problematic, since it is likely to be less reliable than a benchmark standard based on aggregating results of a number of trials (Minami, Serlin, et al., 2008).

Farrell, Schlup and Boschen (2010) used a benchmarking approach to evaluate treatment for youth with Obsessive - Compulsive Disorder (OCD) in a private clinic. Their treatment was based on a standardized manual used in a clinical trial (Barrett, Farrell, Dadds, & Boulter, 2005; Barrett, Healy-Farrell, & March, 2004). They used a common measure of symptoms in OCD - the Children's Yale Brown Obsessive - Compulsive Scale (CY-BOCS; Goodman, Price, Rasmussen, & Mazure, 1989) and compared outcomes of the private clinic against those from five clinical trials that used the CY-BOCS. They found their study group achieved similar or better results than some efficacy trials for youth with OCD (de Haan, Hoogduin, Buitelaar, & Keijsers, 1998; Franklin et al., 1998) but were not as good as one (Barrett et al., 2004). The study sample was deemed broadly comparable to those within the published trials, although the involvement of one of the program developers (Farrell) in the effectiveness trial and the fact that the community sample was actually conducted within a private clinic meant that results would not necessarily generalise to publicly funded community mental health agencies. The authors did not aggregate results of efficacy trials, nor did they use statistical techniques to compare the effect size of their study against the benchmark efficacy studies. This was probably because visual inspection of results revealed substantial overlap between their results and those of the efficacy trials. Relying on visual inspection of effect sizes or confidence intervals, however, is problematic since overlap in confidence intervals does not necessarily mean that there is no significant difference in terms of pre-post improvement between the groups (Wolfe & Hanley, 2002). Further, judgments based on visual inspection of data may be prone to bias.

Hunsley and Lee (2007) developed an interesting adaptation of the efficacy/ effectiveness benchmarking strategy. They reviewed effectiveness RCTs for adult and youth interventions, and grouped these within diagnostic categories. They then systematically compared outcomes from these effectiveness studies to an efficacy study within the same diagnostic category. Comparisons between the effectiveness studies and the benchmark efficacy study were made on a variety of dimensions, including pre-post effect sizes; clinically significant improvement and processes such as the amount of supervision involved. They found that effectiveness studies achieved comparable results to efficacy on most outcome and process dimensions.

The aforementioned studies used benchmarking to evaluate the impact of particular ESTs when delivered in community settings. It is also possible to use benchmarking to evaluate UC – that is, the usual service delivered within community agencies against ESTs. This may be a particularly helpful and relevant strategy for use in settings that routinely collect outcome data pre- and post- treatments. In these situations, information regarding client demographics or symptom profile is generally available, although the particulars of the intervention that they received usually are not. The evaluation is thus based on the premise that it is possible for clients to achieve broadly similar results to those obtained within clinical trials for the same problems, if they receive an appropriate treatment (i.e. an EST). Establishing and using a standard for improvement might be preferable to simply evaluating whether *any* improvements in clients' symptoms from pre- to post-treatment occurred in UC. This is because part of the challenge of evaluating UC is that remission of symptoms can occur with the passage of time alone. Therefore, even if community agencies observe improvements in their client populations between pre- and post-treatment,

this is not enough to be confident of the efficacy of the treatment they have delivered. It is not clear whether this magnitude of improvement reflects the impact of treatment or simply remission that could be expected to occur with the passage of time alone. A benchmarking strategy helps address this question, since outcomes for both empirically supported treatment and wait list controls can be used as comparators. Thus, community agencies can evaluate their services without the need for a comparator group – by generating standards for improvement from wait list groups that reflect improvement expected from the passage of time alone. Thus, the second type of benchmarking study is exemplified in recent innovative research that has examined the impact of UC in community agencies against benchmark standards, rather than evaluation of a specific EST in the community.

Examining results of community agencies against those from clinical trials can act as a catalyst for changing practices if results are weaker than expected. Alternatively, if results meet or exceed expectations, this can help generate ideas for dissemination and implementation of effective practice. In the same way that giving feedback about client progress to individual clinicians can reduce the proportion of clients who fail at treatment (Harmon et al 2007; Lambert, Hansen, & Finch, 2001; Lambert, 2007; Lambert, Harmon, Slade, Whipple, & Hawkins, 2005), providing feedback regarding outcomes of agency performance against a pre-determined research-based standard could also shape practice and hence potentially improve collective outcomes.

In a large-scale effectiveness trial, Westbrook and Kirk (2005) found that 1200 adults with anxiety, depression and eating disorders treated in the National Health Services in the United Kingdom achieved improvements slightly lower than RCTs. This is a remarkable outcome, given the scale of the project, the diversity of clients and ESTs and the limited

opportunity for checks of adherence to treatment or supervision protocols.

Blais et al., (2012) made use of benchmark standards to evaluate outcomes of a group of adults with depression and anxiety receiving UC in an outpatient clinic, based in an academic setting. They used treatment and wait list control benchmarks established by aggregating results of ESTs for anxiety and depression (Minami et al. 2007; Norton & Price, 2007). They concluded that the university clinic achieved results that were superior to a wait list control group, but not as large as the efficacy trials. They did not use statistical analysis to compare the results of their community agency with those of the efficacy trials. Schindler, Hiller and Witthöft (2011) found the pre-post effect size of their subsample of clients receiving UC in a community agency was comparable to those of treatment groups within published RCTs. This study represented an advance in benchmarking methodology since they also included benchmarks for rates of 'response' and 'remission' (that is, a pre-determined level of improvement in symptoms; and movement into 'normal' range of functioning on a symptom measure). They concluded recovery rates were weaker in the community group than in clinical trials. They did not use statistical analysis to draw this conclusion, either, however, nor did they combine results of RCTs to generate a single aggregate benchmark effect size.

Minami, Wampold et al. (2008) conducted one of the more methodologically sophisticated benchmarking studies to date. They used meta-analyses of randomised controlled trials (RCTs) to establish benchmarks for outcomes of treatment for adult depression, and compared the effectiveness of UC delivered in a community service against these. The benchmark was established using a measure of outcome with a similar level of specificity and reactivity as the one used in their community sample (i.e. the

Outcome Questionnaire-30.1 (OQ-30); Lambert et al., 2003). The analysis was further enhanced by considering the outcomes of the community service against both treatment and control group benchmarks. Thus, they were able to consider whether the results of the community agency resembled those of treatment groups or wait list control groups within clinical trials. They found that clients in the managed care agencies that they evaluated improved as much as participants in RCTs receiving ESTs.

To date, there has been limited application of benchmarking to treatment for youth in community mental health services. One study of treatment of youth with depression compared results of youth seen in community services against benchmarks established from treatment and control groups of 17 clinical trials (Weersing & Weisz, 2002). Thus, benchmarks generated by this study were based on substantially more treatment trials than previous benchmarking studies. They found that results of community services were more similar to control groups than treatment groups of clinical trials. That is, the improvement in clients' symptoms in the community service was no different than what might have happened if the youth had received no service. Weersing and Weisz (2002) used a *t*- test to compare the effect sizes obtained for their treatment and control groups and those obtained within the community sample. Minami et al. (2007) point out, however, that this type of statistical analysis might be problematic, because the large number of participants in these comparisons can sometimes mean that *statistically* significant but *clinically* trivial results are obtained. That is, while differences between the groups might be statistically reliable, the magnitude of this difference can be so small that in the 'real' world, it would not reflect a clinically meaningful difference. Instead, they advocate for the use of a 'range null' hypothesis testing procedure (Serlin, 1985; 1993) whereby the difference between two

effect sizes must be larger 0.2 (i.e. a 'small' effect size, within Cohen's 1998 classification system) to be considered clinically significant.

Benchmarking methodology. The review thus far has focussed on the utility of benchmarking as a strategy to evaluate either the transportability of ESTs in the community or the services delivered in UC. There are methodological issues that should be carefully considered when adopting benchmarking as a way to evaluate results of community agencies. These methodological issues relate to measurement and conceptualisation of outcome, as well as the challenges of matching clinical trials with the circumstances within community settings. Details regarding these issues will now be addressed.

All benchmarking studies of outcome reviewed compare the effect sizes of RCTs with those of the community group under consideration. It is important to carefully consider the nature of outcome measures used to compare results of community and RCTs agencies. For instance, the magnitude of effect sizes is moderated by the 'specificity' of the measure used (Minami, Serlin et al., 2008). That is, measures of *specific* constructs (e.g., anxiety) tend to generate larger effect sizes than measures of *general* constructs (e.g., overall psychopathology), particularly when the specific constructs are targeted within the intervention. Measures of specific symptoms such as the Revised Children's Manifest Anxiety Scale (RCMAS; Reynolds & Richmond, 1985) or Children's Depression Inventory (CDI; Kovacs, 1992) would be considered 'high specificity' measures, whereas instruments gauging general internalising psychopathology such as the Child Behavior Checklist-Internalising broadband scale (CBCL-Int; Achenbach, 1991; Achenbach & Rescorla, 2001) could be considered a 'low specificity' measure. Further, there are differences in the magnitude of effect sizes for different constructs (e.g. symptoms versus

functioning; Karpenko, Owens, Evangelista, & Dodds, 2009; Rosenblatt & Rosenblatt, 2002). Thus, when comparing effect sizes in community settings to those of published treatment trials, it is important to match the specificity (i.e., global or symptom domains) and constructs (e.g., symptoms or functioning) of the measures upon which effect sizes are based.

The 'reactivity' of the measure is also important when attempting to generate benchmarks that are comparable across settings. The 'reactivity' of a measure has been operationalised as who rates it - clinician or client (Minami et al., 2007) and in children's mental health research, this can be extended to include whether the rating is completed by parent or child. Clinician-rated measures tend to generate larger effect sizes than client-rated measures (Fava, Evins, Dorer, & Schoenfeld, 2003; Rief et al., 2009). In children with anxiety disorders, parent-reported reductions in symptoms are approximately twice as large as child-reported reductions in symptoms (Prins & Ollendick, 2003) and in fact, child-rated symptoms can resemble effect sizes no larger than wait list controls (Barrett et al., 1996; Hudson et al., 2009).

Concerns regarding the impact of the measure used to establish effect sizes are reflected in the fact that most benchmarking studies make use of the same outcome measure to establish the benchmark standard or, at least, measures with comparable reactivity and specificity (e.g. Dobson et al., 2010; Minami et al., 2009; Schindler et al., 2011; Weersing et al., 2006; Weersing & Weisz, 2002). Thus far, most studies have used a symptom-specific measure of outcome to establish the benchmark and to evaluate outcome in their comparator group. However, establishing a treatment efficacy benchmark using a symptom-specific measure may be of limited use for community agencies conducting

large-scale evaluation of the services they deliver. This is because such large scale efforts tend to mandate use of measures that are broad based enough to capture the diversity of presenting problems typically seen in their clinics. These broad based measures are likely to be less sensitive to change in target symptoms and will therefore typically generate smaller effect sizes than symptom-specific measures. This suggests the effect sizes generated from these broad-based measures are not comparable with those generated from specific measures of the target symptom. Therefore, given measures mandated for use within community agencies are typically broad-based, establishing benchmark standards of outcomes using broad-based measures of psychopathology will likely be extremely helpful. For instance, a common broad-based measure of child psychopathology in youth is the Child Behavior Checklist (CBCL-/6-18) - a measure commonly used in both children's mental health agencies and within many RCTs for children's mental health (Achenbach 1991; Achenbach & Rescorla, 2001).

The focus of discussion thus far has been on use of pre-post effect sizes as the metric most commonly used to generate benchmarks and evaluate the effectiveness of interventions. This information may be of limited utility within clinical settings, however. Many clinicians (and agency managers) do not have a strong understanding of the meaning of effect sizes and they may therefore not make use of this metric of outcome. Further, while effect sizes are relevant to making generalisations about the impact of a treatment on the whole group of participants, they do not provide important information regarding individual differences in treatment outcome (Hunsley & Lee, 2007; Swanson et al., 2001). A potentially important alternative and clinically relevant way of describing outcomes, particularly when trying to establish clinician-friendly benchmarks, is the proportion of

clients who experience either 'clinically significant improvement' or complete recovery from their presenting disorders.

Hunsley and Lee (2007) discuss benchmarks for improvement/ recovery for particular disorders in children and adolescents. The results of recent meta-analyses were used as benchmarks for improvement or recovery rates for each of the target problem areas (e.g. anxiety; depression). This was a challenging task, given that improvement/ recovery rates are not always reported in clinical trials, and/or are operationalized differently across trials. 'Clinically significant improvement' has been operationalized in a number of ways, including the proportion of clients who move from 'pathological' to 'normal' range of functioning on a standardised outcome measure (e.g. Minami et al., 2009). Some studies use a more conservative definition where participants are only considered to have demonstrated 'clinically significant improvement' if they (1) move from 'pathological' to 'normal' range of functioning on standardised measures of pathology (such as the CBCL-/6-18) *and* (2) demonstrate 'reliable' change (Farrell et al., 2010). 'Reliable Change' is a magnitude of improvement that is beyond what might be expected from measurement error alone, and Jacobson and Truax's (1991) 'Reliable Change Index' (RCI) can be used to quantify the amount of improvement required to be considered 'reliable'.

Clinical trials also report results of 'recovery' - a construct similar to 'clinically significant improvement'. This construct is often operationalized as the proportion of participants at post treatment who no longer meet criteria for a DSM diagnosis. However, there may be reasons why 'recovery' from a DSM diagnosis might not be a helpful metric of outcome to use within community agencies. First, DSM diagnoses are often established with different degrees of rigor and standardisation in clinical practice than in research trials

which means change in one might not be comparable to change in the other (Jensen & Weisz, 2002; Lewczyk, Garland, Hurlburt, Gearity, & Hough, 2003). The second reason relates to the first. It is time consuming and costly to generate a reliable, valid DSM diagnosis, and they can only be established by a limited number of highly trained professionals (for instance, physicians or psychologists). Costs mean that sufficiently rigorous diagnostic interviews are unlikely to be routinely conducted for all youth at pre- and post- treatment in community mental health settings. Therefore benchmark standards based on recovery from DSM diagnoses are likely to be of limited usefulness in community settings. For this reason, benchmarks based on 'clinically significant improvement', established from more portable approaches such as from parent-rated measures, may offer greater utility in community agencies than rates of diagnostic recovery.

Discussion, thus far, has focussed on various issues related to measurement when conducting benchmarking studies. Another important issue to address when applying a benchmarking strategy to evaluate community services is how clinical trials are selected when establishing the treatment benchmark standard in the first place. The limited number of RCTs evaluating the treatment of anxiety disorders in youth means that there has been relatively few effectiveness trials conducted. While it might be ideal to compare outcomes of community settings against outcomes of effectiveness studies (rather than efficacy), this would drastically reduce the number of studies available with which to establish the benchmark, and thus render it less reliable. Further, as noted by Hunsley and Lee (2007), it appears that results of effectiveness and efficacy trials are comparable. For these reasons, both effectiveness and efficacy trials can be used to generate benchmark standards.

Related to consideration of selection of trials to establish benchmarks of outcome are strategies to adequately match participants drawn from large-scale community data bases to RCT samples. In essence, the challenge of this process relates to how precisely characteristics of clients drawn from community databases can be matched to those of participants in published trials. Minami, Wampold et al. (2008) outline a process of 'data reduction' – that is, of identifying a comparable sample of clients from the broader participant pool of community mental health agency data bases. There are some challenges with this strategy. Diagnoses based on community assessments can be somewhat different than those typically used in research studies (Jensen & Weisz, 2002; Lewczyk et al., 2003). This may reflect differences in research and community settings with respect to the rigor and standardisation of the diagnostic process. As an alternative to diagnosis, information regarding clinical presentation can be gleaned from results of dimensional measures of psychopathology with population norms (such as the CBCL/6-18; Achenbach, 1991; Achenbach & Rescorla, 2001). Profiles from CBCL/6-18 have been used to identify youth within broad categories of psychopathology (such as 'anxiety disorders'; or 'affective disorders') (Krol, DeBruyn, Coolen, & van Aarle, 2006).

Data reduction is a strategy that can be used to identify subgroups within a broader community population, and this approach may be particularly helpful when there is no single ideal way for identifying youth with a specific target clinical profile. Considering outcomes of subgroups identified in somewhat different ways can assist in interpretation and cross validation of results. For instance, this could include examining subgroups of clients primarily identified by diagnosis or by profile on measures of psychopathology (such as the CBCL). Oei and Boschen (2009) examined outcomes of two subgroups drawn

from a larger data set of adults treated for anxiety disorders in a private hospital in Australia. They considered the improvement of the full group as well as a group with elevated pre-treatment scores on the main outcome measure (the Beck Anxiety Inventory; Beck & Steer, 1990). Schindler et al. (2011) examined outcomes of both their full sample and a subgroup of adults eligible for inclusion in RCTs. Considering outcomes of different subgroups also enriches understanding of factors that might moderate outcome and can allow for matching with particular benchmarks. For instance, Minami et al. (2009) examined outcomes of students seen at a university counselling centre and extracted subpopulations from the larger data set based on characteristics such as treatment completion ('completer' versus 'intent to treat' clients) and then compared outcomes of these to separate benchmarks (i.e. benchmarks based on RCTs reporting 'completer' and 'intent to treat' data).

Strategies have been suggested to improve the reliability of benchmark standards. Weersing and Weisz (2002) suggest it is optimal to establish benchmark standards that are based on meta-analyses rather than selected trials. This contrasts with methodology in some benchmark research that uses only a single trial for comparison (e.g. Gaston et al., 2006) or lists results of a handful of trials rather than aggregating them into a summary statistic (Oei & Boschen, 2009). The strategy of establishing a benchmark based on the aggregation of a number of treatment trials is advantageous because it reflects results across a broader range of client and agency characteristics. Related to this issue is the use of statistical analysis when examining differences between community groups and published trials. It is optimal to use statistical analysis (rather than visual inspection, which

can be biased) when comparing outcomes of the community group against benchmark standards (Minami, Serlin et al., 2008).

The process of measurement and benchmarking may have many potential contributions to improving outcomes in children's mental health services. Systematic measurement and reflection on outcome, as recommended by Bickman (2008), is one of the cornerstones of benchmarking and can lead to improvement in practice (Hodges & Wotring, 2004). Further, benchmarking may be a useful way to evaluate community services in the absence of control groups.

Thus far, discussion has focused on the importance of treating youth with anxiety disorders, the representativeness of conditions in efficacy trials for treating these youth, the generalizability of results of efficacy trials to community settings, the use of benchmarking as a strategy to evaluate community services, and methodological issues that arise when benchmarking. Attention will now turn to groups of agencies providing children's mental health services including youth with anxiety disorders seen in their care – the Comprehensive Community Mental Health Services for Children and their Families (SOC CMHS).

Comprehensive Community Mental Health Services for Children and Their Families Program Systems of Care (SOC CMHS)

In the United States (U.S.), the 'systems of care' (SOC) movement was developed to improve the quality and outcomes of children's mental health services. Over thirty communities, serving more than 50 000 children, have been funded through the Comprehensive Community Mental Health (CMHS) Services for Children and Their Families program (Children's Mental Health Initiative [CMHI]) to support the adoption of

a SOC approach to service delivery (Manteuffel, Stephens, Sondheimer, & Fisher, 2008). The 'system of care' philosophy encourages the implementation of interventions that are evidence-based and emphasizes individualized, strengths-based, coordinated, culturally competent and community-based services for youth with serious emotional disturbance (Holden, de Carolis, & Huff, 2002; Holden, Friedman, & Santiago, 2001).

Part of the strength of the SOC CMHS model includes its mandated evaluation component. This component facilitates (among other things) the examination of outcomes of children who receive services within these agencies. The use of systematic outcome measurement enables communities to track their own effectiveness, and enables the identification of best practices for particular client groups (Hamerlynck, 2005; Hodges, Xui, & Wotring 2004). Mellor-Clark (2006) discusses the helpfulness of consistent feedback regarding outcomes to encourage a 'culture of curiosity'. That is, for clinicians to become curious about the outcomes of the services they deliver, factors that might contribute to strong or poor results and what might be changed to improve or generalize helpful practices.

The development of a comprehensive data set capturing the longitudinal outcomes of clients can facilitate this 'culture of curiosity' within SOC CMHS agencies, as they reflect on the impact of their services (Hodges, Doucette-Gates, & Kim, 2000; Hodges, Doucettes-Gates, & Liao, 1999; Walrath, Mandell, & Leaf, 2001). Current reports of the impact of SOC CMHS services provide information regarding changes in symptoms over time. SOC CMHS agencies aspire to using EBP, and previous research has found the majority of clinicians working in SOC agencies report using EBP (Sheehan, Walrath, & Holden, 2007). However, it has not been clearly demonstrated that these communities achieve results

comparable to those of efficacious treatments, rather than natural remission in youth with anxiety disorders. For instance, it may be that, similar to results of studies of other community settings, results of the SOC CMHS community treatments more closely resemble UC than ESTs (Weisz et al., 2004). The ability of SOC CMHS to achieve results comparable to those of clinical trials would have substantial implications, since it would suggest that it is possible to design and implement large scale system-wide models in community settings which deliver services of comparable impact to those in clinical trials, as opposed to negligible effect sizes typically seen in UC.

As mentioned, anxiety disorders are some of the most prevalent problems of children attending children's mental health agencies, but there are relatively few studies examining the performance of UC in community settings against results of efficacy trials for these disorders. As noted, there is reason to believe that treatment delivered in such settings may not be optimal, with common barriers to dissemination and uptake of EBP being well documented (e.g. Persons, 1997).

Thus, a benchmarking strategy can be applied to evaluating the effectiveness of SOC CMHS agencies, although it is important to exercise caution in particular aspects of the methodology. In the present study, the strategy will be applied to evaluate the outcomes of the SOC CMHS services for youth with anxiety disorders. This strategy will generate benchmarks for improvement in global internalising psychopathology (as measured by the CBCL-Internalising/ 6-18 scale), and for 'clinically significant improvement' (as measured by improvement on the CBCL-Internalising/6-18 scale). In addition, factors associated with treatment response will be explored.

Study Aims

There are several important gaps in the literature regarding benchmarking of standards of care in child and youth mental health services. To the author's knowledge, treatment and natural history benchmarks have not been established for treatment of anxiety disorders in youth using measures of global internalising psychopathology. Further, a benchmarking strategy has not been applied to the evaluation of SOC CMHS services in the U.S. Applying the strategy will be potentially valuable for several reasons. It will enable the establishment of appropriate benchmarks for anxiety disorders for use by other youth mental health services (including Canadian). Further, the examination of the impact of SOC CMHS services could be used to guide training efforts within those services, showcase effective community-based services and may offer evidence of the value of the model in generating outcomes comparable to RCTs, even in community settings.

The present research addressed the following questions;

(i) What are treatment efficacy and natural history pre-post effect size benchmarks for treatment of youth with anxiety disorders?

(ii) What are treatment and natural history benchmarks for 'clinically significant improvement' of youth with anxiety disorders?

(iii) How do pre-post effect sizes of youth with anxiety disorders treated at SOC CMHS agencies compare to treatment efficacy and natural history benchmark standards?

(iv) How does the proportion of youth with anxiety disorders treated in SOC CMHS agencies who evidence 'clinically significant improvement' compare

to treatment and natural history benchmark standards for 'clinically significant improvement'?

(v) What proportion of youth with anxiety disorders treated at SOC CMHS agencies show 'good' or 'poor' treatment response?

(vi) What demographic, family context, child strength and resiliency, child functional impairment and child psychopathology variables are associated with treatment response of youth with anxiety disorders in SOC CMHS agencies? Eleven variables will be considered.

Hypotheses

One of the principles of the SOC approach is a commitment to use of EBP in community services (Holden et al., 2002), with the majority of clinicians in SOC CMHS agencies reporting use of EBP (Sheehan et al., 2007). Further, previous research has noted that when community agencies implement evidence-based interventions, they can achieve outcomes comparable to those of published clinical trials (e.g. Curtis et al., 2009). Based on these findings, it was hypothesised that outcomes of SOC CMHS agencies treating youth with anxiety disorders would meet or exceed the treatment benchmarks generated from research of ESTs, and would be superior to the passage of time alone. The two kinds of treatment benchmarks utilised were (1) those based on both pre-post effect sizes and (2) those based on the proportion of youth evidencing 'clinically significant improvement' (*CSI*).

There has been some concern that clients seen in community settings are more complex than those recruited within efficacy trials and therefore have predictably worse outcomes (Westen et al., 2004). Previous reviews, however, have concluded that there is

generally no significant relationship between outcomes of treatment for anxiety in youth and the presence of comorbid disorders (Ollendick et al., 2008) or adverse social circumstances (Southam-Gerow et al., 2001). Nonetheless, it is important to examine associations between indices of case complexity (such as comorbidity or adverse social circumstances) and treatment response in order to better interpret outcomes of clients of community agencies. Based on research reviews, it was hypothesised that indices of case complexity (living in poverty or having a comorbid affective or externalising disorder) would not be associated with response to treatment of anxious youth served within SOC CMHS agencies. Past studies have also generally found that demographic variables (including age, sex and ethnicity) are not associated with treatment outcomes in youth with anxiety disorders (Silverman, Pina et al., 2008). Nonetheless, these findings have not been entirely consistent and it is important to consider these basic variables when attempting to understand treatment response in the SOC CMHS youth. Based on the preponderance of evidence, it was hypothesised that demographic variables would not be associated with response to treatment of anxious youth served within SOC CMHS agencies. Finally, there are a number of variables that could be associated with the treatment outcomes of anxious children. These include the number of child and family risk factors (such as history of abuse, substance abuse, psychiatric hospitalisation), caregiver stress, family functioning, child strengths and child functioning at baseline. While these factors have not been explicitly evaluated in previous research, studies on related factors (such as parent psychopathology) have found significant relationships with treatment outcomes (c.f. Berman et al., 2000; Crawford and Manassis, 2001; Liber et al., 2010). Based on these related findings and an intuitive understanding of the relationship between family and child

liabilities and treatment outcome, it was hypothesised that children and families with fewer risk factors, experiencing less caregiver stress, with a greater number of child strengths and better functioning would respond to treatment better than those with more risk factors, stress, fewer strengths and more impaired functioning at baseline.

Thus, the study hypotheses were as follows;

1. Outcomes of youth with anxiety disorders served in the SOC CMHS agencies will surpass the natural history benchmark and meet or exceed the treatment efficacy benchmark for pre-post effect size.
2. Outcomes of youth with anxiety disorders served in the SOC CMHS agencies will exceed the natural history benchmark for *CSI* and will meet or exceed treatment benchmarks for *CSI*.
3. The response to treatment of youth in the SOC CMHS data set will not be associated with indices of case complexity (the presence of a comorbid externalizing disorder, the presence of a comorbid affective disorder, poverty status).
4. The treatment response of youth in the SOC CMHS data set will not be associated with demographic variables (sex, ethnicity, age).
5. 'Good' treatment response of youth in the SOC CMHS data set will be associated with fewer child and family risk factors, less caregiver stress, better family functioning, a greater number of child behavioural and emotional strengths and better child functioning at baseline.

Method

Base Longitudinal SOC CMHS Database

The sample for the present study was drawn from the base data set of the national evaluation of SOC CMHS agencies. This data set consists of the descriptive and outcome data of clients drawn from approximately 30 communities within the United States (US) that received initial programme funding grants between 1997 and 2000. Baseline data was collected between 1995 and 2006. These communities were within approximately 16 states and included rural and urban settings (Manteuffel, Stephens, & Santiago, 2002).

To be included in the SOC CMHS data set, youth had to meet at least one of the following criteria; "(1) a clinical DSM-IV (American Psychiatric Association, 1994) Axis I diagnosis, (2) a score in the clinical range on either the Child Behavior Checklist (CBCL) (Achenbach, 1991) or the Child and Adolescent Functional Assessment Scale (Hodges, 1990), (3) a history of multiple system services (e.g., juvenile justice, child welfare, special education), (4) a history of out-of-home placement, or (5) participation in a special education programme for students with serious emotional disturbance" (Stephens, Holden, & Hernandez 2004, p.182). Thus, the data set consisted of youth drawn from agencies that served youth with complex needs and/or experiencing substantial levels of difficulties.

The SOC CMHS longitudinal data set contained data for 4563 youth in the age range of interest in the present study (6-15 yr olds). Table 1 provides descriptive data from this age matched base SOC CMHS longitudinal data set ('Base longitudinal SOC CMHS'). In addition, Table 1 provides details of two subsets of youth from the longitudinal SOC CMHS who were matched to characteristics of published treatment trials. The two subsets of this longitudinal sample were selected on the basis of a number of inclusion and

exclusion criteria, in an effort to maximise the match between them and the characteristics of youth from published treatment trials. These criteria will be detailed shortly. Selection of the two subsets differed on one inclusion criterion. One subset was selected based on a combination of Child Behavior Checklist profile and DSM diagnosis or presenting problem and (CBCL/6-18; Achenbach, 1991; Achenbach & Rescorla, 2001) (SOC_{CBCL}). The other subset was selected on the basis of DSM diagnosis alone (SOC_{diag}). These two subsets of youth form the participant pool for the present study.

The base longitudinal SOC CMHS data set consisted of youth with a mean age of 11.6 (*SD* = 2.6; Table 1). The majority were male 3130 (68.6%), almost half (41.6.8%) lived in 'poverty' (i.e. annual household income less than US$15 000) and almost half identified themselves as belonging to an ethnic minority (47.5%). The most common custody arrangement was houses headed by a lone biological mother (46.6%, *n* = 2110). Families with a biological parent and second biological or step-parent were the next most common custody arrangement (23.8%, *n* = 1076). A sizeable minority of youth in the base longitudinal sample were Wards of the State (8.2%, *n* = 372).

A sizeable minority of the base longitudinal sample did not have a recorded DSM diagnosis (12.3%; *n* = 559). Examination of the clinical profile of participants with recorded DSM diagnoses revealed a small proportion presented with an anxiety disorder (5.8%, *n* = 232), Post Traumatic Stress Disorder (9.8%, *n* = 392), and there was a larger proportion with a mood disorder (34.9%, *n* = 1396). A substantial proportion of the base population had externalising disorders, most commonly Attention Deficit/ Hyperactivity Disorder (ADHD) (46.4%, *n* = 1859) or Oppositional Defiant Disorder (30.0%, *n* = 1202)

with a smaller proportion with Conduct Disorder (7.6%, *n* = 303). The mean number of diagnosed DSM mental health disorders was 1.71 (*SD* = .77).

The CBCL/6-18 profile of the base longitudinal SOC CMHS youth is also summarised in Table 1. Details regarding the nature of CBCL/6-18 broadband and DSM-oriented scales will be outlined later. It should be noted briefly, however, that the CBCL-DSM Anxiety (CBCL-DSM Anx) scale of the CBCL reflects symptoms of anxiety disorders and the CBCL-Externalising reflect broad symptoms of externalising psychopathology. Around one third of the base longitudinal sample lay in the 'clinical' range on CBCL-DSM Anx DSM-oriented scale (37.4%, *n* = 889) (*T* >= 70), and around three quarters fell in the 'deviant' range on the CBCL- Externalising broadband scale (*T* > = 64) (70.2%). Of those with data, around one third of the sample fell within the 'clinical' range on the CBCL-DSM Anx and 'deviant' range on the CBCL-Externalising (CBCL-Ext) broadband scale (33.4%; *n* = 793). That is, around one third of the base sample was significantly elevated on scales measuring symptoms of anxiety and externalising behaviour.

Most youth received mental health services (81.5%, *n* = 3393) with a sizable portion receiving services through the child welfare (30.0% *n* = 1254) or juvenile justice (17.0%, *n* = 709) sectors. The majority of youth received individual therapy (78.9%; *n* = 3188), and around one third received group (34.1%, *n* = 1367) and/or family therapy (38.9%, *n* = 1572). Well over half indicated that they had received medication as treatment for a behavioural or emotional disorder in the six months prior to the initial interview (66.4%; *n* = 2963).

Youth in the data set had received a wide range of therapy 'dose' (i.e. number of sessions). In the six months between the initial and second data collection points, the median number of sessions for clients in the longitudinal data set was 14 for individual therapy (range 1-210), 20 for group therapy (range 1-540) and 8 for family therapy (range: 1 - 200). The data set did not contain information regarding the content or focus of treatment, but this large number of sessions may reflect receipt of help from multiple services. For instance, the vast majority of youth who reported a large number of sessions in the six month period (>= 27 sessions) received assistance from three or more services (97.9%). This might account for the extremely high number of sessions.

SOC CMHS Subsets

For the purposes of the present study, the base longitudinal SOC-CMHS data set was systematically reduced via a set of inclusion and exclusion criteria. This 'data reduction' strategy was applied in an attempt to maximise the correspondence between youth in the SOC CMHS subsets and participants of published clinical trials. The aim of maximising this correspondence was to improve the validity of comparison between outcomes of youth in SOC CMHS subsets with treatment efficacy benchmarks (established from the outcomes of published clinical trials). Two subsets were used, as a way of cross validating results of two alternate strategies for identifying youth with DSM anxiety disorders from the broader longitudinal data set. Inclusion and exclusion criteria were identical for both subsets, with the exception of the first inclusion criteria, where youth with anxiety disorders were identified somewhat differently. Inclusion criteria for the subsets were as follows;

SOC_{CBCL} subset*;*

- A DSM diagnosis of an anxiety disorder and a score on the CBCL-DSM Anx scale in at least the 'borderline clinical' range of functioning (CBCL-DSM Anx T >= 65)

  or

  A presenting problem of being 'anxious' and a score on the CBCL-DSM Anx scale in the 'clinical' range of functioning (CBCL-DSM Anx T >= 70).

These criteria were used to maximise the likelihood that youth had a clinical profile comparable to those in clinical trials, particularly since the correspondence between clinic and research-derived diagnoses is not always strong (Jensen & Weisz, 2002; Rettew, Lynch, Achenbach, Dumenci, & Ivanova, 2009). The CBCL-DSM Anx scale was developed to correspond to diagnoses of anxiety disorders (Achenbach & Rescola, 2001), and although there has been some mixed results (Ferdinand, 2008), most research suggests it is helpful in identifying youth who meet criteria for a DSM anxiety disorder (Ebesutani, Bernstein, Nakamra, Chorpita et al., 2010; Krol et al., 2006; Nakamura, Ebesutani, Bernstein, & Chorpita, 2009; Seligman, Ollendick, Langley, & Baldacci, 2004). As mentioned, details regarding the psychometric properties of the CBCL-DSM Anx scale of the CBCL-6/18, including discriminant validity, will be discussed in further detail in the 'measurement' section of the Method.

Scores in the 'clinical' range on the CBCL-DSM Anx were supplemented with a presenting problem of 'anxiety', because while many youth in the sample scored in the 'clinical' range on the CBCL-DSM Anx, not all of these youth would have been seeking treatment for difficulties with anxiety. Using presenting problem to supplement CBCL profile increased the likelihood that the youth were experiencing symptoms of anxiety severe enough to warrant a diagnosis and that they were seeking treatment for this issue. The more stringent range of the CBCL-DSM Anx scale (i.e. CBCL-DSM Anx in 'clinical' range) was used, because a presenting problem of 'anxiety' is not a conduit for a DSM diagnosis of an anxiety disorder.

SOC_{diag} subset*;*

- A DSM diagnosis of an anxiety disorder.

Using a DSM diagnosis of an anxiety disorder, without reference to anxiety profile on the CBCL, was used for two reasons: 1) because almost all RCTs for treatment of anxiety in youth use diagnosis alone to identify target youth, without reference to scores on dimensional measures of psychopathology, 2) relying only on a DSM diagnosis meant that the pre-treatment mean and standard deviation of scores on the measure used to establish the pre-post effect size (i.e. CBCL-Internalising scale) were not artificially influenced by aspects of study design and were therefore more likely to mirror those of RCTs. This is important, considering the possible impact of elevated pre-

treatment mean and restricted pre-treatment standard deviation on pre-post effect size.

It should be noted that in the SOC CMHS data set, Post Traumatic Stress Disorder (PTSD) was coded separately from other anxiety disorders. Although PTSD is a DSM IV TR anxiety disorder, a diagnosis of PTSD was not used as an inclusion diagnosis in the present study. This is because none of the published treatment trials that were used to establish outcome benchmarks specifically treated youth with PTSD (see Table 2).

The remainder of inclusion and exclusion criteria were identical for both subsets, and were as follows;

- Youth were aged 6-15 years, inclusive. This criterion was used because examination of clinical trials treating children with mean age 6-12 years (which was the target mean age range) revealed most of these included children up to 15 years old (see Table 2). Thus, the range of youth in the SOC CMHS was extended to be commensurate with the clinical trials used to benchmark outcomes.
- The respondent completing the main outcome measure (the CBCL/6-18) was a caregiver. This criterion was applied because all treatment trials used parent responders to rate the CBCL/6-18, and paid workers are likely to have differing perspectives than parents or guardians (c.f. Achenbach & Rescorla, 2001).
- The child was living with the caregiver providing information in the time leading up to the baseline assessment. This criterion was applied because

thorough knowledge of the child (from having the child live with the respondent) was necessary for a valid completion of assessment materials.

- Youth (or their families) had received individual, group or family therapy. This criterion was applied as an attempt to match the type of treatment tested within published trials to that received at SOC CMHS agencies. While the content of these therapies was not necessarily the same as that within clinical trials, the mode of treatment provided was broadly comparable.

Most exclusion criteria reflected those typically used in clinical trials. Exclusion criteria for the sample were defined as follows;

- Clients with a diagnosis of a pervasive developmental disorder or Mental Retardation. While these co-morbidities occur in clinical settings, they are typical exclusion criteria for clinical trials because they are likely to represent substantial moderators of outcome (e.g. Kendall et al., 2008; Nauta et al., 2003, Shortt, Barrett, & Fox, 2001).
- Related to this issue, participants in the SOC CMHS data set were excluded if they had a DSM diagnosis of Conduct Disorder, Bipolar Disorder, Psychotic Disorder, or a substance related disorder or if their reported presenting problems would likely be a higher treatment priority than anxiety (e.g. fire setting; sexually assaultive behaviour; substance abuse, suicide attempt, homicide threat). Further, they were excluded if their CBCL Externalising broadband scale score was more than one standard deviation higher than their CBCL Internalising broadband scale score (i.e. a T score 10 or more higher than their CBCL-Int T score) (Achenbach & Rescorla, 2001). This criterion

was applied to exclude youth with issues likely to be more pressing than the anxiety disorder. In a treatment setting, the problems of participants with a profile including these criteria would likely be prioritised over their difficulties with anxiety disorders.

Some additional inclusion and exclusion criteria were considered but not used. First, youth who received service from juvenile justice agencies (n = 6 in SOC_{CBCL}; n = 4 in SOC_{diag}), or were in therapeutic foster care (n = 2 in SOC_{CBCL}) were not excluded. Excluding these youth reduced the sample size but did not affect conclusions. Therefore, they were not excluded. Also, while many treatment trials require children to refrain from taking medication or to be on stable doses of medication during the course of treatment (e.g. Kendall, 1994); detailed information regarding medication dose was not readily accessible for the youth in the SOC CMHS. Uncontrolled medication use has been acknowledged as a reality in community research (Weersing, Iyengar, Kolko, Birmaher, & Brent, 2006). Thus, youth were included regardless of the nature of their medication use.

A final consideration for inclusion and exclusion criteria for the SOC subsamples related to treatment 'dose'. There have been mixed findings regarding the impact of 'dose' of therapy on outcomes in UC (Andrade, Lambert, & Bickman, 2000; Bickman, 1999; Bickman, Andrade, & Lambert, 2002; Hansen, Lambert, & Forman, 2002). However, because most RCTs included in the present research reported results of treatment 'completers' – that is, participants who complete treatment and thus receive a reasonable 'dose' of therapy (see Table 3) – a minimum 'dose' of therapy was considered as an inclusion criteria. Examination of both the SOC_{CBCL} and SOC_{diag} subsets revealed that

using only treatment 'completers' reduced sample sizes considerably (n = 63 in SOC_{CBCL}; n = 42 in SOC_{diag}) (see Figure 1), which meant analyses were less reliable and hence the critical values for treatment efficacy benchmark were more stringent (Minami et al., 2009). However, conclusions of analysis were not altered (details for the 'full' and 'completer' subsamples are outlined in the Results section). For these reasons, the full SOC_{CBCL} and SOC*diag* subsets were used in the present study.

The process of data reduction is illustrated in Figure 1. The final subset samples represent a small proportion of the full SOC CMHS longitudinal age matched data set (n = 101, 2.2%, SOC_{CBCL}; n = 70, 1.7%, SOC_{diag}). Each step of the data reduction process was an attempt to match the present subsets to samples within clinical trials. Reduction of the sample was substantial when youth with issues that would take clinical priority (mostly externalising problems) were excluded (reduced from n = 2215 to 929). There was also substantial reduction when only youth with significant problems related to anxiety were included (reduced from n = 622 to n = 101, SOC_{CBCL} or n = 70, SOC_{diag}).

The demographic and treatment profiles of youth in SOC CMHS subsets are outlined in Table 1. As can be seen, the demographic profile and treatment of participants in the matched subsets are similar to those in the base longitudinal subset. Consistent with the inclusion and exclusion criteria of the present study, there was a higher proportion of youth with anxiety disorders and scores in the 'clinical' range of the CBCL-DSM Anx in the subsets than in the full sample and there was no youth with excluded diagnoses such as Psychotic Disorder or a substance related disorder. All youth had either a DSM diagnosis of an anxiety disorder, or a presenting problem of anxiety. All youth in the SOC_{CBCL} and the majority of youth in the SOC_{diag} scored in the 'borderline' or 'clinical' range of the

CBCL-DSM Anx (100%, n = 101, SOC_{CBCL}; 74.0%, n = 37, SOC_{diag}), a third in the SOC_{CBCL} (36.6%, n = 37) and all of the SOC_{diag} (100%, n = 70) had a diagnosis of an anxiety disorder. A substantial proportion of both subsets were diagnosed with ADHD (49.5%, n = 50 in SOC_{CBCL}; 32.9%, n = 23 in SOC_{diag}). The pre-treatment CBCL-Externalising (CBCL-Ext) scores were high in both SOC CMHS subsets (SOC_{CBCL} mean (SD) = 70.0 (9.2); SOC_{diag} mean (SD) = 64.73 (11.0). A substantial proportion of youth in the SOC_{CBCL} (77.2%, n = 78) and somewhat fewer in the SOC_{diag} (50.0%, n = 25) fell within the 'clinical' range of scores for both the CBCL-DSM Anx and CBCL-Ext. That is, they had clinically elevated symptoms of both anxiety and externalising psychopathology. There was overlap between the two subsets, with 37 youth belonging to both groups.

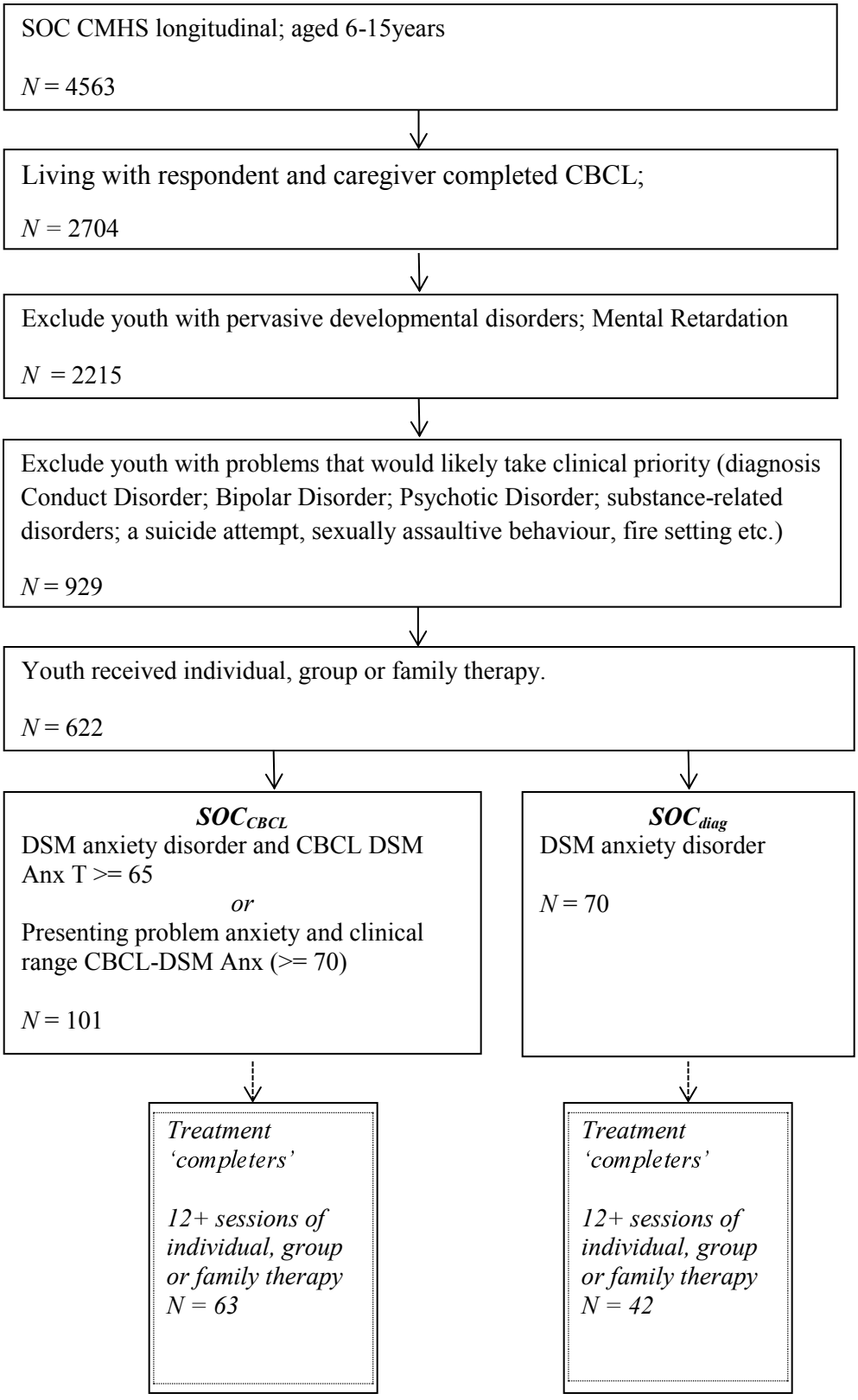

Figure 1. Data reduction from base longitudinal age matched SOC CMHS to SOC CMHS subsets.

Table 1

Profile of Youth from Base Longitudinal SOC CMHS and SOC CMHS Subsets

	Base longitudinal SOC CMHS	SOC CMHS$_{CBCL}$	SOC CMHS$_{diag}$
Demographics			
Youth *n*	4563	101	70
Male *n* (%)	3130 (68.6)	71 (70.3)	50 (71.4)
Age *M* (*SD*, Range)	11.6 (2.6, 6-15)	11.0 (2.5, 6-15)	11.1 (2.6, 6-15)
Ethnic minority *n* (%)	2168 (47.5)	39 (38.6)	31 (44.3)
Annual household income <US$15 000 *n* (%)	1896 (41.6)	42 (41.6)	27 (38.6)
Custody			
Two parents (biological/ step) *n* (%)	1076 (23.8)	28 (27.7)	27 (38.6)
Bio mother only *n* (%)	2110 (46.6)	46 (45.5)	29 (41.4)
Bio father only *n* (%)	180 (4.0)	1 (1.0)	2 (2.9)
Adoptive *n* (%)	242 (5.3)	8 (7.9)	1 (1.4)
Grandparents *n* (%)	325 (7.2)	10 (9.9)	7 (10.0)
Ward of the State *n* (%)	372 (8.2)	2 (2.0)	1 (1.4)
Other *n* (%)	196 (4.3)	6 (5.9)	3 (4.3)
Family and child risk factors			
None *n* (%)	348 (7.7)	5 (5.0)	5 (7.1)
One *n* (%)	417 (9.2)	13 (12.9)	6 (8.6)
Two *n* (%)	516 (11.4)	11 (10.9)	14 (20.0)
Three or more *n* (%)	3238 (71.7)	72 (71.3)	45 (64.3)
Sector of services received			
Mental health *n* (%)	3393 (81.5)	94 (93.1)	64 (91.4)
Education *n* (%)	2097 (50.2)	62 (61.4)	39 (55.7)
Health *n* (%)	751 (18.0)	28 (27.7)	10 (14.3)
Social services/ welfare *n* (%)	1254 (30.0)	27 (26.7)	19 (27.1)
Juvenile Justice *n* (%)	709 (17.0)	6 (5.9)	4 (5.7)
Other *n* (%)	793 (19.0)	16 (15.8)	10 (14.3)
Three or more services *n* (%)	4052 (97.1)	96 (95.0)	65 (92.9)

Treatment[a]			
Group therapy *n* (%)	1376 (34.1)	21 (20.8)	16 (22.9)
Individual therapy *n* (%)	3188 (78.9)	96 (95.0)	67 (95.7)
Family therapy *n* (%)	1572 (38.9)	42 (41.6)	31 (44.3)
Number of sessions[a]			
Group therapy median (1st quartile, 3rd quartile)	20 (7, 35)	10 (10, 24)	12 (5, 19.5)
Individual therapy median (1st quartile, 3rd quartile)	14 (6, 24)	15 (6, 24)	15 (6, 24)
Family therapy median (1st quartile, 3rd quartile)	8 (4, 20)	6 (4,12)	6 (3, 24)
Medication for behavioural/ emotional: Yes *n* (%)	2963 (66.4)	83 (82.2)	50 (71.4)
Diagnosis[a]			
Anxiety disorder *n* (%)	232 (5.8)	37 (36.6)	70 (100.0)
PTSD *n* (%)	392 (9.8)	11 (10.9)	4 (5.7)
Mood disorder *n* (%)	1396 (34.9)	34 (33.7)	21 (30.0)
Adjustment Disorder *n* (%)	424 (10.6)	12 (11.9)	6 (8.6)
Oppositional Defiant *n* (%)	1202 (30.0)	24 (23.8)	6 (8.6)
ADHD *n* (%)	1859 (46.4)	50 (49.5)	23 (32.9)
Conduct Disorder *n* (%)	303 (7.6)	0 (0)	0 (0)
Bipolar Disorder *n* (%)	208 (4.6)	0 (0)	0 (0)
Substance related disorder *n* (%)	280 (5.9)	0 (0)	0 (0)
Psychotic Disorder *n* (%)	154 (3.2)	0 (0)	0 (0)
PDD/Autism *n* (%)	118 (2.9)	0 (0)	0 (0)
Mental Retardation *n* (%)	150 (3.7)	0 (0)	0 (0)
Comorbid anxiety and externalising disorder *n* (%)	100 (2.5)	17 (16.8)	29 (41.4)
Comorbid anxiety and mood disorder *n* (%)	80 (2.0)	9 (8.9)	21 (30.0)
Number of mental health diagnoses recorded *n* (*SD*)	1.7 (0.8)	1.8 (0.8)	2.0 (0.7)

Child Behavior Checklist			
CBCL-Internalising T score mean *(SD)*	65.38 (11.0)	73.3 (6.6)	68.4 (9.0)
CBCL-Externalising T score mean *(SD)*	70.2 (10.4)	70.0 (9.2)	64.7 (11.0)
CBCL–DSM Anx scale in 'borderline' or 'clinical' range at baseline *n* (%)	889 (37.4)	101 (100)	37 (74.0)
In 'clinical' range on both CBCL–DSM Anx and CBCL-Ext at baseline *n* (%)	793 (33.4)	78 (77.2)	25 (50.0)

Note. Percentages of participants with available data.
[a]Not mutually exclusive.

Measures

A standard protocol was used for data collection in the SOC CMHS national evaluation study. Measures included both standardised instruments and those developed specifically for the project. Instruments used in the present study included the Descriptive Information Questionnaire (DIQ; Center for Mental Health Service, 2004), the Multi Sector Service Contract Questionnaire (MSSC), the Child Behavior Checklist (CBCL/6-18; Achenbach, 1991; Achenbach & Rescorla, 2001), the Child and Adolescent Functional Assessment Scale (CAFAS; Hodges, 1990; 1996; 2005), the Behavioral Emotional Rating Scale (BERS; Epstein & Sharma, 1998), the Caregiver Strain Questionnaire (CGSQ:7; Heflinger, & Bickman, 1998) and the Family Assessment Device – General Functioning Scale (FAD-GFS; Epstein, Baldwin, & Bishop, 1983).

Descriptive Information Questionnaire (DIQ; Center for Mental Health Service, 2004). The DIQ is a semi-structured interview that was designed for the national evaluation study (Center for Mental Health Service, 2004)**.** The DIQ gathers descriptive information

regarding client demographics, presenting problems, medications and child and family risk factors. 'Child risk factors' include previous psychiatric hospitalisation, a history of being physically abused, a history of being sexually abused, a history of having run away, a suicide attempt, a history of substance abuse or a history of being sexually abusive toward others. 'Family risk factors' include a history of domestic violence/ spousal abuse, a history of mental illness in biological family, whether biological parents have ever been in a psychiatric hospital; whether biological parents have ever been convicted of a crime, whether there is a history of substance abuse among biological family, and/or whether the child's biological parents have received treatment for substance abuse. The DIQ was administered by the clinician, using caregivers as respondents.

Multi Sector Service Contact Questionnaire (MSSC). The MSSC was developed for the national evaluation study. The MSSC assesses the youth and families' use of services and whether caregivers perceived services to have met the child/ family's needs. 'Use of services' included gathering information regarding the amount of service (i.e. how many sessions), type of service (e.g. individual, group or family therapy) and sector of services received (e.g. mental health, education, health, social services/welfare, juvenile justice). The MSSC does not collect information regarding the content or focus of service.

Child Behavior Checklist (CBCL/ 6-18; Achenbach 1991; Achenbach & Rescorla, 2001). The CBCL/6-18 is a widely used parent-report measure of child psychopathology that gives a standardized measure of symptomatology for children aged 6 to 18 years (Achenbach, 1991; Achenbach & Rescorla, 2001)**.** There are youth self-report (YSR) and teacher-report (TR) versions of the CBCL. The CBCL/6-18 consists of 118 items rated on a three point scale (0, 1, or 2). Items can be organised into nine empirically-

derived syndrome scales (Aggressive Behavior; Anxious/Depressed; Attention Problems; Rule-Breaking Behavior; Social Problems; Somatic Complaints; Thought Problems; and Withdrawn/Depressed). Of all syndrome scales, the most commonly used as a measure of anxiety symptomatology is the CBCL Anxious/Depressed (CBCL-AnxDep), which is made up of items reflecting both anxiety and depressive symptoms (which consistently covary). The CBCL AnxDep syndrome scale has been used to measure the impact of treatments of anxiety disorders in youth (e.g. Rapee et al., 2006). Six CBCL DSM oriented scales have also been developed (Affective Problems; Anxiety Problems; Somatic Problems; Attention Deficit/ Hyperactivity Problems; Oppositional Defiant Problems; and Conduct Problems). In contrast to the syndrome scales, the item pools for the DSM-oriented scales were rationally derived and based on consensus judgments of expert diagnosticians regarding their consistency with DSM diagnoses. The CBCL-DSM oriented scale for anxiety disorders (CBCL-DSM Anx) is made up of six items judged to be consistent with Generalised Anxiety Disorder; Specific Phobia and/or Separation Anxiety Disorder (Achenbach & Rescorla, 2001; Achenbach, Dumenci, & Rescorla, 2003). Lastly, scores on the CBCL/6-18 can be organised into two broad-band scales: CBCL-Internalising (CBCL-Int) and CBCL-Externalising (CBCL-Ext). The CBCL-Int broadband scale consists of 32 items reflecting a variety of somatic, depressive, withdrawn or anxious behaviours. The six items of the CBCL-DSM Anx scale are all drawn from these 32 items.

The forgoing scales reflected measures of psychopathology on the CBCL/6-18. There are also three social competence scales (Activities; Social; School) that reflect child functioning at home, with peers and in school. Lastly, the Total Problem scale summarises

results of the syndrome scales and the Total Competence scale summarises the competence scales.

Raw scores on scales are converted to T scores. On syndrome and DSM-oriented scales (including CBCL-DSM Anx) T scores 65-69 fall in the 'borderline clinical' range and those > 69 in the 'clinical' range of functioning. The ranges are slightly different for broadband scales (i.e. CBCL-Int and CBCL-Ext), where T scores 60-63 fall in the 'borderline clinical' and those >= 64 in the 'clinical' range of functioning.

Cut off scores were determined by analyses of Receiver Operating Characteristics (ROC) (Swets & Pickett, 1982). The scores of demographically similar clinical and non-referred populations were compared on these scales; and cut off scores were chosen that "minimised the percent of referred children below it and non-referred children above it" (Achenbach & Rescorla, 2001, p.95). CBCL/6-18 raw scores are preferred for data analysis of the syndrome and DSM-oriented scales, because the T scores of syndrome and DSM oriented scales are truncated, reducing their variability (Achenbach & Rescorla, 2001). Either raw scores or T scores can be used for analysis of broadband scales (i.e. CBCL-Int and CBCL-Ext).

Although no information regarding the psychometric properties of the CBCL/6-18 have been established with youth in the SOC CMHS data set, there has been extensive research regarding its reliability and validity with other populations. The CBCL/6-18 has established reliability (Achenbach, 1991; Achenbach & Rescorla, 2001). Test-retest reliability, Cronbach's alpha and correspondence between inter-parent ratings have been examined. Mean Cronbach's alpha is strong for CBLC-Int (.90), CBCL-Ext (.94) and CBCL-Total (.97) broadband scales. Test-retest reliabilities across broadband scales are

satisfactory (mean *r* = .94), as is inter-parent reliability for CBCL-Int (*r* = .72) and for CBCL-Ext (*r* = .80) broadband scales (Achenbach, & Rescorla, 2001). The mean internal consistency of CBCL-DSM Anx is = .72, and mean test-re test reliability is *r* = .80 (Achenbach & Rescorla, 2001).

There has been extensive research regarding the validity of the scales. The CBCL/6-18 has strong convergent validity with the Conners Parent Questionnaire (Conners, 1973) (*r* = .56 to .86) and the Quay-Peterson Revised Behavior Problem Checklist (Quay & Peterson, 1993) (*r* = .52 to .88) (Achenbach, 1991; Achenbach & Rescorla, 2001), and can discriminate between those referred for mental health services and those who are not (Achenbach, 1991; Achenbach & Rescorla, 2001).

Criterion-related and construct validity for the CBCL/6-18 have been examined. CBCL/6-18 syndrome scales have been replicated in a number of cultures, and are significantly associated with genetic and biochemical markers (Achenbach & Rescorla, 2001). Of relevance to the present research is the ability of the CBCL- DSM Anx to discriminate between youth with and without anxiety disorders. One study found the CBCL-DSM Anx scale to be better than the corresponding CBCL-AnxDep syndrome scale at discriminating youth with anxiety disorders (Ebesutani et al., 2010) while others have found them to be comparable (Achenbach et al., 2003). Some studies have found the CBCL-DSM Anx to have only 'fair' discriminant validity when distinguishing youth with anxiety disorders from those without (Ferdinand, 2008); while others have found the scale to have 'good' (or 'strong') discriminant validity (Ebesutani et al., 2010; Nakamura et al., 2009) including being able to distinguish youth with anxiety disorders from those with externalising disorders. Seligman et al. (2004) found the CBLC-DSM Anx was able to

distinguish youth with anxiety disorders from those with no anxiety disorders; and from those with externalising disorders, but not from those with affective disorders (Seligman et al., 2004). Ebesutani et al. (2010), however, found the CBCL-DSM Anx was able to distinguish youth with anxiety disorders from those with affective disorders, using Receiver Operator Characteristic curves (ROCs). Krol et al. (2006) compared profiles of youth on the CBCL/6-18 against DSM diagnoses generated from an established semi-structured diagnostic interview (Diagnostic Interview Schedule for Children, DISC IV; Shaffer, Fisher, Lucas, Dulcan, & Schwab-Stone, 2000). They found good specificity when using the 'strict' cut offs (i.e. 'clinical' range, T >= 70) for CBCL-DSM Anx (specificity = .87 - .89) and better sensitivity when using the 'lenient' cut offs (i.e. 'borderline' and above, T >= 65) (sensitivity = .27 - .50). Positive predictive power (PPP) (i.e. the chances that a youth with an anxiety disorder will be in 'deviant' range on the CBCL-DSM Anx) ranged from .29 - .43 when 'strict' scoring rules were used and .21 - .36 when 'lenient' rules were used. Negative predictive power (NPP) (i.e. the chances that a youth without an anxiety disorder will fall in the 'normal' range on the CBCL-DSM Anx) ranged from .89 - .94 for 'strict' scoring and .79 - .97 for 'lenient' scoring. These results are similar to Lengua, Sadowski, Friedrich, & Fisher (2001), who found PPP = .50 and NPP = .86. That is, the CBCL-DSM Anx might fail to identify some anxious children, but does not identify many youth as clinically 'anxious', who are in fact 'normal', particularly when 'strict' scoring criteria are used.

Child and Adolescent Functional Assessment Scale (CAFAS; Hodges 1990; 1996; 2005). The CAFAS is designed to evaluate impairments in functioning resulting from emotional, behavioural, psychological or substance misuse problems in youth aged

5.5 - 17.5 years (Hodges, 1990; 2005). It is generally completed by the youth's clinician and consists of eight domains measuring functioning in: School/Work, Home, Community (reflecting delinquent behaviour), Behavior Toward Others, Moods/Emotions (primarily depression and anxiety), Self-Harmful Behavior, Substance Use and Thinking (reflecting thought problems). Research suggests that following training, raters can achieve good rates of reliability (range .63 - .78) (Hodges & Wong, 1996), and inter-rater reliabilities tend to be high (Pearson's *r* correlations above .92) (Rosenblatt & Rosenblatt, 2002). The CAFAS has been tested with youth receiving services within systems of care agencies and has demonstrated concurrent validity (Hodges et al., 1999; Hodges & Wong, 1996; Hodges & Wong, 1997; Manteuffel et al., 2002), convergent validity (Hodges & Wong, 1996) and predictive validity (Hodges et al., 1999; Hodges et al., 2000; Hodges & Kim, 2000; Hodges & Wong, 1997; Quist & Matshazi, 2000).

Behavioral Emotional Rating Scale (BERS; Epstein & Sharma, 1998). The BERS was designed to identify a youth's behavioural and emotional strengths and resiliencies. It consists of a 52 item checklist, rated on a 4 point Likert scale. These items are organised into five dimensions/ subscales; Interpersonal Strength, Family Involvement, Intrapersonal Strength, School Functioning, and Affective Strength and are combined into an overall BERS Strengths Quotient. The BERS has demonstrated test-retest reliability, inter-rater reliability and internal consistency in other populations of youth (Epstein, Harniss, Pearson & Ryser, 1999). Convergent validity has been established with moderate to high correlations with questionnaires measuring similar constructs (Harniss, Epstein, Ryser, & Pearson, 1999).

Caregiver Strain Questionnaire (CGSQ; Brannan et al., 1997). The CGSQ is a measure of caregiver strain resulting from their child's behavioural or emotional difficulties (Brannan et al., 1997). The questionnaire consists of 21 items rated on a 5 point Likert scale, with scores combining to form three subscales (Objective Strain, Internalising Subjective Strain, Externalising Subjective Strain). The CGSQ has demonstrated reliability, internal consistency (Heflinger, Northrup, Sonnichsen, & Brannan, 1998), construct validity (Brannan et al., 1997) and predictive validity (Foster, Saunders, & Summerfelt, 1996) with other populations of youth.

Family Assessment Device-General Functioning Scale (FAD-GFS; Epstein et al., 1983; Byles, Byrne, Boyle, & Offord, 1988). The FAD-GFS is a 12 item measure of family functioning, including items related to how families interact, communicate and work together (Epstein et al., 1983; Byles et al., 1988). Items are rated on a 4 point Likert scale. The test has demonstrated test-retest reliability (Kabacoff, Miller, Bishop, Epstein, & Keitner, 1990; Miller, Epstein, Bishop, & Keitner, 1985), construct validity (Byles et al., 1988; Epstein et al., 1983; Fristad, 1989; Heflinger et al., 1998; Miller et al., 1985) and predictive validity (Epstein et al., 1983; Fristad, 1989) in other populations of youth.

Data Collection

Participants were recruited for the national evaluation study when they entered services at SOC CMHS agencies. Researchers (employed by the SOC CMHS agencies) explained the study and obtained informed consent (or assent) and then guided them through data collection (Holden et al., 2001). Data collection involved face-to-face interviews and completion of questionnaire outcome measures (Stephens et al., 2004).

Data for the national evaluation project were collected at intake to clinical services, then six months and twelve months later. The baseline and six month data were used for the present study. Standardised time periods were used rather than standardising data collection to post treatment (or follow up after treatment), because in clinical settings, the length of treatment is not fixed. Thus, intervals spacing the timing of data collection were standardised instead.

The standardised protocol that was administered included interview and questionnaires. The DIQ, CBCL/6-18, CAFAS, BERS, CGSQ and FAD-GFS, were administered at intake to the SOC CMHS agencies and six months afterward by agency staff. DSM diagnoses were obtained from management information systems, case records or clinician assessments. Outcome data were entered into a data file by community evaluation staff and sent to the national evaluation office. Results were aggregated across settings.

For the purposes of the present study, access to the national evaluation data set was obtained following a formal application to Macro International Inc.- the company contracted to collect and manage the SOC CMHS national evaluation information.

Ethics Clearance

Ethical clearance for the national evaluation project was obtained from the local Institutional Review Boards of the agencies involved in the project. Guardians of children under 10 years old gave informed consent and their children assented to participate in the SOC CMHS national evaluation study. Both guardians and children over 10 years old gave informed consent for participation in the SOC CMHS national evaluation study. Participant data were submitted to the CMHS without identifying information.

The Research Ethics Board at Lakehead University waived ethical clearance for the present study (see Appendix 1).

Selection of Clinical Trials for Treating Anxiety Disorders

The present study adapted Minami, Serlin's et al. (2008) strategy for benchmarking the impact of community treatments against published trials. After the 'target' population was identified (i.e. children with anxiety disorders, mean age 6-12 years) appropriate clinical trial research studies to establish treatment and natural history benchmarks for clinical practice were identified. Interventions aimed at addressing anxiety disorders were targeted.

Inclusion criteria. Studies were screened, based on the following inclusion criteria:

Age. Mean age of participants in the study was between 6 and 12 years, inclusive. As mentioned, it was noted that many RCTs with mean age in this range included youth up to 15 years old (see Table 2). For this reason, the range of youth included in the SOC subsets was 6-15 years.

Clinically significant anxiety disorders. Participants had clinically significant symptoms of an anxiety disorder as evidenced by a DSM diagnosis (e.g., DSM–IV TR; American Psychiatric Association, 2000), established with a formal diagnostic interview using instruments such as the 'Anxiety Disorders Interview Schedule' (ADIS P-C; Silverman, 1987; Silverman & Albano, 1996; Silverman & Nelles, 1988) or a score falling in the 'clinical' range of impairment in symptoms as measured by a standardised measure of anxious psychopathology.

Comorbid conditions. The intervention was not targeted at youth with developmental disabilities or pervasive developmental disabilities. Further, the major focus of the trial was addressing anxiety disorders, not a comorbid condition (e.g. a personality disorder or

medical condition). Studies of treatments for youths with comorbid conditions were included if the primary defining presenting problem was an anxiety disorder.

Study design. The study had to include the following features; participants were randomly allocated to treatment condition, and parent-rated CBCL-Int pre and post T scores were reported. Criteria related to inclusion of the CBCL/6-18 were applied for at least two major reasons. First, the SOC CMHS national evaluation study used the CBCL/6-18, and the measure is commonly used within clinical trials for the treatment of anxiety disorders in youth. As noted, it is preferable to use the same measure to establish the benchmark and to measure outcome in the comparator community group (Minami, Serlin et al., 2008), as is common practice within previous benchmarking studies (e.g. McEvoy & Nathan, 2007; Minami, Serlin et al., 2008; Weersing et al., 2006; Weersing & Weisz, 2002). This is important, considering, as mentioned, the magnitude of an effect size can be impacted by the nature of the measure used to calculate it (Minami, Serlin, et al., 2008). Second, other community agencies are more likely to have aggregate information on a generic measure such as the CBCL/6-18 than the plethora of target - specific measures used within studies of anxiety disorders (such as the Multidimensional Anxiety Scale for Children (MASC; March, Parker, Sullivan, & Stallings, 1997). A treatment efficacy benchmark based on this broad-based measure of psychopathology is therefore more useful for those agencies than a target-specific measure.

Clinic-referred. At least some participants were referred or self-referred for treatment (i.e., not all participants were recruited to the study through advertising). This criterion was applied because advertising-recruited participants may have better outcomes than those referred to community agencies (Brent et al., 1998; Lincoln & Rief, 2004).

Bona-fide psychosocial treatment. The treatment delivered within the RCT was a psychosocial treatment that had been identified as at least 'probably' efficacious within a recent review (Davis III & Whiting, 2011 or Silverman, Pina, et al., 2008). This criterion was applied so that only interventions with empirical support were used when generating benchmark standards. Further, the treatment delivered was manualised (Minami et al., 2007). Manualised treatments facilitated tests of treatment fidelity within the trials and replication of the intervention in the community.

Identification of Clinical Trials for Inclusion

Consistent with previous benchmarking studies (e.g., Minami, Wampold et al., 2008), several methods were used to identify appropriate clinical trials for treatment of youth with anxiety disorders. First, meta-analyses and reviews of youth psychotherapy were considered as a means of identifying potentially appropriate trials (c.f. Cartwright-Hatton, Roberts, Chitsabesan, Fothergill, & Harrington, 2004; Chorpita, Daleiden et al., 2011; Compton, Burns, Egger, & Robertson, 2002; Compton et al., 2004; David-Ferdon & Kaslow, 2008; Davis III & Whiting, 2011; Hunsley & Lee, 2007; Silverman, Ortiz et al., 2008; Silverman, Pina, et al., 2008; Verdeli, Mufson, Lee, & Keith, 2006; Weisz et al., 2004; Weisz, Jensen-Doss et al., 2006). Second, PsychInfo and Dissertation data-bases were searched for appropriate clinical trials published between January 1990 and September 2011. Date of earliest publication was limited to when the Child Behavior Checklist was first published (Achenbach, 1991) and to contain the number of volumes that had to be searched manually. Searches were conducted using terms including "treatment," "trials", "psychotherapy," "intervention" and "anxiety", "trauma", "Obsessive Compulsive Disorder", "internalising", "phobia", and "emotional disorders", to generate a selection of

potentially appropriate studies. Also, the reference lists of treatment trials were searched. Lastly, data-bases where ongoing RCTs are registered – 'Current Controlled Trials' and 'ClinicalTrials.gov'- were searched to identify randomized control trials that were underway but not published. While both published and unpublished research studies generated from the above search strategies were considered, no unpublished research studies were identified that met all inclusion and exclusion criteria.

Five studies which otherwise met criteria for inclusion in the present research only reported data generated from CBCL-Int raw scores (Bodden, et al., 2008; Cohen, Deblinger, Mannarino, & Steer, 2004; Deblinger, Mannarino, Cohen, Runyon, & Steer, 2011; Levy, Hunt, & Heriot, 2007; Liber et al., 2008). Preliminary analysis revealed studies using CBCL-Int raw scores generated effect sizes that were consistently smaller than the CBCL-Int T scores, possibly due to differences in scaling. Thus, it appeared that effect sizes generated from the CBCL-Int raw scores were not commensurate with those from CBCL-Int T. Therefore, only studies reporting CBCL/6-18 T scores were included. No studies treating youth with PTSD met all inclusion and exclusion criteria. Therefore, youth in the SOC CMHS data set identified with PTSD (but not another anxiety disorder) were not specifically included in the present study.

Using these criteria, 18 clinical trials were identified to establish the clinical benchmarks (Barrett et al., 1996; Barrett, 1998; Beidel et al., 2000; Cartwright-Hatton et al., 2011; Cobham et al., 1998; Flannery-Schroeder & Kendall, 2000; Heyne et al., 2002; Kendall, 1994; Kendall et al., 1997; Kendall, Hudson, Gosch, Flannery-Schroeder, & Suveg, 2008; Lyneham & Rappee, 2006; Nauta et al., 2003; Rapee et al., 2006; Shortt et

al., 2001; Silverman et al., 1999a; Silverman et al., 1999b; Southam-Gerow et al., 2008; Spence, Holmes, March & Lipp, 2006) (see Table 2).

Some RCTs report results of treatment 'completers' whereas others report results of the whole sample including those who did not receive a full 'dose' of therapy because they dropped out before completion of the therapy protocol (i.e. 'ITT' samples). This distinction is relevant, because clients sometimes leave treatment early because they are not benefitting - meaning their sub-optimal results are not included when only 'completer' results are reported. Therefore, completer samples may overestimate treatment effectiveness. Some research suggests that the effect sizes of 'intent to treat' (ITT) samples are around 10% lower than 'completer' samples (Eddy, Dutra, Bradley, & Westen, 2004; Westen & Morrison, 2001). Some previous benchmarking studies have only included used ITT results (i.e. including both treatment completers and non-completers) from RCTs to generate benchmarks (Minami, Serlin et al., 2008). This is because these are considered more appropriate for comparison with results from community data-sets which include both treatment completers and non-completers. However, excluding studies that report only results of treatment completers meant that 13/18 studies would be excluded from analysis. Therefore, studies reporting either 'completer' or 'ITT' data were included when establishing treatment efficacy and natural history benchmarks. In studies where both were reported, the ITT results were included in the present analysis, thus generating a more conservative estimate of treatment efficacy benchmark effect size. It should also be noted that some studies conducted analyses with ITT data but only reported pre- and post-CBCL Int means for treatment completers. These were classified as 'completer' samples since it was the 'completer' means that were used for analyses in the present study. As mentioned,

conclusions reached using the results of SOC CMHS treatment completers (SOC_{CBCL} n = 63; SOC_{diag} n = 42) were not different from those using the SOC CMHS who started but did not receive a full 'dose' of therapy (SOC_{CBCL} n = 101; SOC_{diag} n = 70) (see Results section). Thus the full samples of the SOC CMHS subsets were used in the present study.

Clinical Trial Characteristics

Characteristics of the 18 clinical trials (including client demographics; treatment length and format; treatment setting) are detailed in Table 2. As can be seen, the mean age for treatment participants fell within six to 12 years, and most included youth up to 15 years (and some included youth up to 18 years). The proportion of boys and girls in each study was fairly even. Of the 11 studies reporting minority status of participants, only one sample comprised more than 50% of youth from ethnic minorities (53.6%), whereas most samples had relatively few youth from minority groups (0 - 39.0%) (k = 10). Table 2 also provides details of the key characteristics of therapy length and format, therapy settings, therapists and recruitment strategies. Overall, treatment consisted of 8-18 sessions (median = 12 sessions). While RCTs evaluating treatment for any type of anxiety disorders were eligible for inclusion, only trials addressing either anxiety disorders in general, or specifically Generalised Anxiety Disorder, Separation Anxiety Disorder and phobic disorders (including Social Phobia Disorder) met inclusion and exclusion criterion. RCTs for treatment of Obsessive Compulsive Disorder (OCD) or Post Traumatic Stress Disorder (PTSD) did not. The majority of studies were conducted in university or research-based clinics with graduate students and university-based practitioners (usually programme developers) (15 out of 18), although a substantial minority (4 out of 18) were based in community or hospital settings with community practitioners (one study included both

university and community sites). The studies that provided information regarding supervision usually reported incorporating two-hour weekly group supervision sessions. Treatment fidelity was monitored in the majority of studies, using checklists and video or audio-recording of a random selection of sessions. These checks were usually conducted by independent observers. Most trials used therapists with a Masters degree in a mental health discipline. Some of these were doctoral students. Most studies reported only using referral or self-referral for recruitment to receive services, although seven studies used a mixture of advertising and referral to recruit participants. Thus, a number of studies had at least some of the characteristics of effectiveness trials and made use of levels of supervision and training comparable to many community settings.

Table 2.

Demographic, Treatment, and Recruitment Characteristics of Clinical Trials

Study	Inclusion diagnoses	Total *N*[a]	Age Range (mean)	Boys %	Min-ority %	Treatment	Setting and Therapists	Suprvn and Trtmnt Fidelity	Recrt
Barrett et al. (1996)	OAD SOP SAD	85	7-14 (-)	57.0	-	12 sessions, child individual CBT 12 sessions, child individual CBT + family anxiety management	University Clinical psychts	- Fidelity checkd	Advrtng and referral
Barrett (1998)	OAD SOP SAD	60	7-14 (-)	53.3	-	12 sessions, group child CBT 12 sessions, group child CBT+ family anxiety management	University Clinical psychts	- -	Advrtng and referral
Beidel et al. (2000)	SOP	67	8-12 (10.5)	-	-	12 individual child + 12 group child CBT	Research -	- -	Advrtng and referral
Cartwright-Hatton et al. (2011)	Any anxiety disorder + CBCL Int/ PBCI	74	2.7-9 (6.6)	43.2	25.7	10 sessions, parent group CBT	Hospital Clinical psychts	- Fidelity checkd	Referral and self referral

Cobham et al. (1998)	GAD OAD SOP SAD Simp Ph Agorphb	67	7-14 (9.6)	50.8	-	10 sessions, child group CBT 10 sessions child group CBT + 4 parent anxiety management sessions	University	- Fidelity checkd	Referral
Flannery-Schroeder and Kendall (2000)	GAD SAD SOP	45	8-14 (-)	51.0	11.0	18 sessions, child individual CBT 18 sessions, child group CBT	University Doctoral stdnts	2hr, weekly Fidelity checkd	Referral
Heyne et al. (2002)	School refusal and anxiety diagnosis	65	7-14 (11.5)	54.1	8.2	8 individual child CBT 8 Parent/teacher CBT training 8 child CBT therapy + 8 parent/teacher CBT training.	Medical Centre MA psychts	- Fidelity checkd	Referral and self referral
Kendall (1994)	OAD SAD Avoidant	60	9-13 (-) -	60.0	24.0	17 sessions, individual child CBT	University Doctoral stdnts	- Fidelity checkd	Referral
Kendall et al. (1997)	Any anxiety disorder except Spec Ph as primary	118	9-13 (-) -	62.0	15.0	16 sessions, individual child CBT	University Doctoral stdnts	- Fidelity checkd	Referral

Kendall et al. (2008)	GAD SAD Spec Ph	161	7-14 (-)	56.0	15.0	16 sessions, individual child CBT 16 sessions, family CBT	University Phd psychts MA thrpsts	2 hr, weekly group Fidelity checkd	Advrtng referral and self referral
Lyneham and Rapee (2006)	Any anxiety disorders	100	6-12 (9.4)	51.0	10.0	12 weekly modules CBT bibliotherapy + phone contact 12 weekly modules CBT bibliotherapy + email 12 weekly modules CBT bibliotherapy + client initiated contact	University Grad stdnts	- -	Advrtng referral and self referral
Nauta et al. (2003)	GAD SAD SOP Panic D	79	7-18 (11.0)	49.4	0.0	12 sessions individual child CBT 12 sessions individual child CBT and 7 sessions CPT	University and community Clin psychts Grad stdnts	Weekly group -	Advrtng and referral
Rapee et al. (2006)	Any anxiety disorder	267	6-12 (-)	39.3	-	9 Sessions group CBT Bibliotherapy	University Clin psych Grad stdnts	- -	Referral and self referral

Shortt et al. (2001)	GAD SAD SOP	71	6-10 (7.8)	40.8	1.4	10 sessions (with an extra 2 booster) group child CBT + family skills component	University Doctoral stdnts	Weekly Fidelity checkd	Advrtng and referral
Silverman et al. (1999a)	GAD OAD SOP	56	6-16 (10.0)	60.7	53.6	14 sessions group CBT child and parent separate groups; meet 15 minutes together.	University Doctoral stdnts	- Fidelity checkd	Referral and self referral
Silverman et al. (1999b)	Spec Ph SOP Agoraph	104	6-16 (9.8)	51.9	39.0	10 sessions individual child and parent self control therapy 10 sessions, individual child and parent contingency management	University Doctoral stndts Post docs	Weekly group Fidelity checkd	Referral
Southam-Gerow et al. (2010)	GAD SAD SOP Spec Ph	48	8-15 (10.9)	43.8	-	Mean 14.0 sessions individual child CBT	Community Community practitioners psychts + intern	Weekly group Fidelity checkd	Referral

Spence et al. (2006)	Any anxiety disorder	72	7-14 (-)	58.3	-	Combined 10 individual child CBT; 6 parent CBT group in clinic (+ 2 booster sessions) Combined 5 individual child sessions CBT via internet + 5 in clinic; 3 parent group CBT sessions via internet + 3 group parent in clinic	University Psychts + clinical interns	Bi - weekly group Fidelity checkd	Referral

Note. A dash indicates that the information was not reported. *N* = total number of participants at randomisation; Suprvn= supervision; Trtmnt Fidelity = treatment fidelity; Recrt = recruitment; OAD = Overanxious Disorder; SOP = Social Phobia; SAD = Separation Anxiety Disorder; GAD = Generalised Anxiety Disorder; Spec Ph = Specific Phobia; Panic D = Panic Disorder; Sim Ph = Simple Phobia; Agoraph = Agoraphobia; CBCL Int = score in 'clinical' range on Child Behavior Checklist – Internalising; PBCI = score in 'clinical' range of Preschool Behavior Checklist Internalising; CBT = cognitive behavioural therapy; CPT = parent training; Clinical psychts = clinical psychologists; Grad stdnts = graduate students; Doctoral stdnts = doctoral students; MA psychts = Masters qualified psychologists; Psychts = psychologists; MA thrpsts = Masters qualified therapists; post doc = post doctoral fellow; Fidelity checkd – fidelity of session content checked using checklists and audio or video-tapes of sessions; Advrtng = advertising.

Data Analysis

The first major focus of data analysis was to establish pre-post effect size treatment efficacy and natural history benchmarks. This was achieved by aggregating results of treatment trials using random effects meta-analytic methods described by Hedges and colleagues (Borenstein, Hedges, Higgins, & Rothstein, 2009; Hedges & Vevea, 1998). A fully random effects model was used because there were likely to be a range of true effect sizes both within and between studies (Borenstein et al., 2009) and random effects models generate results that are more generalizable than fixed effects models (Hedges & Vevea, 1998). When aggregating results, studies with more precise effect size estimates were given more weight (Borenstein et al., 2009). Some meta-analyses also attempt to weight the contribution of studies according to their quality (c.f. Moher et al., 1998). This practice is considered problematic, however, for statistical and methodological reasons (Emerson, Burdick, Hoaglin, Mosteller, & Chalmbers, 1990; Greenland & O'Rourke, 2001; Higgens, Altman & Stern 2011; Juni, Witschi, Bloch, & Egger, 1999; Lipsey & Wilson, 2001). Using quality ratings to weight studies generates a compound weight for each study (consisting of weight according to precision and quality). An adaptation of the statistical theory upon which the data analysis is based would be required, to take these quality ratings into account (Lipsey & Wilson, 2001). Further, quality ratings themselves are problematic, because scoring of particular items is somewhat arbitrary and the scales consistent of items reflecting different constructs (such as the quality of reporting practices of studies and the quality of study design per se) (Greenland, 1994). This may explain (in part) why quality scales are not consistently associated with systematic differences in study outcomes (Emerson et al., 1990; Juni et

al., 1999). As part of the Cochrane Collaboration Handbook Review, Higgens & Green (2006) conclude that weighting according to quality within meta-analyses is time consuming, not empirically validated and can lead to misleading results. Thus, in the present research, studies were not weighted according to quality ratings. Rather, characteristics of study design that might systematically influence internal validity and outcome (e.g. randomization of participants; participant recruitment strategy; use of treatment manual) were incorporated in selection criteria for studies and/or were considered in moderator analysis (e.g. 'completer' versus 'intention to treat' analysis).

It is important to recognize that treatment efficacy benchmarks can be affected by publication bias. Publication bias can include the tendency for studies with significant results to be published, while studies with null findings are not (Sackett, 1979). Studies with large samples sizes are likely to have significant results due to increased power. Smaller studies with null findings might not be submitted for publication; or might be less likely to be published if submitted. Thus, publication bias is assessed by considering the distribution of effect sizes relative to the precision of effect sizes (which is related to sample size). Three strategies were used to examine publication bias in the studies selected for the present research. First, a funnel plot charting study effect sizes against their standard errors was generated (Light & Pillemer, 1984). In the absence of bias, the distribution of study effect sizes should be symmetrical around the mean effect size. It is expected that large studies (at the top of the graph) will congregate around the summary mean effect size, and the distribution of effect size estimates for smaller studies (at the base of the graph) will also be symmetrical around the mean but will be more spread out. Hence, a 'funnel' shape will be formed. Conversely, publication bias is consistent with

(although not proven by) an asymmetrical distribution of effect size estimates around the mean, typically featuring a higher concentration of studies exhibiting large standard errors (i.e. less precise outcomes) and large effect sizes, without corresponding studies with small effect sizes and large standard errors (Borenstein, 2009). Duval and Tweedie's 'trim and fill' analysis was used to quantify asymmetry in the distribution of effect sizes (relative to their standard errors) (Duval & Tweedie 2000a; 2000b). The test generates a mean effect size estimate that is adjusted to compensate for the hypothesized publication bias by imputing values of studies assumed to be 'missing'. Lastly, Begg & Mazumdar's rank order correlation test was used to assess evidence of publication bias. This test is based on the assumption that studies with larger effect size estimates are more likely to be published than those with similar sample sizes but smaller effect sizes. Thus, publication bias is consistent with a significant correlation between effect size estimates and the standard errors of the effects (Begg & Mazumdar, 1994). Calculations for synthesizing results of studies and for examining publication bias were performed using Comprehensive Meta-Analysis v.2 (CMA: Biostat, 2006) software.

The second major focus of data analysis involved comparing results of SOC CMHS subsets against these benchmarks, using a strategy adapted from Minami, Serlin et al. (2008). Results from the SOC CMHS subsets were compared to these benchmark standards to establish whether levels of improvement (1) were clinically equivalent to the treatment efficacy benchmark (2) were better than the natural history benchmark but inferior to the treatment efficacy benchmark, or (3) were clinically equivalent to the natural history benchmark (Minami, Serlin et al., 2008). Thus, two comparisons were made for each SOC CMHS subset, the first between the SOC CMHS subset and the

treatment efficacy benchmark, and the second between the SOC CMHS subset and the natural history benchmark.

A 'range null' hypothesis testing procedure was used in this benchmarking analysis. This approach was developed by Serlin and Lapsey (1983; 1985) as way of identifying differences between effect sizes that were both statistically and clinically meaningful. Consistent with previous research, 0.2 (a 'small' effect size; Cohen, 1998) was nominated as a clinically meaningful magnitude of difference between effect sizes (Curtis et al., 2009; Minami, Serlin et al., 2008; Minami et al., 2009). In practice, this meant that differences between the true effect sizes of SOC CMHS subsets and the true benchmark effect sizes that were less than 0.2 standard deviations were considered clinically equivalent. The 'range null' hypothesis testing procedure allows for a comparison between the effect size of the comparator group and a range of effect sizes, rather than the more traditional 'point null' test. The 'range null' procedure utilises a non-central *t* statistic when making these statistical comparisons (Serlin & Lapsley, 1985; 1993), and generates critical values that identify the effect size the community group must exceed to be considered either clinically equivalent to (or better than) the treatment benchmark or significantly better than the natural history benchmark. The R statistics program was used to perform calculations for the 'range null' hypothesis testing procedure, including identification of critical values.

A third set of analyses were conducted to establish treatment and natural history benchmarks for 'clinically significant improvement' (*CSI*) and to compare results of youth in the SOC CMHS subsets to these *CSI* benchmarks. Two-sample *z* tests for proportions were used for these comparisons. Consistent with the definition used in

published clinical trials, 'clinically significant improvement' was operationalized as falling from 'clinical' to 'subclinical' functioning on the CBCL-Int scale (Achenbach, 1991; Achenbach & Rescorla, 2001) – that is from CBCL-Int T >= 65 to CBCL-Int T < 65 (e.g. Cartwright-Hatton et al., 2011). This operationalization of 'clinically significant improvement' does not include a requirement for statistically reliable improvement, which has been used as part of the definition of 'treatment response' by other scholars (c.f. Jacobsen, Roberts, Berns, & McGlinchey, 1999). Further, CBCL-Int T = 65 is not identified as a cut off point for the CBLC-Int by the scales' developers (Achenbach & Rescorla, 2001). Nonetheless, this definition of 'clinically significant improvement' was used in the present research. This is because it mirrored definitions used within published studies and hence allowed for direct comparison of results of the SOC subsets with those of clinical research trials. The StarStat: Significance Testing Calculator (DataStat Inc., 1995-2011) was used to conduct the two-sample *z* tests for proportions.

A fourth set of analyses was conducted to identify 'good' and 'poor' treatment responders. Only the SOC_{CBCL} subset was used for this analysis, because of the small *n* in the SOC_{diag} subset. Youth were classified as 'good' treatment responders if they satisfied two conditions. First, youth whose pre-treatment scores fell from the 'deviant' (CBCL-DSM Anx T >= 65) to 'normal' (CBCL-DSM Anx T < 65) range of functioning on the CBCL-DSM Anx scale AND who demonstrated 'reliable change' on the CBCL-DSM Anx were considered 'good' treatment responders. Youth whose initial CBCL-DSM Anx score was in the 'deviant' range pre-treatment, and who did not meet these two criteria, were classified as 'poor' treatment responders. Use of the CBCL-DSM Anx T = 65 as a cut off was consistent with recommendations in Achenbach and Rescorla (2001)

for dichotomising samples into groups of 'deviant' or 'normal' functioning. Youth whose CBCL- DSM Anx scores improved by a statistically reliable amount were classified as 'reliable improvers'. Those whose scores did not, were classified as 'non-improvers'. 'Reliable change' was established using the 'Reliable Change Index' (RCI; Jacobson et al., 1999; Jacobson & Truax, 1991). Essentially, the RCI establishes a threshold for improvement that is beyond the measurement error of the instrument used - in this case, the CBCL-DSM Anx. Since Jacobsen and colleagues introduced the concept of 'reliable change', a number of methods for establishing the RCI have been developed (Wise, 2004). Wise (2004) reported that the five major methods he reviewed generated comparable results. This led him to conclude that the method recommended by Jacobson et al. (1999) should be the default approach because it is easier to understand than the other methods. This makes it a better choice for use in community agencies, where a straightforward calculation of the RCI is preferable. Thus, Jacobson's et al. (1999) method for calculation of the RCI was used in the present study, and was based on the following formula;

$$RCI_{CBCL\ DSM\ Anx} = \left(Pre_{CBCL-DSM\ Anx} - Post_{CBCL-DSM\ Anx}\right) \Big/ SD\sqrt{1 - r_{xx}} \quad (1)$$

where;

$RCI_{CBCL\ DSM\ Anx}$ = Reliable Change Index for the CBCL-DSM Anx DSM oriented scale

$Pre_{CBCL-DSM\ Anx}$ = Pre-treatment CBCL-DSM Anx raw score

$Post_{CBCL-DSM\ Anx}$ = 6month CBCL-DSM Anx raw score

SD = Pre-test standard deviation of CBCL-DSM Anx raw scores for SOC_{CBCL} youth in the 'clinical' range of CBCL-DSM Anx at pre-test = 1.77.

r_{xx} = test-retest reliability of the CBCL-DSM Anx (test-retest reliability = 0.80; Achenbach & Rescorla, 2001)

Participants whose scores improved beyond a RCI of 1.96 (i.e. $p < .05$) were considered to have demonstrated reliable change. In practice, this meant youth whose CBCL-DSM Anx raw score decreased by 1.55 or more during the six months after treatment initiation were considered to have demonstrated 'reliable' change.

The fifth major focus of data analysis was an examination of factors associated with treatment response. First, a logistic regression to examine prediction of 'treatment response' was planned. Logistic regression assesses the combined prediction of a dependent variable from a set of variables, as well as the unique prediction of each variable relative to the others. Thus, the unique predictive value of each variable is influenced, in part, by which other variables are included within the analysis (Tabachnik & Fidell, 2013). Various methods have been suggested for deciding the appropriate ratio of cases to predictor variables within logistic regression (e.g. Tabachnick & Fidell, 2013). Research by Peduzzi, Concato, Kemper, Holford & Feinstein (1996) indicated that the number of predictors should be limited by the size of the smallest of the two outcome groups, with a recommended ratio of approximately 1 predictor for every 15 participants in the smaller of the groups ('event to predictor;' ratio; Peduzzi, et al., 1996). The sample size (n = 101) meant that the maximum number of youth in the smaller of the groups ('good' versus 'poor' treatment responders) would be 50, and hence a maximum of three

variables could be used to predict treatment response. Poverty status, externalizing comorbidity and affective comorbidity were chosen as predictor variables, since they represent possible confounds impacting outcome of the SOC CMHS group (c.f. Westen et al., 2004). 'Poverty status' was operationalized using household income, as recorded in the DIQ, consistent with the definition used by the national evaluation study. Youth were classified as either 'living in poverty' (when household income was less than $15 000) or 'not living in poverty' (when household income was $15 000 or more). Externalising comorbidity was operationalized using DSM diagnosis, as recorded in the DIQ. Youth were classified as either 'comorbid externalising disorder' (having a DSM diagnosis of either ADHD or ODD at baseline) or 'no comorbid externalising disorder' (not having a DSM diagnosis of ADHD or ODD at baseline). Affective comorbidity was operationalized using DSM diagnosis, as recorded in the DIQ. Youth were classified as either 'comorbid affective disorder' (having a DSM diagnosis of a mood disorder at baseline) or 'no comorbid affective disorder' (not having a DSM diagnosis of a mood disorder at baseline). The three variables were entered into the logistic regression concurrently, since there was no a priori reason to enter them in separate steps.

Univariate analyses were also used to examine factors associated with treatment response. Unlike multivariate analysis, the relationship between factors in prediction of outcome is not taken into account in univariate analysis. However, it does allow for examination of a larger number of factors than does multivariate analysis (with correction for Type 1 error). These features of univariate analysis are particularly helpful in the present study, given that the investigation of factors associated with response to anxiety treatment in children is currently largely exploratory. For instance, some variables being

considered in the present research (e.g. child strengths; child functional impairment) have not been examined at all in previous studies (to the authors' knowledge). 'Good' and 'poor' treatment responders were compared on 11 variables related to demographics, family context, child strengths, child functional impairment and child psychopathology. Independent samples *t*-tests were used for comparisons of continuous variables and *Chi squared* tests were used for comparisons of categorical variables. Demographic variables were drawn from the DIQ questionnaire, and included age, gender, and ethnicity. Family context variables included the total number of family and child risk factors at baseline (from the DIQ), caregiver stress (total raw score on CSQG), general family functioning (total FAD-caregiver score) and poverty status (from the DIQ). Child strengths were measured using the BERS strength quotient. Child functional impairment across settings was measured using CAFAS total score. Child psychopathlogy was assessed using DSM diagnoses recorded in the DIQ. These were examined using univariate analyses to ascertain their relationship with treatment response, independent of other variables. Limited research on factors associated with treatment response in children with anxiety disorders meant a relatively large number of factors were examined. Because of the large number of analyses conducted, results were vulnerable to Type 1 errors. To reduce Type I error, a Bonferroni correction was applied to alpha, which was set at .005.

Lastly, a secondary analysis of moderators of effect sizes for the treatment trials was conducted using a *Q* test, which is comparable to an analysis of variance in primary research (Borenstein et al., 2009). A fully random effects model was used in the moderator analysis. The analysis examined seven possible moderators of effect size outcomes: treatment completion status ('completer' or 'intent to treat'); research setting

(community clinic or research); recruitment (referral only or both advertising and referral); method of intervention (in-person or via bibliotherapy, phone, internet, or email); method of delivery (group or individual), persons involved (parent, child, or both) and age of youth (7 or older or 6 or younger). Due to the number of analyses being conducted, a Bonferroni correction to the alpha was applied. The alpha was set at .007. Calculations for this analysis were performed using the CMA software (Biostat, 2006).

Pre-Analysis Preparations

Calculation of treatment efficacy benchmark. Results of clinical trials were aggregated to establish a pre-post effect size treatment efficacy benchmark. A single summary effect size from each study was used to contribute to the overall benchmark, so that studies with multiple subgroups did not have a disproportionate influence on the generation of the benchmark (Minami, Serlin, et al., 2008). The process of generating a single effect size per study progressed through a series of stages, depending on features of study design, such as the number of raters and treatment subgroups. In essence, the process involved aggregating effect sizes across parents within each subgroup, then across subgroups within each study, then effect sizes across studies, to generate the pre-post effect size treatment efficacy benchmark (Borenstein et al., 2009). Details of this process follow.

Calculation of pre-post effect size estimates. Consistent with previous studies (Minami, Serlin et al., 2008; Oei & Boschen, 2009; Westbrook & Kirk, 2005), effect size estimates were calculated by standardising change scores with the standard deviation (*SD*) of the pre-test group. The pre-treatment *SD* was used to standardise treatment gain, rather than the more commonly used pooled *SD*, because it meant that the estimated

effect size reflected the magnitude of change against the distribution of untreated youth. This made greater conceptual sense for the purposes of the present study than use of standardised gain scores (Lipsey & Wilson, 2001). Also, the true effect size of a treatment might be better estimated using the standard deviation of the pre-test score, rather than a pooled standard deviation, because it is not influenced by repeated measures and the impact of treatment (Morris, 2000).

Because small sample sizes can result in inflated effect size estimates, a correction for small sample size was utilised. Consistent with Minami, Serlin et al. (2008) and Curtis et al. (2009), each estimated effect size was multiplied by the correction for small sample size developed by Hedges (1981), 'c'. Thus, the general formula for calculating the corrected pre-post effect sizes estimate was as follows;

$$Y = \left[(M_{pre} - M_{post}) \Big/ SD_{pre} \right] * (c) \quad (2)$$

where;

Y = the corrected pre-post effect size estimate.

M_{pre} = the mean CBCL-Int T score at pre-treatment.

M_{post} = the mean CBCL-Int T score at post treatment.

SD_{pre}= the standard deviation of the pre-treatment CBCL-Int scores.

c = adjustment for small sample size = 1 - $\frac{3}{4n-5}$

n = sample size

The variance of this estimated effect size is as follows (Lipsey & Wilson, 2001);

$$\sigma_Y^2 = {[4(1-r) + Y^2]}/{2n} \quad (3)$$

where;

σ_Y^2= variance of effect size estimate

r = correlation between pre- and post- CBCL-Int scores = 0.45 (based on Bodden et al., 2008).

Y = corrected pre-post effect size estimate

n = sample size.

Aggregating effect size estimates across raters. In studies where both mother's and father's ratings were reported, the first step in generating a single effect size per study was to aggregate the effect sizes across parents' ratings. These outcomes were not independent (since they reflected the behaviour of the same child) and the aggregation process had to take the dependency into account so that the error in the estimate was not underestimated and the precision of the summary effect size was not overestimated (Borenstein et al., 2009). The composite parent effect size estimate ($\bar{Y}$) is the mean of mother's and father's effect sizes and the variance of this composite, was calculated as follows (Borenstein et al., 2009);

$$V_{\bar{Y}} = \left(\frac{1}{4}\right) \quad \left(V_j + V_k + 2r_{jk}\sqrt{V_j}\sqrt{V_k}\right) \quad (4)$$

where;

$V_{\bar{Y}} =$ variance of the composite effect size, $\bar{Y}$

$V_j =$ variance of rater j

$V_k =$ variance of rater k

r_{jk} = correlation between mother's and father's effect size estimates, estimated as r = 0.7, based on aggregation of results of studies measuring both.

Where the number of mothers and/or fathers in each (sub) group was not reported in the study or could not be established from contact with the authors, numbers were estimated using information within the trial (such as the number of mothers and fathers in the entire study or the degrees of freedom within analyses).

Aggregating effect sizes across subgroups within studies. The second step in generating a single effect size for each treatment study was to aggregate effect sizes across subgroups in studies that include multiple 'bona fide' treatments. In studies reporting 'bona fide' and experimental subgroups, only results of 'bona fide' subgroups were used. Results of all 'bona fide' treatments were aggregated (rather than being examined separately) because the focus of the present study was on generating a benchmark of any efficacious treatment, not on comparing the relative efficacy of different treatments. Composite effect sizes estimates (Y_s) were generated for each clinical trial, and are outlined in Table 3.

Aggregating effect size estimates across studies. Third, having combined effect sizes across raters and then across 'bona fide' treatment subgroups within studies, the next step was to aggregate each studies' summary effect size estimate (Y_s) to generate the treatment efficacy benchmark Y_{TE}. The summary effect size estimates for each study (Y_s) and the treatment efficacy benchmark (Y_{TE}) are outlined in Table 3.

Table 3

Study Treatment Effect Size Estimates, Treatment Efficacy Benchmark and Effect Size Estimates for SOC CMHS Subsets

Study	n[a]	*CBCL-Int* M_{pre} *(SD)*	*CBCL-Int* M_{post} *(SD)*	Y_s *(SE)*
Barrett et al. (1996)	53	-	-	1.14 (0.2)
Barrett (1998)	56	-	-	3.02 (0.4) *1.75*[b] *(0.4)*
Beidel et al. (2000)	30	68.4 (7.2)	60.2 (8.1)	1.11 (0.2)
Cartwright-Hatton et al. (2011)	34	66.6 (7.6)	59.4 (6.6)	0.92[a] (0.2)
Cobham et al. (1998)	67	-	-	0.96[a] (0.1)
Flannery-Shroeder and Kendall (2000)	25	-	-	1.68 (0.3)
Heyne et al. (2002)	57	-	-	1.28 (0.2)
Kendall (1994)	27	70.7 (7.0)	58.1 (10.3)	1.75 (0.3)
Kendall et al. (1997)	60	-	-	1.32 (0.2)
Kendall et al. (2008)	111	-	-	0.73[a] (0.1)
Lyneham and Rapee (2006)	78	-	-	1.07[a] (0.2)
Nauta et al. (2003)	76	71.5 (9.5)	61.7 (9.5)	0.90 (0.1)
Rapee et al. (2006)	180	-	-	0.65[a] (0.1)
Shortt et al. (2001)	48	-	-	6.91 (0.9)
Silverman et al. (1999a)	25	72.9 (7.6)	61.6 (8.4)	1.45 (0.3)
Silverman et al. (1999b)	65	-	-	0.74 (0.2)

Southam-Gerow et al. (2010)	15	66.5 (9.2)	58.9 (9.0)	0.79 (0.3)
Spence et al. (2006)	45	-	-	0.91 (0.2)
SOC_{CBCL}	101	73. 3 (6.6)	68.0 (11.2)	0.79 (0.11)
SOC_{diag}	70	68.4 (9.0)	63.7 (12.1)	0.52 (0.14)
Treatment efficacy benchmark[c]: Y_{TE} = 1.05 (*SE* = .08)				

Note. Dashes indicate single mean (*SD*) not available because of multiple informants or subgroups; *n* = Number of youth on which mean CBCL-Int is based, established from number of CBCL-Int questionnaires completed (where available) or number of youth whose data was reported; *CBCL-Int* M_{pre} = Mean pre-treatment CBCL-Int T score; *SD* = standard deviation of mean pre-treatment CBCL-Int T score; *CBCL-Int* M_{post} = Mean post-treatment CBCL-Int T score; Y_S = Study effect size estimate for each study; SE_S = standard error of effect size estimate.
[a]Effect size based on intent to treat data. [b]Winsorized effect size estimate [c]Treatment efficacy benchmark with results of Shortt et al. (2001) excluded from analysis and results of Barrett (1998) winsorized.

Within meta-analyses, studies with extreme outlying values can distort findings, leading to spurious or misleading conclusions (Lipsey & Wilson, 2001). There is some variability in practice, but outliers are commonly identified as those effect size estimates that are 2 or 3 standard deviations from the mean summary effect size (Lipsey & Wilson, 2001). Examination of the results listed in Table 3 revealed that Shortt et al. (2001) generated an effect size estimate that was substantially larger than any other study (Y_s = 6.9, *SE* = 0.9). Results of this study were so disparate from the others that it did not appear to fall within the same population of effect sizes and it was eliminated from further analysis (Lipsey & Wilson, 2001). The magnitude of the effect size estimate of the Barrett (1998) study (Y_s = 3.02) was almost 2 standard deviations higher than the

summary effect size that was generated after Shortt et al. (2001) was eliminated ($Y_{TE\ Barrett\ incl.}$ = 1.12). The next highest effect size was 1.75 (see Figure 2). Including such an unusually large effect size appeared inconsistent with the goal of generating realistic and representative treatment efficacy benchmarks for use in community agencies, but eliminating the study would lead to a loss of data. In order to limit the impact of outliers but minimise data loss, previous meta-analyses have 'winsorized' outcomes of studies with extreme results (e.g. Cook, Williams, Guerra, Kim, & Sadek, 2010; Crepaz et al., 2009; Derzon 2001; Durlak, Weissberg, & Pachan, 2010; Kobayashi, 2005; Malouff, Thorsteinsson, Rooke, Bhullar, & Schutte 2008; Shadish & Baldwin, 2005). 'Winsorizing' involves replacing the extreme values of outliers with more moderate ones. The next highest value in the distribution can be used as the replacement value (Lipsey & Wilson, 2001). Following this rationale, results of the Barrett (1998) study were 'winsorized' – and the study was assigned the same effect size estimate as the study with the next largest effect size (i.e. 1.75) (Lipsey & Wilson, 2001). With these changes made, synthesising results of the remaining treatment studies revealed an overall summary estimate effect size for all studies of Y_{TE} = 1.05 (SE_{TE} = 0.08). This value became the pre- post effect size treatment efficacy benchmark (see Table 3).

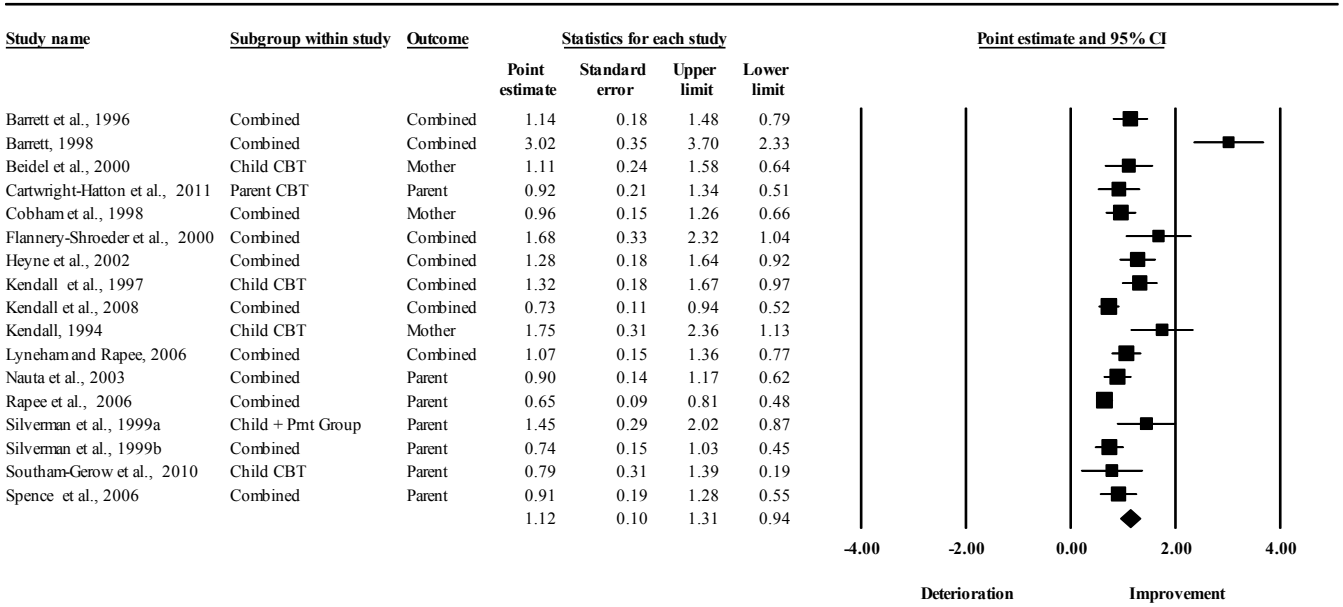

Figure 2. Study and summary effect size estimate(s) for treatment groups of clinical trials with Shortt et al. (2001) removed

Publications bias. Visual inspection of the pattern of distribution of effect size estimates illustrated on the funnel plot is consistent with publication bias (see Figure 3). As can be seen, there are four studies with large standard errors and relatively large effect sizes but no corresponding studies with similar standard errors but small effect size estimates. Duval and Tweedie's 'trim and fill' was used to impute values for these studies (see dark spots, Figure 3). Using this analysis, an estimated unbiased adjusted effect size of $Y_{TEadj} = .95$ was generated. Begg and Mazumdar's rank correlation test also yielded a significant result, indicating that the relationship between the magnitude of effect size estimates and standard errors of the effect sizes was significant, Kendall's $\tau = .54$; $p = .001$. Hence, results of the three tests were consistent with publication bias.

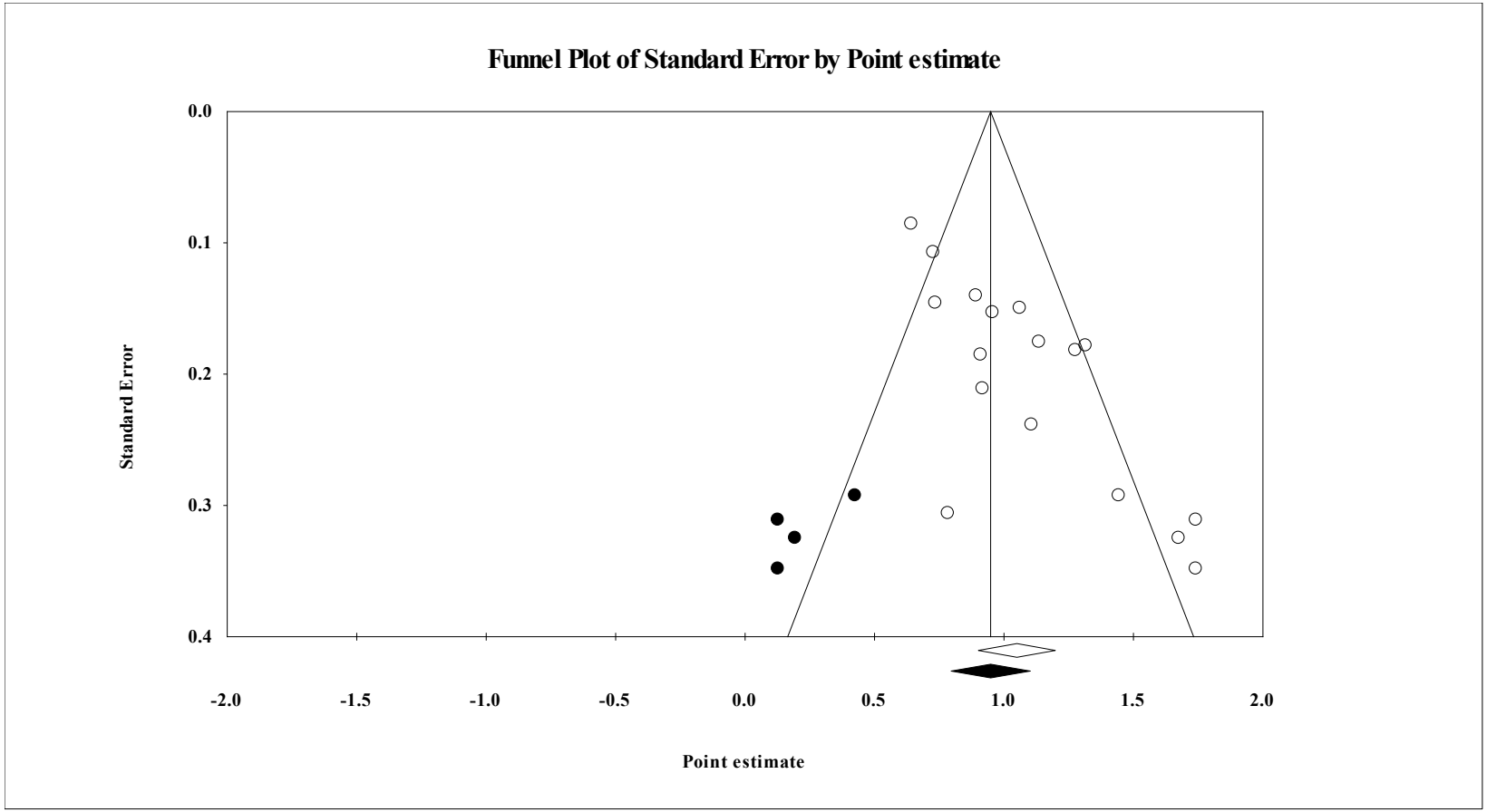

Figure 3. Funnel plot of standard error by effect size estimates for treatment groups of clinical trials, with effect size estimates of 'missing' studies imputed

Calculation of natural history benchmark. Consistent with Minami, Serlin et al. (2008), a 'natural history' effect size benchmark was established based on results of wait list control groups of the clinical trials. The 10 studies that used wait list control groups (i.e. Barrett et al., 1996; Barrett, 1998; Cartwright-Hatton et al., 2011; Flannery-Shroeder & Kendall, 2000; Kendall, 1994; Kendall et al., 1997; Rapee et al., 2006; Silverman et al. 1999a; Spence et al., 2006) were aggregated to establish this benchmark, which reflects the magnitude of improvement that might be expected with the passage of time alone. Mother's and father's ratings were synthesised using techniques identical to those within treatment groups (Borenstein et al., 2009). As can be seen in Table 4, most effect size estimates for wait list control groups fell between 0.2 — 0.5. Only one study showed deterioration in scores (Silverman et al.,1999a), and results of this study appeared somewhat disparate from the others (Figure 4). However, the difference between the effect size estimate of this study and the overall mean, did not approach 2 standard deviations, thus winsorizing was not considered (Lipsey & Wilson, 2007). The natural history effect size benchmark was calculated as Y_{NH} = 0.30; SE_{NH} =.07.

Table 4

Study Wait List Control Effect Size Estimates, Natural History Benchmark and Effect Size Estimates for SOC CMHS Subsets

Study	n[a]	CBCL-Int M_{pre} (SD_{pre})	CBCL-Int M_{post} (SD_{post})	Y_{WLC} (SE_{WLC})
Barrett et al. (1996)	23	-	-	0.38 (0.21)
Barrett (1998)	16	-	-	0.38 (0.25)
Cartwright-Hatton et al. (2011)	36	71.0 (7.5)	67.0 (9.4)	0.51 (0.18)
Flannery-Shroeder and Kendall (2000)	12	-	-	0.17 (0.18)
Kendall (1994)	18	71.6 (7.7)	69.3 (7.4)	0.28 (0.25)
Kendall et al. (1997)	34	-	-	0.18 (0.18)
Lyneham and Rapee 2006	22	-	-	0.10 (0.22)
Rapee et al. (2006)	87	68.4 (7.7)	65.1 (8.8)	0.42 (0.12)
Silverman et al. (1999a)	16	67.5 (9.1)	71.3 (6.8)	-0.40(0.27)
Spence et al. (2006)	23	68.7 (5.6)	66.8 (8.6)	0.33 (0.22)
SOC_{CBCL}	101	73. 3 (6.6)	68.0 (11.2)	0.79 (0.11)
SOC_{diag}	70	68.4 (9.0)	63.7(12.1)	0.52 (0.14)
Natural History effect size benchmark Y_{NH} = 0.30 $(SE_{NH}$ = .07)				

Note. Dashes indicate single mean not available because of multiple informants; N = Number of questionnaires used to establish mean (where available) or number of youth in group*;* *CBCL-Int* M_{pre} = Mean pre-waitlist CBCL-Int T score; *SDpre* = standard deviation of pre wait-list mean CBCL-Int T score; *CBCL-Int* M_{post} = Mean post wait list CBCL-Int T score; *SDpost* = standard deviation of post wait-list mean; Y_{WLC} = study effect size estimate for wait list control groups for each study*;* SE_{WLC} standard error of study wait list control estimated effect size.

Study name	Subgroup within study	Outcome	Statistics for each study			
			Point estimate	Standard error	Upper limit	Lower limit
Barrett et al., 1996	WLC	Combined	0.38	0.21	0.80	-0.04
Barrett, 1998	WLC	Combined	0.38	0.25	0.88	-0.12
Cartwright-Hatton et al., 2011	WLC	Parent	0.51	0.18	0.87	0.15
Flannery-Shroeder et al., 2000	WLC	Combined	0.38	0.32	1.01	-0.25
Kendall et al., 1997	WLC	Combined	0.17	0.18	0.53	-0.18
Kendall, 1994	WLC	Mother	0.28	0.25	0.78	-0.21
Lyneham and Rapee, 2006	WLC	Combined	0.10	0.22	0.53	-0.34
Rapee et al., 2006	WLC	Parent	0.42	0.12	0.65	0.20
Silverman et al., 1999a	WLC	Parent	-0.40	0.27	0.13	-0.93
Spence et al., 2006	WLC	Parent	0.33	0.22	0.77	-0.11
			0.30	0.07	0.44	0.16

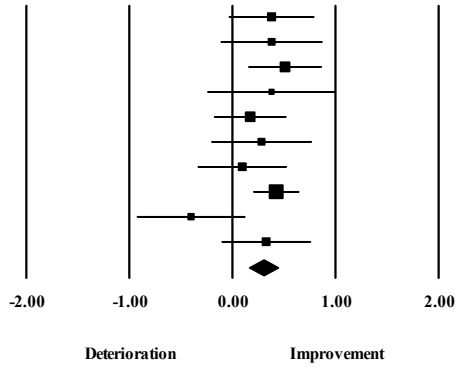

Figure 4. Study and summary effect size estimate(s) for wait list control groups of clinical trials

Calculation of SOC CMHS subset effect size estimates. The effect size estimates for each of the SOC CMHS subsets was calculated, using the same basic formula that was used to estimate the effect sizes within the clinical trials. Specifically, the effect size estimates were calculated as follows;

$$Y_{SOC} = \left[\left(M_{CBCL\ Int\ pre} - M_{CBCL\ Int\ 6mnth} \right) \Big/ SD_{CBCL\ Int\ pre} \right] * [c] \quad (5)$$

where;

Y_{SOC} = effect size estimate for SOC CMHS subset (adjusted for sample size)

$M_{CBCL\ Int\ pre}$ = the mean CBCL-Int T score at pre-treatment.

$M_{CBCL\ Int\ 6mnth}$ = the mean CBCL-Int Tscore six months post treatment-initiation.

$SD_{CBCL\ Int\ pre}$ = the pre-treatment standard deviation of the CBCL-Int T scores.

c = adjustment for small sample size (see formula 2).

Note that the second outcome measurement is taken six months post treatment-initiation, rather than directly post-treatment. As mentioned, this is because in community settings, the duration of treatment is inconsistent. Thus, the timing of the second outcome measurement is standardised by amount of time following treatment initiation, rather than by length of treatment. The six month time period encompasses the duration of all clinical trials included within the present study. The pre-treatment CBCL-Int T mean (*SD*) of youth in the SOC_{CBCL} matched subset was 73.3 (6.6) at pre-treatment and 68.0 (11.2) at six months post treatment initiation. Hence, when adjusted for sample size, the SOC_{CBCL} subset generated a pre-post estimated effect size of $Y_{SOC_{CBCL}} = 0.79$. The mean (*SD*) CBCL-Int T of the youth in SOC_{diag} subset was 68.4 (9.0) and 63.7 (12.1) at six

months. Hence, with adjustment for small sample size, the effect size estimate for the SOC_{diag} subset was $Y_{SOCdiag} = 0.52$.

The variance of the effect size estimate was calculated as follows (Lipsey & Wilson, 2001);

$$V_{SOC} = [4\ (1 - r_{SOC}) + Y_{SOC}^2] \Big/ 2n_{SOC} \quad (6)$$

where;

V_{SOC} = variance of the effect size estimate of the SOC subset

r_{SOC} = the pre-post correlation of CBCL-Int T scores for the SOC subset ($r_{SOC\ CBCL} = 0.49$; $r_{SOC\ diag} = 0.47$)

Y_{SOC} = the pre-post estimated effect size for the SOC subset ($Y_{SOCCBCL} = .79$; $Y_{SOCdiag} = .52$).

n_{SOC} = the sample size of the SOC subset ($SOC_{CBCL} = 101$; $SOC_{diag} = 70$)

Using this formula, the variance of the SOC_{CBCL} subset effect size estimate was $V_{SOC\ CBCL} = 0.01$ and the variance for the SOC_{diag} subset was $V_{SOC\ diag} = 0.02$.

Summary information regarding treatment efficacy benchmark, natural history benchmark and SOC subset effect sizes are summarised in Table 5 and their relative distributions are illustrated in Figure 5.

Table 5
Benchmark and SOC CMHS Subset Effect Size Estimates

Group	*N*	*K*	*Y*	V_Y	$CI_{.025}$	$CI_{.975}$
Treatment efficacy benchmark (Y_{TE})	1004	17	1.05	.006	0.90	1.20
Natural history benchmark (Y_{NH})	287	10	0.30	.005	0.16	0.44
SOC CMHS $Y_{SOC\ CBCL}$	101	-	0.79	.01	0.57	1.01
SOC CMHS $Y_{SOC\ diag}$	70	-	0.52	.02	0.26	0.77

Note. N = number of youth whose data were used to generate effect size estimate; K = Total number of studies; Y =Pre-post effect size estimate; V_Y = Variance of effect size estimate; $CI_{.025}$ = Lower confidence interval at 95 percent; $CI_{.975}$ = Upper confidence interval at 95 percentile

Study name	Point estimate	Standard error	Lower limit	Upper limit
Treatment efficacy benchmark	1.05	0.08	0.89	1.21
Natural history benchmark	0.30	0.07	0.16	0.44
SOC CMHS_CBCL	0.79	0.11	0.57	1.01
SOC_CMHS_diag	0.52	0.14	0.25	0.79

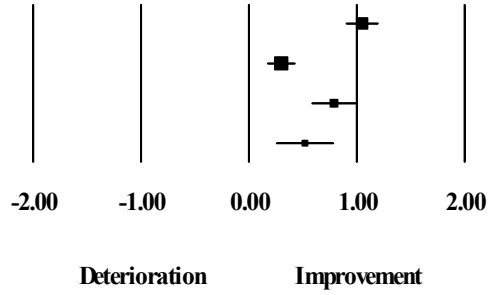

Figure 5. Treatment efficacy benchmark, natural history benchmark and effect size estimates for SOCcbcl and SOCdiag subsets

Calculation of benchmark for 'clinically significant improvement'. The proportion of participants falling from 'clinical' range on the CBCL-Int to 'subclinical' range has been reported in a minority of studies, with some using cut off scores of CBCL-Int T = 65 (Flannery-Schroeder & Kendall, 2000; Kendall et al., 2008) and others CBCL-Int T = 70 (Kendall, 1994; Kendall et al., 1997; Heyne et al., 2002; Silverman et al., 1999a; Silverman et al., 1999b; Spence et al., 2006) to differentiate 'clinical' and 'nonclinical' ranges. A cut off of CBCL Int T = 65 was chosen for the present study, since this was closer to the cut off of CBCL Int T score = 64 recommended by Achenbach and Rescorla (2001) for classification of clinical status (CBCL Int T = 60-63 'borderline' clinical range; CBCL-Int T >= 64 'clinical' range). Results of studies are reported in Table 6 based on information obtained from published results (k = 3) or from the authors (k = 5) for treatment groups (total: k = 8 studies). Results were aggregated across groups where studies reported results for multiple groups or raters (e.g. Flannery-Schroeder & Kendall, 2000; Heyne et al., 2002; Kendall et al., 2008; Spence et al., 2006).

Table 6

'Clinically Significant Improvement' (CSI) across Treatment Groups of Clinical Trials and 'Clinically Significant Improvement' Treatment Benchmark

	% *CSI*	*N* in 'clinical' range pre-treatment
Beidel et al. (2000)	66.7[a]	30
Cartwright-Hatton et al. (2011)	70.3[a]	34
Flannery-Schroeder and Kendall (2000)	77.8	19[b]
Heyne et al. (2002)	63.0[a]	46
Kendall et al. (1997)	55.3[a]	34
Kendall et al. (2008)	57.9[c]	38
Rapee et al. (2006)	44.2[a]	102
Spence et al. (2006)	60.7	28
'Clinically significant improvement' treatment benchmark CSI_{TE} = 57.5 (*SD* = 10.0)		

Note. CSI = 'Clinically significant improvement' i.e. proportion of youth who were in 'clinical' range of CBCL-Int pre-treatment (CBCL-Int T score >= 65) who fell to 'subclinical' range post-treatment (CBCL-Int T score < 65); *N* = number of youth in 'clinical range' of CBCL-Int pre-treatment (CBCL-Int T score >=65); *SD* = standard deviation of 'clinically significant improvement' benchmark.
[a]Based on data from author [b]Estimate based on information reported in publication
[c]Weighted mean across mother's and father's ratings.

Table 6 shows that only a minority of studies reported 'clinically significant improvement', and a weighted mean of CSI_{TE} = 57.5% (range: 44.2 - 77.8%) of youth moved from the 'clinical' range on the CBCL-Int at pre-treatment to 'sub-clinical' range post-treatment. This weighted mean will act as the 'clinically significant improvement' treatment benchmark (CSI_{TE} = 57.5%, *SD* = 10.0).

Table 7

'Clinically Significant Improvement' (CSI) for Wait List Control Groups of Clinical Trials and 'Clinically Significant Improvement' Natural History Benchmark

	% *CSI*	*N* in 'clinical' range pre-wait list
Cartwright-Hatton et al. (2011)	38.9[a]	36
Flannery-Schroeder and Kendall (2000)	20.5	9[b]
Kendall et al. (1997)	43.5[a]	23
Rapee et al. (2009)	32.3[a]	52
Spence et al. (2006)	10.0	20
'Clinically significant improvement' natural history benchmark CSI_{NH} = 31.9 (*SD* = 10.6)		

Note. *CSI* = 'clinically significant improvement' i.e. proportion of youth in wait list control group who were in 'clinical' range on the CBCL-Int pre-wait list (CBCL-Int T score >= 65) who fell to 'sub clinical' range post-wait list (CBCL-Int T score < 65). [a]Information supplied by author [b]Estimated from information in publication.

Table 7 illustrates that almost one third of youth in wait list control groups who were in the 'clinical' range on the CBCL-Int pre-waitlist fell to 'subclinical' range of functioning post wait-list. This weighted mean will be used as the 'clinically significant improvement' natural history benchmark (CSI_{NH} = 31.9%; *SD* = 10.6).

Calculation of 'clinically significant improvement' in SOC CMHS subsets. At pre-treatment, 89.1% (i.e. *n* = 90) of youth in the SOC_{CBCL} subset and 70.0% (i.e. *n* = 49) in the SOC_{diag} subset were in the 'clinical' range of the CBCL-Int (CBCL-Int T >= 65). Of these, $CSI_{SOC\ CBCL}$ = 23.3% (*n* = 21) and $CSI_{SOCdiag}$ = 30.6% (*n* = 15) fell to 'sub clinical' range six months after the start of services at SOC CMHS agencies.

Table 8

'Clinically Significant Improvement' in Benchmarks and SOC CMHS Subsets

Group	*N*	*K*	*CSI*	*SD*
Treatment CSI_{TE}	331	8	57.5%	10.0
Natural history CSI_{NH}	140	5	31.9%	10.6
SOC_{CBCL}	90	-	23.3%	-
SOC_{diag}	49	-	30.6%	-

Note. A dash indicates information not applicable. *N* = total number of youth at pre-treatment in 'clinical' range on CBCL-Int; *K* = total number of studies; *CSI* = 'clinically significant improvement' i.e. proportion of sample in 'clinical' range of functioning on CBCL-Int who fell to 'sub-clinical' range post- treatment or post-wait time; *SD* = standard deviation of *CSI*; CSI_{TE}= treatment benchmark for 'clinically significant improvement'; CSI_{NH} = natural history benchmark for 'clinically significant improvement'.

Hypothesis Testing

Testing the SOC CMHS subsets against the treatment efficacy benchmark.

First, the true effect size estimate of each SOC CMHS subset was tested against the true treatment efficacy effect size benchmark. The null hypothesis (H_0) was that the difference between the true treatment efficacy benchmark and the true effect size of the SOC subset was equal to or greater than the maximum margin allowed (i.e. $\Delta = 0.2$) to be considered clinically equivalent. The alternative hypothesis (H_1) was that the difference between the true SOC subset and the true treatment efficacy benchmark was less than the maximum allowed margin (i.e. $\Delta = 0.2$) (Serlin & Lapsey, 1985; 1993; Minami et al., 2009);

$$H_o : \delta_{B\ (TE)} - \delta_{Y\ (SOC)} \geq \Delta \quad (7)$$

$$H_1 : \delta_{B\ (TE)} - \delta_{Y\ (SOC)} < \Delta \quad (8)$$

where;

$\delta_{B\ (TE)}$= true treatment efficacy benchmark

$\delta_{Y(SOC)}$ = true effect size of SOC CMHS subset.

Δ = 0.2 (i.e. maximum difference between true effect size of SOC CMHS subset and true treatment efficacy benchmark allowed, for them to be considered clinically equivalent).

That is, to be considered clinically equivalent to the true treatment efficacy benchmark, the true SOC CMHS subset effect size couldn't fall more than 0.2 of a standard deviation below this benchmark (Serlin & Lapsley, 1985; 1993). The test statistic $t_{(TE)}$ for the hypothesis test follows a non-central t distribution, with degrees of freedom $(\nu) = N - 1$, where N = sample size of the SOC subset. The formula for the non-centrality parameter $(\lambda_{(TE)})$ for the treatment efficacy benchmark is as follows (Minami, Serlin, et al., 2008);

$$\lambda_{(TE)} = \sqrt{N}\ (\delta_{B\ (TE)} - \Delta) \quad (9)$$

where;

N = sample size of the SOC subset (SOC_{CBCl} = 101; SOC_{diag} = 70)

$\delta_{B\ (TE)}$= true treatment efficacy benchmark

Δ = 0.2

Using the 'range null' hypothesis testing procedure, a critical value ($Y_{CV\,[TE]}$) can be identified, which is the value the observed effect size for the community samples (Y_{SOC}) has to exceed to be considered clinically equivalent to the treatment efficacy benchmark (with an alpha of 0.05). If the observed effect size of the SOC subset (Y_{SOC}) exceeds this critical value, then the null hypothesis is rejected and the impact of the community treatment can be considered clinically equivalent (or better) than the treatment efficacy benchmark. If it does not exceed the critical value, this would indicate that there is no evidence that the effectiveness of treatment at SOC agencies is clinically equivalent to that delivered in clinical trials. The critical value is identified using the sample size (N) and non-centrality parameter (λ_{TE}). The formula for the critical value ($Y_{CV\,[TE]}$) is as follows;

$$Y_{CV\,[TE]} = [t_{(TE),\nu\ \lambda:.95}]/\sqrt{N} \quad (10)$$

where;

$t_{(TE)\nu,\lambda:.95}$ = the 95th percentile of the non-central t distribution

N = size of the SOC subset ($SOC_{CBCL;}$ = 101, SOC_{diag} = 70)

It should be noted that the smaller sample of the SOC_{diag} subset means the observed effect size is less reliable, hence the critical values are slightly more stringent for this subset.

Testing the SOC CMHS subsets against the natural history benchmark. The 'range null' hypothesis test was also used to compare the effect size estimates of each SOC CMHS subset with the natural history effect size benchmark. The strategy was the same as that used for testing against the treatment efficacy benchmark, except the signs in the null and alternative hypotheses were reversed. That is, the true effect size of the SOC

subset had to exceed the true natural history benchmark by more than 0.2 standard deviations to be considered significantly different and hence more clinically impactful than the passage of time alone. The null and alternative hypotheses can be stated as follows (Minami, Serlin et al., 2008; Minami et al., 2009);

$$H_o : \delta_{Y\,(SOC)} - \delta_{B\,(NH)} \leq \Delta \tag{11}$$

$$H_1 : \delta_{Y\,(SOC)} - \delta_{B\,(NH)} > \Delta \tag{12}$$

where;

$\delta_{Y\,(SOC)}$= true effect size of the SOC CMHS subset.

$\delta_{B\,(NH)}$= true natural history benchmark

Δ = 0.2 (minimum difference required to be considered significantly different).

The test statistic for this hypothesis also follows a non-central *t* distribution, with degrees of freedom (ν) = *N* -1. The formula for the natural history non-centrality parameter (λ_{NH}) is as follows (Serlin & Lapsley, 1985; 1993);

$$\lambda_{NH} = \sqrt{N}\ (\delta_{B\,(NH)} + \Delta) \tag{13}$$

where;

N = size of the SOC subset (SOC_{CBCl} = 101; SOC_{diag} = 70),

$\delta_{B\,(NH)}$= true natural history benchmark

Δ = 0.2

The formula for the critical value ($Y_{CV\,(NH)}$) is as follows (Serlin & Lapsley, 1985; 1993; Minami, Serlin et al., 2008);

$$Y_{CV\,(NH)} = t_{(NH)\nu,\lambda:.95}/\sqrt{N} \tag{14}$$

where;

$t_{(NH)\nu,\lambda:.95}$ = the 95th percentile of the non central *t* distribution

N = size of the SOC CMHS subset (SOC_{CBCL} = 101; SOC_{diag} = 70),

λ_{NH} = non-centrality parameter for the natural history benchmark.

If the observed pre-post effect size of the SOC subsets (Y_{soc}) exceeds this critical value ($Y_{CV\,(NH)}$), then the null hypothesis is rejected and it can be concluded that treatment effectiveness in SOC CMHS agencies is both clinically and statistically superior to the passage of time alone. If it does not exceed this critical value, this would indicate that there was no evidence that improvement in the SOC CMHS subset is more substantial than the passage of time alone. That is, there is no evidence that the impact of services received at SOC CMHS agencies is any greater than the natural history of symptom remission.

Results

Overview of Results Section

The results section reports findings of five sets of analyses. First, both SOC CMHS subsets were evaluated against the treatment efficacy effect size benchmark and the natural history effect size benchmark. These analyses were conducted to establish whether the magnitude of improvement of youth receiving treatment at SOC CMHS agencies was comparable to those of youth in treatment trials, or mirrored a magnitude of improvement commensurate with the passage of time alone. The pre-post effect size estimates for the SOC CMHS subsets were compared to the treatment efficacy and natural history benchmarks by establishing whether the difference between these groups'

effect sizes was larger than = 0.2 - that is, a difference considered clinically significant. This comparison was made using the 'range null' hypothesis testing procedure, developed by Serlin and Lapsley (1985; 1993). Second, the evaluation of each SOC CMHS subset against treatment and natural history benchmarks for 'clinically significant improvement' was considered. A two-sample *z* test for proportions was used to compare the proportion of youth in SOC CMHS subsets who evidenced 'clinically significant improvement' to those in treatment and wait list groups of treatment trials. Third, details of analysis of the 'completer' subsamples are outlined. Fourth, a logistic regression was conducted, examining the ability of 3 variables to predict reliable treatment improvement. Also, pre-treatment differences between 'good' or 'poor' treatment responders on 11 variables were examined. Lastly, a secondary analysis used a *Q* test to examine seven moderators of outcome within the clinical trials was conducted.

Hypothesis 1: Evaluating the SOC CMHS Data Against Pre-Post Effect Size Benchmarks

Evaluating SOC$_{CBCL}$ data against the treatment efficacy benchmark. First, the SOC$_{CBCL}$ subset effect size was tested against the treatment efficacy benchmark. The treatment efficacy benchmark was Y_{TE} =1.05. Taking reliability into account, the critical value for the treatment efficacy benchmark for the SOC$_{CBCL}$ subset was identified as $Y_{CV\ (TE\ SOC\ CBCL)}$ = 1.06 (see Table 9). Because the observed SOC$_{CBCL}$ subset effect size ($Y_{SOC\ CBCL}$ = .79) did not exceed this critical value ($Y_{CV(TE\ SOC\ CBCL)}$ = 1.06), it was concluded that there was no evidence that youth in the SOC_{CBCL} subset improved as much as youth in clinical trials who received ESTs, t (100) = 10.6, $\lambda_{TE} = 8.5$, $p > .05$. That is,

there was no evidence the effectiveness of treatment at SOC CMHS agencies could be considered clinically equivalent to that of ESTs.

Evaluating SOC$_{CBCL}$ data against the natural history benchmark. The SOC$_{CBCL}$ subset pre-post effect size was tested against the natural history effect size benchmark (Y_{NH}). The natural history benchmark was Y_{NH} = 0.30. Taking reliability into account, the critical value for the natural history effect size benchmark for the SOC$_{CBCL}$ subset was identified as $Y_{CV\ (NH\ SOC\ CBCL)}$ = .68. Because the observed SOC$_{CBCL}$ subset effect size ($Y_{SOC\ CBCL}$ = .79) exceeded the critical value for the natural history benchmark ($Y_{CV\ (NH\ SOC\ CBCL)}$ = .68), it was concluded that youth in the SOC$_{CBCL}$ subset improved significantly more than youth in wait list control groups of published clinical trials, t (100) = 6.9, $\lambda_{NH} = 5.0, p < .05$ (see Table 9). That is, the magnitude of improvement in the SOC$_{CBCL}$ subset was greater than would be expected from the passage of time alone.

Evaluating SOC$_{diag}$ data against the treatment efficacy benchmark. Next, the SOC$_{diag}$ subset effect size was tested against the treatment efficacy benchmark (Y_{TE}). As mentioned, the treatment efficacy benchmark was Y_{TE} =1.05. Taking reliability into account, the critical value for the treatment efficacy benchmark for the SOC$_{diag}$ subset was identified as $Y_{CV(TE\ SOC\ diag)}$ = 1.10 (see Table 9). Because the observed SOC$_{diag}$ subset effect size ($Y_{SOC\ diag}$ = .52) did not exceed the critical value of the treatment efficacy benchmark ($Y_{CV(TE\ SOC\ diag)}$ = 1.10), it was concluded that there was no evidence that the improvement of youth in the SOC$_{diag}$ subset was clinically equivalent to that of youth in clinical trials receiving ESTs, t (69) = 9.2, $\lambda_{TE} = 7.1, p > .05$.

Evaluating SOC_{diag} data against the natural history benchmark. The SOC_{diag} subset effect size was tested against the natural history benchmark (Y_{NH}). As mentioned, the natural history benchmark was $Y_{NH} = 0.30$. Taking reliability into account, the critical value for the natural history effect size benchmark for the SOC_{diag} subset was identified as $Y_{CV(NH\,SOC\,diag)} = .72$ (see Table 9). Because the SOC_{diag} subset effect size estimate ($Y_{SOC\,diag} = .52$) did not exceed this critical value ($Y_{CV(NH\,SOC\,diag)} = .72$), it was concluded that there was no evidence improvement of youth in the SOC_{diag} subset was any greater than that of youth in wait list control groups, $t\,(69) = 6.0, \lambda_{NH} = 4.2, p > .05$ (see Table 9). That is, improvement of youth in the SOC_{diag} subset could be considered clinically equivalent to the natural remission of anxiety symptoms.

Table 9

Effect Size Estimates of SOC CMHS Subsets Tested Against Critical Values of Treatment Efficacy and Natural History Benchmarks.

			Treatment Efficacy				Natural History			
Subset	*N*	Y_{SOC} *(SE)*	$Y_{CV[TE]}$	*t*	ν	λ_{TE}	$Y_{CV[NH]}$	*t*	ν	λ_{NH}
SOC_{CBCL}	101	0.79 (0.11)	1.06	10.6	100	8.5	0.68	6.9	100	5.0
SOC_{diag}	70	0.52 (0.14)	1.10	9.2	69	7.1	0.72	6.0	69	4.2

Note. *N* = number of youth in subset; Y_{SOC} = observed effect size estimate for subset; *SE* = standard error of effect size estimate; $Y_{CV[TE]}$ = critical value for treatment efficacy benchmark; *t* = non-central *t* test statistic; ν = degrees of freedom; λ_{TE} = noncentrality parameter for treatment efficacy benchmark; $Y_{CV[NH]}$ = critical value for natural history effect size benchmark; λ_{NH} = non-centrality parameter for natural history effect size benchmark.

Evaluating SOC CMHS subsets against unbiased adjusted effect size estimate. If the adjusted unbiased effect size estimate ($Y_{TEadj} = .95$) was used as the treatment efficacy benchmark, both the SOC_{CBCL} subset, $t\,(100), 9.54, \lambda = 7.54, p > .05$ and the

SOC_{diag}, t (69) = 8.31, λ = 6.27, $p > .05$ would still fail to exceed the critical value for the treatment efficacy benchmark. Thus, use of this adjusted unbiased effect size estimate would not alter conclusions of the study and was not used as the treatment efficacy benchmark.

Hypothesis 2: Evaluating SOC CMHS data against 'clinically significant improvement' benchmarks.

The proportion of youth in the SOC_{CBCL} and SOC_{diag} subsets who fell from the 'clinical' to the 'subclinical' range on the CBCL-Int was compared to both treatment and natural history *CSI* benchmarks (CSI_{TE} and CSI_{NH}, respectively) (see Table 8). Results revealed that compared to youth in the treatment groups of clinical trials, there were significantly fewer youth in the SOC_{CBCL} subset, $z = 6.47$, $p < 0.001$ and in the SOC_{diag} subset, $z = 3.7$, $p < .001$ who evidenced 'clinically significant improvement'. Further, there was no significant difference between the natural history *CSI* benchmark and the proportion of youth who showed 'clinically significant improvement' in the SOC_{CBCL} subset, $z = 1.36$, $p = 0.07$ nor in the SOC_{diag} subset, $z = .17$, $p = 0.43$. That is, the proportion of youth evidencing 'clinically significant improvement' in both the SOC_{CBCL} and SOC_{diag} subsets was significantly lower than treatment groups and not significantly different from wait list control groups in clinical trials.

Hypotheses 1 and 2: Evaluating SOC CMHS 'completer' data against benchmarks.

As noted in the method, the full SOC CBCL and SOCdiag subsets were used in the present study, since conclusions from these full subsets did not differ from those of the 'completer' subsamples. For the $SOC_{CBCL\ completer}$ subset, pre-post effect size was 0.83 (compared to 0.79 for the full SOC_{CBCL} subset). Because of smaller sample size and

hence reduced reliability, the critical value for the treatment efficacy benchmark for the $SOC_{CBCL\ completer}$ subset was $CV_{TE\ SOC\ CBCL\ cmpltr} = 1.12$ and the critical value for the natural history benchmark for the $SOC_{CBCL\ completer}$ subset was $CV_{NH\ SOC\ CBCL\ cmpltr} = .74$. Hence, the $SOC_{CBCL\ completer}$ subset, like the full SOC_{CBCL} subset, was significantly greater than the natural history benchmark, $t\ (62) = 5.19$, $\lambda = 3.28$, $p < .05$, but not the treatment efficacy benchmark, $t\ (62) = 8.89$, $\lambda = 6.74$, $p > .05$. Further, the proportion of youth evidencing 'clinically significant improvement' (*CSI*) in the $SOC_{CBCL\ completer}$ subset was 17.5% (compared to 23.3% for the full SOC_{CBCL} subset). Hence, like the full SOC_{CBCL} subset, the $SOC_{CBCL\ completer}$ subset was not greater than the *CSI* natural history benchmark (*CSI* natural history benchmark = 31.9%) and significantly lower than the treatment *CSI* benchmark (*CSI* treatment benchmark = 57.5%), $z = 7.22$, $p < .001$. Thus, conclusions of analysis of the $SOC_{CBCL\ completer}$ subset did not differ from those of the full SOC_{CBCL} subset. Hence, the full SOC CMHS subsets were used in the present study.

For the $SOC_{diag\ completer}$ subset, pre-post effect size was 0.57 (compared to 0.52 for the full SOC_{diag} subset). The treatment efficacy critical value was $CV_{TE\ SOC\ diag\ cmpltr} = 1.19$ and the natural history critical value was $CV_{NH\ SOC\ diag\ cmpltr} = 0.80$. Thus, like the full SOC *diag* subset, there was no evidence that the $SOC_{diag\ completer}$ subset surpassed the natural history benchmark critical value, $t\ (41) = 5.15$, $ncp = 3.24$, $p > .05$. Further, for this subset, the proportion of youth evidencing *CSI* was 33.3% (compared to 30.6% for full SOC_{diag} subset). There was no significant difference between this value and the *CSI* natural history benchmark, $z = 0.14$; $p = 0.44$. Thus, conclusion of analysis of the SOC $_{diag\ completer}$ subset did not differ from those of the full SOC_{diag} subset.

Hypotheses 3, 4 and 5: Factors Associated with Treatment Response

Of the 101 youth in the SOC_{CBCL} subset in the 'deviant' range of functioning on the CBCL-DSM Anx at baseline (i.e. CBCL-DSM Anx $T >= 65$), 21 (20.8%) were classified as 'good' treatment responders, while the majority (79.2%, $n = 79$) were classified as 'poor' treatment responders ($n = 1$, not classified due to missing data). That is, only a minority of the SOC CMHS youth fell from 'deviant' to 'normal' range of functioning on the CBCL-DSM Anx AND improved by a statistically reliable amount (i.e. CBCL DSM Anx raw score reduced by 2 or more). However, of the 101 youth in the 'deviant' range pre-treatment, 48.5% ($n = 49$) demonstrated 'reliable' improvement and 21.8% ($n = 22$) were in 'normal' range of functioning on the CBCL-DSM Anx post-treatment (i.e. CBCL-DSM Anx $T < 65$).

With so few youth classified as 'good' treatment responders, it was not possible to conduct a logistic regression, since a sample of this size would only allow for one predictor of outcome (c.f. Pelluzzi et al., 1996). For this reason, a logistic regression was conducted using 'reliable improvement status' as an alternative dependent variable (since there were 49 'responders' identified). 'Reliable improvement status' was not the first choice as a dependent variable, since it did not incorporate the clinically relevant component of 'falling to normal range' of internalising symptomatology within its operationalization. However, using this less conservative operationalisation of response to treatment allowed a multivariate analysis of factors associated with a positive treatment outcome,

Before the logistic regression was conducted, data were screened for accuracy of data entry, missing values and multivariate outliers. Frequency plots were used to

examine data entry accuracy; no apparent errors were detected. Data for co-morbid externalizing diagnosis and comorbid affective diagnosis status were available for all youth. A small proportion of participants were missing information regarding poverty status (3.0%, $n = 3$) and reliable improvement status (1.0%, $n = 1$). No significant patterns in the variables used within the logistic regression were identified using Missing Values Analysis in SPSS (Tabachnick & Fidell, 2013). Because there were relatively few cases with missing data, and no apparent pattern to their missingness, these were deleted from the logistic regression analysis. Multivariate outliers were examined using Cook's distance (D) (Tachnick & Fiddell, 2013). Cook's D is a measure of the influence of particular cases on an analysis, and values greater than 1 may be cause for concern (Cook & Weisberg, 1982). None were detected.

Thus, following deletion of cases with missing information, the logistic regression was conducted using data from 97 youth (47 'reliable improvers' and 50 'non improvers'). Results of the logistic regression indicated that the model was not significant, *Chi square* = 0.40 (3), $p = .94$ and accounted for almost no variance in outcome, *Nagelkerke R squared* = 0.6%. None of the variables were significant unique predictors of reliable improvement status in the context of the others. With the full model, 53.6% of cases were correctly classified using these variables, which only improved base rate classification (51.5%) by a marginal amount.

Assumptions underlying logistic regression were tested. There was no evidence of multi-collinearity among predictor variables, based on Tolerance values and results of collinearity diagnostic tests (range of Tolerance values .97-.99; variance proportions for dimension with smallest eigenvalue = .01, .08; 94) (Field, 2009; Tabachnick & Fiddell,

2013). Standardised residuals (*z*) were used to examine evidence for outliers in the logistic regression solution, with scores larger than 3 indicative of a cause for concern (c.f. Field, 2009; Tabachnick & Fiddell, 2012). None were detected. Lastly, examination of independence of errors using methods outlined by Field (2009) revealed no evidence of over dispersion, since the dispersion parameter (Φ) was less than one.

Results of univariate analyses comparing 'good' and 'poor' treatment responders revealed there were no significant differences on any of the 11 variables examined in, t (98) = -1.94 to 1.55, *ns*; range of *Chi square* (1) = -1.00 to 0.62, *ns*. Further, even without a Bonferroni correction to alpha, there were no significant differences between 'good' and 'poor' responders on any of the 11 variables examined.

Moderators of Effect Size Within Clinical Trials

There was significant heterogeneity within studies' summary effect size estimates, Q (16) = 48.2, $p < .001$. Secondary analysis was conducted to examine sources of heterogeneity using a Q test. When a Bonferroni correction was applied (alpha set at .006), no significant moderators of outcome were identified. One moderating effect approached significance. There was a non-significant trend for outcomes of studies reporting 'ITT' data to differ from those reporting 'completer' data, $Q\ (1_{bet})$ = 7.0, p = 0.008. The six treatment trials reporting 'intent to treat' data generated marginally lower effect size estimates than the 11 trials reporting data from 'completer' samples mean (mean Y_{ITT} = .85, *SE*= .09; mean $Y_{completer}$ = 1.20, *SE* = .09 for ITT and completer trials respectively). There were no significant differences in effect size estimates in treatment trials based on research setting, $Q\ (1_{bet})$ = .05, p = .83; recruitment, $Q\ (1_{bet})$ = 0.0, p = .88; mode of delivery, $Q\ (1_{bet})$ = 1.10, p = .30; persons involved, $Q\ (2_{bet})$ = 2.2, p = .34;

method of delivery, Q (1_{bet}) = 1.71, p = .43; or minimum age of youth, Q (1_{bet}) = 22, p = .64. That is, there was no evidence that studies conducted in research settings (as opposed to the community), where recruitment included advertising (rather than referral alone), where therapy was conducted in person (rather than via technology) had better outcomes. Further, there was no clear advantage for treatment to be delivered in a particular format (individual or group), with particular people involved (parent, child or both) or with children in a certain age range (6 years or less; 7 years or more).

Discussion

The aim of the present research was threefold. Firstly, to generate benchmark standards for pre-post effect sizes and rates of 'clinically significant improvement' that can be used to evaluate outcomes of treatment of anxiety disorders in youth. These benchmarks were established from results of clinical trials and based on a common broadband measure of psychopathology (the CBCL-Int/6-18). Given the increasing emphasis on implementation of EBP in applied settings, the present study was an attempt to bridge the gap between research and clinical practice by establishing standards for treatment effectiveness that are accessible and useful for treatment agencies in the community.

The second objective of the present study was to examine outcomes of youth with anxiety disorders treated at SOC CMHS agencies against these benchmarks. It is important to evaluate the impact of UC within SOC CMHS services, particularly given previous findings that UC is sometimes no more effective than natural remission in treating psychopathology, and that ESTs can be implemented in community agencies with success that is comparable to that of published efficacy trials.

The third objective was to identify 'good' and 'poor' treatment responders within youth treated at the SOC CMHS agencies, and to attempt to identify factors associated with improvement or treatment response in these youth. Having an empirically based understanding of what factors are associated with outcome can inform understanding for particular populations, including youth seen at SOC CMHS agencies. For instance, understanding factors associated with treatment response or reliable improvement can help tailor case management (by attempting to influence variables associated with treatment response), prevent treatment failure, and ultimately enhance treatment success. Further, it can provide information regarding what variables, though intuitively appealing, are not associated with treatment response, and may be less valuable to focus intervention upon. Lastly, it can inform understanding of mechanisms influential in the development and maintenance of anxiety disorders (Kendall, Settipani, & Cummings, 2012).

The first part of the discussion will address each of these objectives and will then explore issues secondary to the main findings that emerged during the course of the research.

Treatment Efficacy and Natural History Benchmarks for Pre-Post Effect Sizes

Attempts were made to be appropriately conservative when generating effect size estimates for each study, including adjusting for small sample size and making allowances for dependency between outcomes. Further, including both father's and mothers' outcome ratings, while increasing the statistical complexity of analysis, meant both perspectives were taken into account when generating estimates of effect sizes. Lastly, eliminating one study (Shortt et al., 2011) and 'winsorizing' the results of another

(Barrett, 1998) also contributed to a more conservative treatment efficacy benchmark. Consistent with recommendations (Higgens et al., 2011; Lipsey & Wilson, 2001), quality ratings were not used to weigh the contribution of studies to the summary mean effect size. However, inclusion/ exclusion criteria were adopted that meant a minimum standard of rigor in design was required for studies to be included in the analysis. Further, the influence of features of research design that might exert a systematic influence on results (e.g. ITT versus completer analysis) were examined.

There was significant heterogeneity within studies' effect size estimates although most were larger than 0.8. Secondary analysis was conducted to examine sources of heterogeneity within the treatment efficacy studies – that is, possible moderators of outcome in the clinical trials. After a Bonferroni correction was applied, no significant moderators of outcome were identified. However, treatment 'completer' status approached significance. Trials reporting 'intent to treat' data generated marginally lower effect size estimates than those reporting data from 'completer' samples ($p < .008$). This trend is consistent with previous research which has found effect sizes generated from studies reporting 'intent to treat' data are lower than those from studies reporting results of 'completer' samples (Eddy et al., 2004; Westen & Morrison, 2001). The finding could reflect the impact of common research methodology, where the baseline symptom scores of treatment drop-outs are carried forward and used as an estimate of their post-treatment symptoms. Hence (by definition) these participants do not improve and the treatment effect sizes from these studies are therefore likely to be lower than those reporting 'completer' samples. Given that in the present study, even youth in wait list control groups improved by a moderate amount over time, adopting this strategy might yield an

excessively conservative result for drop-outs from anxiety treatments. It does reflect the importance, however, of considering attrition when evaluating the outcomes of community samples and of considering the nature of data reported (ITT or completer) when aggregating studies to form benchmarks.

Treatment setting (research versus community), minimum age of participant (7 years or older versus 6 years or younger), delivery format (group versus individual), parent involvement (parent versus child versus both), delivery mode (in person versus via technology), recruitment method (referral only versus both referral and advertising) were not significant moderators of the outcome. Further, none of these moderators would have been significant, even without a Bonferroni correction to alpha. These findings are consistent with previous research, which found effectiveness studies achieved comparable results to efficacy (Hunsley & Lee, 2007), and that parent involvement and treatment delivery format were not significant moderators of outcome (Liber et al., 2010; Reynolds et al., 2012; Silverman, Pina et al., 2008). Results were inconsistent with some previous findings. For instance, previous research has found that effect sizes from samples recruited via advertising were larger than for those recruited from referral (Brent et al., 1998). This might be because studies that recruited exclusively through advertising were excluded from the present research and, as a result, the potential impact of recruitment method may have been diluted.

Heterogeneity of effect size estimates in the treatment trials highlights the importance of using a larger number of studies to generate benchmarks, since any single study is unlikely to represent the range of possible outcomes. Compared to previous benchmarking research (e.g. Curtis et al., 2009; Farrell et al., 2010) the present study

used far more studies to generate the benchmark (k =17), improving confidence in the reliability of the result.

The treatment effect size benchmark generated from the present research, while using a broadband measure of psychopathology, is comparable in magnitude to the summary mean effect size of a recent meta analysis of treatments of anxiety disorders in youth (In Albon & Shneider, 2006), even though this meta-analysis reported pre-post effect sizes generated from a symptom-specific measure of anxiety. Analysis of findings of clinical trials indicated there was some evidence of asymmetry in the distribution of effect sizes (in relation to standard errors). There were a disproportionate number of studies with small samples, but relatively large effect sizes, without a commensurate number of studies with small samples and small effect sizes. It is possible that this asymmetry was caused by publication bias. That is, it is possible that studies with small samples were less likely to be published if they had small effect sizes than if they had large ones. If the asymmetry was due to publication bias, analyses suggest that the best estimate of the unbiased treatment efficacy effect size composite was $Y_{adjRCT} = 0.95$. Using the adjusted effect size would not alter conclusions of the present study (since the effect size estimates of the SOC subsets still would not surpass this benchmark), but does demonstrate the importance of taking the possibility of publication bias into account when generating benchmarks. When attempting to interpret results of tests of publication bias, alternative reasons to account for findings (including heterogeneity) can be investigated. This includes considering systematic methodological differences between smaller and larger studies that might account for sources of heterogeneity in the distribution of effect size estimates (Sterne, Gavaghan & Egger, 2000). While only one

marginally significant moderator was established to account for heterogeneity of effect size estimates in the present study (i.e. completer versus ITT status), future research could continue to examine possible sources of heterogeneity in study outcomes. A greater number of studies will allow examination of factors that interact to systematically influence outcome and hence heterogeneity. Several strategies have been suggested to attempt to minimise publication bias in research. These include requiring that clinical trials are registered before they commence, in order to adequately track the proportion of studies that are published and to decrease the possibility that only studies with significant findings are published (Ioannidis, 2005). Further strategies to decrease publication bias include becoming more open to publishing null or non significant findings and having minimum sample size requirements (Ioannidis, 2005). Using minimum sample size requirements would reduce the likelihood of studies with null findings being withheld from publication due to inadequate power.

It was notable that even youth who were in wait list control groups (on average) improved by a moderate amount (summary effect size = 0.30). This finding highlights two issues. First, when evaluating outcomes of community agencies, it is important to be aware that youth with anxiety disorders are likely to experience moderate improvement with the passage of time alone. Results show that even a reasonable amount of improvement might be considered clinically equivalent to natural remission. The second issue relates to the first. The results show that when evaluating treatment of youth with anxiety disorders in community agencies, improvement per se is not enough to demonstrate adequate impact. A fairly substantial magnitude of improvement is required

to show improvement superior to wait list control groups and an even more substantial one is required to be considered clinically equivalent to EST treatments.

The second major benchmark standard considered in the present study was 'clinically significant improvement' (*CSI*) – operationalized as the proportion of the sample in the 'clinical' range on the CBCL-Int at pre-treatment who moved to 'subclinical' range at post-treatment. While this improvement does not take measurement error into account, generating a benchmark based on this information was helpful because this was the way CSI was reported in published trials and this could therefore be compared directly to results of treatment trials. Further, this operationalization of *CSI* is an easily understood and calculated measure of improvement, making it potentially more helpful in community settings. Fewer trials contributed to these benchmarks than the pre-post effect size benchmarks, meaning they may be less reliable. Ultimately, aggregating results of eight studies showed that a mean of 57.5% of youth in treatment groups within RCTs fell from 'clinical' to 'sub-clinical' functioning post treatment. Thus, even when youth receive ESTs, only a little over half of the group can be expected to fall to the 'sub-clinical' range of functioning on the CBCL-Int post-treatment. The relatively modest proportion of youth evidencing 'clinically significant improvement', even after receiving efficacious treatments, suggests that *CSI* may be a less sensitive measure of change than pre-post effect sizes. Nonetheless, this metric of outcome may be more accessible and clinically relevant to community services than effect sizes, and should be considered when evaluating agencies' results.

Data Reduction

As can be seen in Figure1, the ultimate samples of SOC_{CBCL} n = 101 or SOC_{diag} n = 70 were considerably reduced from the 4500 or so youth in the full age matched SOC CMHS longitudinal data set. Further, these (relatively) small subsets could have been reduced even more, by using additional inclusion criteria (such as treatment completion; or including only youth treated within the mental health treatment sector). The ultimate size of the subsets was similar to previous benchmarking studies of youth (e.g. Weersing et al., 2006) and the reduction partly reflects the fairly stringent inclusion and exclusion criteria that were used to identify the subsets. While using these criteria reduced the size of the subsets, it also increases confidence in the validity of comparison between results of SOC CMHS youth and those of participants in clinical trials. Another reason for the substantial reduction in sample size may be because of the nature of the agencies contributing information to the national evaluation SOC CMHS data set. A substantial portion of youth in the SOC CMHS national evaluation study were served within agencies in the special education, child welfare and youth justice sectors. Youth served in these sectors would be disproportionately affected by the exclusion criteria of the present study - such as excluding youth with a diagnosis of a developmental disability or with severe externalising behaviour. This could account for the large reduction in the proportion of the sample eligible for participation in the present study and is not necessarily a reflection of the generalizability of treatment conditions within clinical trials compared to community mental health outpatient settings.

A third possible reason for the large reduction in sample size relates to the challenge of matching SOC CMHS clients to youth treated within clinical trials,

particularly with respect to clinical profile. Only some of the youth with 'borderline' or even 'clinical' range scores on the CBCL-DSM Anx had a clinician-designated DSM diagnosis of an anxiety disorder (see Table 1). The low prevalence of youth with identified anxiety diagnoses in the SOC_{CBCL} subset (even amongst those with substantially elevated symptoms of anxiety and/or a presenting problem of anxiety) may reflect a tendency for clinicians to under-diagnose anxiety and other internalising disorders in youth (Richardson, Russo, Lozano, McCauley, & Katon, 2010). Previous research has found that youth are less likely to be diagnosed with an anxiety disorder by community clinicians than when they are assessed by researchers using standardised diagnostic tools (Jensen & Weisz, 2002; Rettew et al., 2009). This may be because the externalising problems are more obvious or disruptive and hence more likely to be identified by community clinicians. Alternatively, there may be other factors that influence community-generated diagnoses such as insurance or availability of resources for particular disorders but not others. For instance, there may be practitioners available to prescribe medication for ADHD but not to deliver CBT for anxiety disorders.

Selection of participants with significant problems related to anxiety was a challenging process. Selection by anxiety diagnosis (i.e. the SOC_{diag} subset) yielded a smaller sample and was likely less psychometrically reliable (c.f. Jensen & Weisz, 2002) than selection by CBCL profile. However, pre-treatment mean and *SD* of CBCL-Int scores in this subset were more comparable with those of clinical trials (see Table 2), making interpretation of results somewhat easier. Selection based on a combination of diagnosis, presenting problem and CBCL-DSM Anx profile (i.e. SOC_{CBCL}) was (likely) more psychometrically sound, but confounded interpretation of results due to a higher

pre-treatment mean CBCL-Int and a smaller pre-treatment standard deviation than those seen in clinical trials (see Table 2). The implications of these issues will be discussed in the sections to follow, in the context of results of the benchmarking process.

Hypothesis 1: Comparison of SOC CMHS Subsets with Treatment Efficacy and Natural History Benchmarks.

The present study made use of the 'range null' hypothesis testing procedure to compare the SOC subsets to effect size benchmarks (Serlin, 1975; 1983). This meant that evaluation of results from the SOC agencies could incorporate consideration of both statistically and clinically meaningful outcomes, which is of particular relevance to community agencies.

Comparison of each SOC subset with pre-post benchmarks yielded somewhat different results. The SOC_{CBCL} subset evidence moderate gains that surpassed the critical value of the natural history benchmark but not the treatment efficacy benchmark. That is, there was no evidence that gains in this subset were as substantial as might be expected after receiving ESTs, although treatment gains were significantly better than natural remission. Similar to the SOC_{CBCL} subset, improvement in youth in the SOC_{diag} subset also failed to exceed the critical value for the treatment efficacy benchmark. However, the magnitude of improvement of youth in the SOC_{diag} subset did not even surpass the critical value of the natural history benchmark (Table 9). Hence, there was no evidence that the impact of SOC services for this subset was clinically any different from what might be expected from natural remission of anxiety symptoms (Minami, Serlin et al., 2008). The magnitude of improvement in the SOC_{diag} youth was comparable to that seen in large-scale evaluations of UC (Trask & Garland, 2012). Further, it is consistent with

previous research that has found UC in community settings is associated with a similar degree of improvement as that seen within wait list control groups (Weersing & Weisz, 2002).

Thus, inconsistent with hypothesis 1, neither subset achieved improvement as substantial as that evidenced in youth who received ESTs for anxiety disorders in RCTs. Improvement in SOC_{CBCL} youth was superior to wait list control groups, whereas there was no evidence the SOC_{diag} subset was more impactful than natural remission. These conclusions were also true for the 'completer' subsamples of the SOC CMHS subsets.

There are several reasons why results of the two subsets, which were identified using similar inclusion and exclusion criteria, differed from each other (where one surpassed the natural history benchmark and the other did not). First, it could be that the SOC_{CBCL} subset had more obvious signs of anxiety than the SOC_{diag} subset and hence were more likely to receive appropriate treatment and this was why they improved more than the other subset. Youth presenting with severe internalising symptoms might be more likely to receive treatment specifically to address anxiety, rather than competing externalising diagnoses. Alternatively, differences in the observed effect sizes of the two subsets may relate to their pre-treatment CBCL-Int mean and *SD* scores. Examination of the mean and standard deviation of the SOC_{CBCL} subset (mean = 73.3, $SD = 6.6$) revealed that the mean of this group was higher than the pre-treatment mean of any of the clinical trials, and the standard deviation was lower (see Table 3). The pre-treatment mean and standard deviation of the SOC_{diag} subset (mean = 68.4, $SD = 9.0$), however, were more commensurate with those of clinical trials. Hence, while attempts were made to match these two subsets to clinical trials, the SOC_{CBCL} was not as comparable on pre-treatment

mean and standard deviation of the dependent variable from which effect size estimates were generated. As mentioned, more severe symptoms at pre-treatment have been associated with larger pre-post effect sizes (e.g. Kley et al., 2012; Oei & Boschen, 2009). Further, even small differences in standard deviations can result in large differences in pre-post effect size estimates (Lueger & Barkham, 2010). In combination, these factors might mean that the elevated effect size for this subset was an artefact of study design rather than truly reflecting the magnitude of the impact of treatment for youth with anxiety problems seen at the SOC CMHS agencies. That is, the effect size estimate of the SOC_{CBCL} may have been larger because selection criteria meant the pre-test mean was higher and standard deviation was lower than the SOC_{diag}, rather than because treatment was more effective for this group. Results pertaining to the second type of benchmark – 'clinically significant improvement' - can help to clarify interpretation of findings related to pre-post effect sizes, by allowing consideration of the consistency of outcomes across the two types of benchmarks. This in turns allows integration of findings and increases confidence in conclusions.

Hypothesis 2: Comparison of SOC CMHS subsets and 'Clinically Significant Improvement' Benchmarks.

Inconsistent with hypothesis 2, the proportion of youth in the SOC CMHS subsets who evidenced 'clinically significant improvement' on the CBCL-Int was significantly lower than the proportion of youth in treatment groups of clinical trials who evidenced this improvement. In fact, inconsistent with hypothesis 2, the proportion of youth evidencing 'clinically significant improvement' in the SOC subsets was not significantly greater than those of wait list control groups. Consistent with results of pre-post effect

size estimates, these rates suggest that the improvement of youth in the SOC_{diag} subset were not significantly different from wait list control groups and could be considered clinically equivalent to rates of natural remission in youth with anxiety disorders. While previous research suggests lower pre-treatment mean is associated with smaller effect sizes but better rates of recovery, this was not found in the SOC_{diag} subset, suggesting that results were not explained by pre-treatment CBCL profile. Results for the SOC_{CBCL} are consistent with previous research that has found that higher initial symptom severity is associated with greater improvement (as measured by effect sizes) but less recovery (from pathological functioning) (e.g. Liber et al., 2010).

There may be a number of reasons for the relatively weak results of both SOC CMHS subsets compared to treatment efficacy benchmarks for both effect sizes and rates of *CSI*. Firstly, it may be that treatments for anxiety being used within SOC CMHS agencies are not as efficacious as those ESTs tested within clinical trials. Although previous research found that clinicians at SOC CMHS agencies reported using evidence based practice (Sheehan et al., 2007), this finding was based on self-report which may over-estimate incorporation of evidence based components in actual practice (Hurlbert, Garland, Nguyen, & Brookman-Frazee, 2010). Further, the definition of 'evidence based practice' was broad (e.g. 'used cognitive behavioural therapy') and did not include a measure of adherence to particular protocols with empirical support for treatment of specific disorders (such as use of Coping Cat). There was no information available regarding the content or focus of therapy in the SOC CMHS subsets, which makes it more difficult to understand the sub-optimal results of the SOC CMHS with any confidence. There are numbers of studies showing that UC does not produce results

comparable with ESTs (e.g. Weisz et al., 2006), and research by Garland and colleagues has found that UC interventions tend to be spread over a relatively long time and are unlikely to include components of evidence-based practice (Hurlbert et al., 2010). Clinical trials evaluating ESTs for anxiety disorders were only beginning to emerge during the time of data collection in the SOC CMHS sample, which makes it less likely that the core components associated with better treatment outcomes were being used within SOC CMHS agencies. This hypothesis was consistent with the observation that, similar to other research on 'dose' of therapy in UC (Trask & Garland, 2012) a full 'dose' of therapy in the subsets did not impact outcome, suggesting that the therapy being delivered was not impactful.

Alternatively, the relatively weak results might not relate to the effectiveness of treatment received, but rather reflect the characteristics or context of youth seen in SOC agencies. Consistent with previous research in community settings (Southam-Gerow et al., 2003), the prevalence of indices of social deprivation and complex clinical presentation were high in both subsets in the present research. A large proportion of the SOC youth lived in poverty, in sole parent homes, and their externalising symptomology was high, even after youth with the most severe comorbid disorders had been excluded. It could be argued that high externalising comorbidity and social deprivation contributed to weaker outcomes (c.f. Southam-Gerow et al., 2001). However, as reviewed, there have been inconsistent findings regarding the impact of social deprivation and externalising symptoms on therapy outcome. Further, multivariate and univariate analyses (discussed later) found no significant associations between treatment response or reliable improvement on any of these potential moderators of outcome. This suggests these

characteristics do not explain the suboptimal outcomes of the SOC CMHS youth. Alternatively, although efforts were made to maximise match between SOC subsets and clinical trials, there may be other ways (not measured) in which populations served within SOC CMHS are fundamentally different from those in clinical trials and these differences mean that they are a more challenging population to treat. For instance, it is possible (and likely) that youth involved in numbers of services (including juvenile justice) have a history of treatment failure. They would be less likely to need to access further services if they experience good treatment response in one. Poor response to previous treatments is likely to predict poor response to future treatments and may account for weaker outcomes of the SOC CMHS samples.

Another plausible explanation for the relatively poor results of SOC CMHS subsets is that although the youth in the subsets were experiencing substantially elevated symptoms of anxiety, this may not have been the focus of the treatment they received. There was high comorbidity in the sample, and the number of children diagnosed with ADHD in the SOC_{CBCL} subset was actually higher than those diagnosed with anxiety. The high prevalence is not surprising, given previous research findings of high comorbidity between ADHD and anxiety, that having comorbid ADHD increases the likelihood that a child will access treatment (Hammerness et al., 2010; Kendall et al., 2010), and that externalising disorders may be more likely to be diagnosed than internalising. It is possible that treatment was focused on addressing other difficulties, and for this reason, symptoms of anxiety did not reduce to the same degree as those of youth receiving ESTs for anxiety. It is difficult to ascertain whether this was the case, since the SOC national evaluation data set did not contain information regarding the

focus of treatment. Future studies might include this information in order to aid interpretation of results.

Lastly, it is possible that SOC subsets failed to achieve results comparable to the treatment benchmarks because they were generated using results of both 'completer' and 'intent to treat' (ITT) samples from clinical trials. Moderator analysis revealed that clinical trials using ITT samples had marginally lower effect sizes than those using 'completer' samples. Clinical trials reporting ITT data generated an aggregate effect size of $Y_{ITT} = 0.85$ which is fairly close to the observed effect size of the SOC CBCL ($Y_{SOC\ CBCL} = .79$), though not the SOC_{diag} subset ($Y_{diag} = .52$). However, this is unlikely to account for results. First, as mentioned, the observed effect sizes for the 'completer' subsamples of both subsets were almost identical to the one generated for the full ITT sample for both subsets. Thus, using the 'completer' subsample of the SOC_{CBCL} would not exceed the treatment efficacy benchmark, nor would the effect size estimate of full subsets exceed the critical values of a treatment efficacy benchmark based exclusively on ITT studies. Either way, the conclusions of the present study would remain unchanged.

Hypothesis 3, 4 and 5: Factors Associated with Treatment Response

A logistic regression found three variables reflecting case complexity (poverty status; externalising comorbid diagnosis; affective comorbid diagnosis) did not predict reliable improvement in the SOC CMHS youth. Further, comparison of 'good' and 'poor' treatment responders failed to find significant differences on 11 variables at pre-treatment. Previous research examining prediction of treatment response in anxious youth has also generated relatively few significant findings (Southam-Gerow et al., 2001; Liber et al., 2010). The absence of significant differences in pre-treatment variables in the

present research might be due to several factors. First, it may be that the variables available for consideration, while intuitively appealing, are not associated with treatment responsivity. Second, it could be that any one variable alone is not sufficient to impact outcome, but that their impact is cumulative (c.f. Lincoln & Rief, 2004). For instance, a child with ADHD and anxiety, with few emotional resiliencies, being raised by a poor, single parent who suffers from depressive and anxiety disorders might be more likely to show a 'poor' treatment response than a youth not challenged by any of these factors. Third, it could be an artefact of the operationalization of the dependent variables (i.e. 'reliable improvement' or 'treatment response'). While dividing groups on either side of a cut off means that there are more participants in the analysis, which potentially improves power to detect differences, making the division in this manner reduces differences between the groups since participants close to the border whose scores may not be substantially different are included in the analysis. For instance, examination of differences between youth at the more extreme ends of the 'treatment response' continuum, while excluding numbers of participants, might identify factors critical in predicting either extreme success or non-response. This could not be examined in the present research due to small sample size but could be in future research with larger samples. Lastly, the lack of significant findings may simply reflect an under-powered analysis. Using a conservative definition of 'treatment response' contributed to quite an uneven split in the number of 'good' and 'poor' responders, with 'good' treatment responders having had a relatively small sample size (n = 21) compared to 'poor' treatment responders (n = 79). A larger sample size would have increased power to detect factors associated with treatment response.

Implications of Results

The initial focus of discussion will be on the clinical implications of findings of the present study. Results suggest that an understanding of UC at SOC CMHS agencies is needed to provide context to the sub-optimal outcomes of anxious youth receiving services in these organisations. First, the degree to which difficulties with anxiety are under-diagnosed should be considered, and efforts to promote recognition of anxiety as a comorbid or underlying issue in the presence of other disorders should be undertaken. Improved recognition of anxiety disorders might be particularly important, since they can act as a 'gateway' to a host of other problems in adolescence and adulthood (Kendall et al., 2012). The extent to which ESTs for the treatment of anxiety disorders are being utilised in SOC CMHS agencies should also be considered, particularly since results of the moderator analysis of published trials found there were no significant differences in outcomes of ESTs delivered in the community compared to research settings. If it emerges that ESTs are not routinely used (as previous research by Garland and colleagues regarding usual practice in child and youth services would predict), strategies to optimise dissemination and implementation of EBP could be employed. To assist in this process, implementation of ESTs in the community should consider both flexible adaptations to local conditions as well as fidelity to core components of the interventions (c.f. Beidas, Benjamin, Puleo, Edmunds, & Kendall, 2010; Kendall, Gorsch, Furr, & Sood, 2008; Mazzuchelli & Sanders, 2010). Emerging treatment and training models, designed to assist with incorporation of EBP in the context of the realities of community mental health work (including comorbidity), might also be helpful in this context

(Chorpita, Bernstein, & Daleiden, 2011; Chorpita & Daleiden, 2009; Chorpita, Daleiden, & Weisz, 2005; Kendall et al., 2012; Southam-Gerow et al., 2012; Weisz et al., 2012).

The moderator analysis found no significant differences between the outcomes of treatments that make use of alternative technology (i.e. bibliotherapy, telephone counselling, or the internet) and those that used more traditional therapy models. This finding should be interpreted with caution, since it could reflect an under-powered analysis and includes treatments with a broad range of intensity. Nonetheless, in light of recent work emphasising the importance of parsimony in delivery of treatment, the absence of clear advantage of interventions requiring more intense therapist involvement suggests that future research should be invested in examining whether treatment protocols delivered using these alternative approaches could be considered ahead of more costly ones (Cougle, 2012; Kendall et al., 2012).

Results also have methodological implications. The magnitude of data reduction illustrates how challenging it was to retrospectively identify youth in SOC agencies with anxiety problems that were commensurate with the samples of clinical trials. The correspondence between DSM diagnosis and profile on CBCL-DSM Anx was not perfectly consistent (see Table 1). That is, not all youth with an elevated CBCL-DSM Anx profile had a recorded DSM diagnosis, and not all youth with DSM diagnoses had elevated CBCL-DSM Anxiety profiles. Further, almost no youth presented with a diagnosis of an anxiety disorder, a presenting problem of anxiety *and* a CBCL-DSM Anx score in the 'borderline' or 'clinical' range. It is likely these differences primarily relate to the issue of measurement in community settings. Specifically, the process of establishing a formal DSM diagnosis in the SOC CMHS communities was likely much

less systematised than that used in clinical trials. This reality necessitated the use of two strategies to identify youth with anxiety disorders within the broader SOC CMHS data set. These two strategies each have both advantages and pitfalls. The process illustrates that there is no single algorithm available to perfectly match youth seen at SOC agencies with samples from published research particularly with respect to clinical profile. The implications of this reality are that results of different data reduction strategies can be used to complement each other, with the advantages and disadvantages of each borne in mind when interpreting findings.

There was substantial heterogeneity in effect sizes across treatment trials, even though these were established using the same measure to evaluate broadly similar treatment approaches (i.e. CBT). This suggests that combining results of effect sizes based on measures with different constructs, levels of specificity and metric are likely to yield findings that are increasingly difficult to interpret. Consistent with Minami et al.'s recommendations, results confirm the importance of generating benchmarks based on either identical measures, or those with similar reactivity and specificity. Further, consistent with Minami et al. (2009), it is also worth considering whether specific benchmarks can be generated to match particular subgroups within community datasets (e.g. clients with comorbid conditions; 'intent to treat' samples).

Results also illustrate the importance of considering publication bias when generating benchmarks. The effect size estimate adjusted for possible publication bias generated a smaller treatment effect size benchmark than that from the original calculation. While using this adjusted effect size as the treatment benchmark would not have altered conclusions of the present research, it illustrates that it is possible and likely

that there is a bias for small studies to be published only if they have reasonably large positive findings. The possibility of bias toward publishing small studies with large effect sizes (but not small studies with null findings) should be considered in any efforts to generate future benchmarks.

Lastly, there were marginal (though not significant) differences in outcomes of research reporting 'completer' and 'ITT' data, suggesting that future benchmarking studies should take ITT status into consideration when evaluating outcomes of community groups against those of published trials.

Limitations of Study

There was some ambiguity regarding the nature, content and focus of treatment received through SOC CMHS agencies which made it more challenging to interpret results. The length of treatment was potentially greater than was typical in clinical trials (up to 26 weeks), and for some youth involved far more sessions. Further, the content might not have reflected components of ESTs for anxiety, or the focus of treatment may have been on comorbid conditions. It should be noted, however, that the point of benchmarking comparisons is to understand whether youth seen at SOC CMHS agencies who presented with similar difficulties and received treatment in a similar format (i.e. therapy) achieved treatment gains as optimal as those seen in youth receiving empirically supported treatments. Further, even in the presence of comorbid conditions, the potentially powerful outcomes associated with treatment of anxiety disorders might make them the more logical choice for initial treatment focus, and the length of the treatment window (i.e. 26 weeks) could potentially accommodate sequential interventions for more than one disorder.

The possible confound of differences in pre-treatment mean and standard deviation of CBCL-Int scores also made interpretation of results more challenging. Anecdotally, analyses were run numbers of times with different iterations of inclusion and exclusion criteria. Findings were fairly consistent; changes in criteria for subsets based on contextual or diagnostic variables (e.g. treatment sector or DSM diagnoses) tended to have very modest impact on results. Any changes in the inclusion criteria that systematically reduced the range and increased the mean of pre-treatment CBCL-Int scores were associated with increases in pre-post effect size estimates. This highlights the value of using more than one subset and more than one way of operationalizing outcome, and represents an interesting avenue for future enquiry.

Another important difference between the SOC CMHS subsets and clinical trials was the stability of medication use by youth. Unstable medication use may have negatively impacted outcomes of the SOC CMHS subsets. Reducing the sample to include only those with stable medication regimes would have excluded the great majority of an already substantially reduced sample. For this reason, as in other community-based research (e.g. Oei & Boschen, 2009), unstable medication use was an acknowledged confound in the present study's research design.

The final analysis for examining factors associated with treatment response included only a small number of 'good' treatment responders (n = 21, 20.8%). This meant 'reliable improvement' was used as the dependent variable in the logistic regression. 'Reliable improvement' is arguably a less clinically relevant operationalization of response to treatment than 'treatment response', and may have a different relationship with predictor variables. Further, the relatively small number of

'good' treatment responders limited the power of the univariate analyses and may have compromised the detection of important but subtle differences between 'good' and 'poor' treatment responders. Further, the small sample may reduce generalisability of results. This issue can be addressed in future research using a larger number of participants.

The psychometric properties of the CBCL and most other measures have not been specifically evaluated in the context of the SOC CMHS national evaluation study. The CBCL (and most other instruments) were self-report questionnaires, which might lessen the influence of a specific context on their reliability and validity. Nonetheless, ideally, the psychometric properties of measures should be established for the population and context in which they will be used. Future research should investigate the psychometric properties of measures in the context of use in routine clinical work in SOC CMHS agencies. This might be particularly important for instruments such as the CAFAS, which incorporate clinician judgment in scoring.

Lastly, the present benchmarks did not incorporate results of studies treating either PTSD or OCD, which are both classified as DSM IV TR anxiety disorders. This is because there were no RCTs evaluating treatments of these disorders that met criteria for inclusion in the present research. Thus, the present benchmark standards may be limited to the anxiety disorders treated within the clinical trials – mostly Generalised Anxiety Disorder, Separation Anxiety Disorder and phobias. It is possible that broadening inclusion criteria for identifying studies (e.g. by extending age range; considering studies with quasi experimental designs) might identify appropriate treatment studies for these disorders. Alternatively, separate benchmarks (e.g. for PTSD) using raw CBCL scores could be generated.

Implications for Future Research

Future research could use common broadband outcome measures to benchmark a range of presenting problems and diagnoses. These might include externalising difficulties (such as ADHD or Oppositional Defiant Disorder), mood disorders or other anxiety disorders (including PTSD). Further, given many communities organise their outcome data in broad symptom-based categories (e.g. 'internalising'; 'externalising'; 'comorbid'), rather than DSM diagnoses (e.g. CAFAS Ontario, 2010), benchmarks could be generated to mirror these groupings. In addition to broadening the scope of presenting problems or diagnoses, future research could extend population and time frame parameters – for instance, including treatment trials for adolescents or very young children and examining long-term outcomes of treatment. One potential advantage of the present study is that it generated benchmarks based on a commonly used broadband scale of childhood psychopathology (the CBCL/6-18). This is useful for ongoing large-scale community evaluations where broadband measures are more likely to be utilised and future investigators should be encouraged to incorporate this common measure in their research designs. There are contexts, however, in which community agencies might want benchmarks based on symptom-specific measures of outcome (for example, when piloting implementation of an empirically supported treatment), and these could also be generated. Further, these symptom-based benchmarks could be compared to those based on broadband measures, to establish the relative utility of each (for example, consideration of the sensitivity of broad band measures to treatment effects for any given disorder).

The present study focused entirely on symptom reduction as an index of outcome. Other indices of outcome are important, particularly to community agencies, and can be benchmarked. For instance, benchmark standards can be established for child psycho - social functioning, treatment duration and attrition (Hunsley & Lee, 2007). Lastly, future research might seek to establish benchmarks for treatment within specific treatment contexts – for instance welfare, child protection, and juvenile justice settings - to establish effectiveness of empirically supported treatments in those sectors. The relatively small number of studies within each of these contexts might necessitate broadening inclusion criteria for treatment studies to contribute to the benchmark (for example, including non-randomised trials). Benchmarks based on ESTs in these contexts would lead to greater confidence in the comparability of outcomes with clinical populations such as those included within the SOC CMHS subsets.

Treatment response and reliable improvement were not significantly associated with any of the factors examined. As mentioned, this may be because the effects of the variables tested are cumulative, because important moderators were not included in the analyses or because the analyses were underpowered. Future research could test clusters of possible predictors, examine as yet untested factors (e.g. history of service use), and/or include a greater number of participants. This would allow for multivariate analysis examining prediction of treatment response and maximise power to detect differences between groups. Process variables such as therapeutic alliance, engagement and homework completion could also be considered in this examination and findings could be used to extend to understanding of the mechanisms by which variables impact outcomes (Kendall et al., 2012).

Summary and Conclusions

The present study applied a benchmarking strategy to evaluate the outcomes of youth with anxiety disorders treated at SOC CMHS agencies. Results of selected treatment trials were aggregated to generate two different kinds of benchmarks. First, pre-post effect sizes from treatment trials were combined to generate pre-post effect size treatment efficacy (Y_{TE}) and natural history (Y_{NH}) benchmarks. These pre-post effect size estimates were based on a broadband measure of internalising psychopathology in youth (CBCL-Int; Achenbach, 1991; Achenbach & Rescorla, 2001). Second, the proportion of youth demonstrating 'clinically significant improvement' was aggregated across trials to generate treatment (CSI_{TE}) and natural history (CSI_{NH}) benchmarks for 'clinically significant improvement'. 'Clinically significant improvement' was operationalised as falling from 'clinical' (CBCL-Int T score >= 65) to 'subclinical' range of functioning (CBCL-Int T score < 65) on the CBCL-Int. Lastly, factors associated with reliable improvement and treatment response were examined. 'Good' treatment responders were defined as those who (1) moved from 'deviant' to 'normal' range of functioning on the CBCL-DSM Anx scale (pre-treatment CBCL-DSM Anx T >= 65 to CBCL- DSM Anx T < 65) and (2) who demonstrated reliable improvement on the CBCL-DSM Anx scale, as indicated by the Jacobson and Truax (1991), Reliable Change Index (RCI). 'Poor' treatment responders were those youth who were in the 'deviant' range of functioning on the CBCL-DSM Anx at baseline, but who failed to meet both these improvement criteria during the six months of involvement with SOC CMHS services. Because there were too few 'good' treatment responders to allow for multivariate analysis, a logistic regression examined prediction of 'reliable improvement' status by variables related to case

complexity (poverty, comorbid externalising diagnosis and comorbid affective diagnosis). Exploratory univariate analyses were conducted where 'good' and 'poor' treatment responders were compared on 11 variables related to demographics (gender, age, ethnicity), family context (family functioning, caregiver stress, poverty status, number of risk factors), child strengths and resiliencies, child functional impairment and child psychopathology (externalising comorbidity, affective comorbidity).

Results revealed that outcomes of youth from SOC CMHS subsets were significantly worse than those of youth who received empirically supported treatments for both types of benchmarks (effect sizes and rates of 'clinically significant improvement'). The pre-post effect sizes of youth selected on the basis of elevated CBCL-DSM Anx scores did improve more than what might have been expected with the passage of time alone, whereas the pre-post effect sizes of youth selected primarily on the basis of DSM diagnosis did not. Neither subset achieved rates of 'clinically significant improvement' that were significantly different than natural remission. Three indices of case complexity (poverty, externalising and affective comorbidity) failed to predict 'reliable improvement' in the logistic regression. Differences between 'good' and 'poor' treatment on the 11 demographic, family context, child strength and child psychopathology variables that were tested were not significant. Thus, none of the variables examined offered an explanation for the relatively poor outcomes of youth in the SOC CMHS agencies.

In conclusion, establishing benchmark standards for outcomes of evidence based treatment of anxiety disorders in children is potentially extremely helpful to community agencies to contextualise the impact of their services, including evaluating whether they

are commensurate with the optimal but obtainable outcomes of studies evaluating ESTs or with the wait list control groups of those studies. While a 'perfect' match between clinical profile of youth in the community sample and those from research trials was not achieved, results were still informative. The present study showed even moderate improvement in symptoms may not be better than natural remission, and also that methodological nuance can significantly impact magnitude of effect sizes. Outcomes of youth in the SOC CMHS subsets were consistently worse than those of youth receiving ESTs in clinical trials. Outcomes for one subset were generally more comparable with the outcomes of wait list control groups in those trials. The reasons for suboptimal improvement in SOC CMHS agencies are not clear. It may be that the content of therapy delivered in UC in SOC CMHS agencies was not consistent with research-supported interventions for anxiety. Alternatively it might be that the focus of intervention was not on treatment of anxiety and therefore anxiety-related symptoms were not substantially reduced during the course of contact with services. Results show that some differences between community samples and clinical trials (including baseline symptoms) can confound interpretation of findings. This suggests that when there is no ideal way to identify the target population within a broader data set, it may be pragmatic to identify more than one comparator group (e.g. as in the present study, using subsets identified by diagnosis and CBCL profile) and more than one metric of outcome (e.g. effect sizes as well as rates of 'clinically significant improvement') in order to strengthen confidence in findings and allow cross validation of results. The present research did not identify any factors associated with reliable improvement or treatment response and future research might examine clusters of variables or alternative factors that might be associated with

outcome. Results of the present study emphasise the importance of understanding UC in SOC agencies, including whether efforts would best be directed at identifying youth with anxiety, disseminating, implementing, adapting and/or maintaining use of ESTs for treatment of anxiety within the clinical contexts of these agencies. With appropriate attention to methodological and clinical issues, the process of benchmarking can be used as an ongoing strategy to help support these endeavours.

References

References marked with an asterix indicate studies included in the benchmarking aggregation. The in-text citations to studies selected for benchmarking aggregation are not preceded by asterisks.

Accurso, E. C., Taylor, R. M., & Garland, A. F. (2011). Evidence-based practices addressed in community-based children's mental health clinical supervision. *Training and Education in Professional Psychology, 5*(2), 88-96. doi: http://dx.doi.org/10.1037/a0023537

Achenbach, T. M. (1991). *Manual for Child Behavior Checklist/ 4-18 and 1991 Profile*. Burlington, VT: University of Vermont, Dept. of Psychiatry.

Achenbach, T. M., & Rescorla, L. A. (2001). *Manual for the ASEBA school-age forms & profiles: an integrated system of multi-informant assessment*. Burlington, VT: University of Vermont. Research Center for Children, Youth, & Families.

Achenbach, T. M., Dumenci, L., & Rescorla, L. A. (2003). DSM-oriented and empirically based approaches to constructing scales from the same item pools. *Journal of Clinical Child and Adolescent Psychology, 32*(3), 328-340. doi: http://dx.doi.org/10.1207/S15374424JCCP3203_02

Achenbach, T. M., Howell, C. T., McConaughy, S. H., & Stanger, C. (1995). Six-year predictors of problems in a national sample: III. Transitions to young adult syndromes. *Journal of the American Academy of Child & Adolescent Psychiatry, 34*(5), 658-669. doi: http://dx.doi.org/10.1097/00004583-199505000-00018

Addis, M. E., Hatgis, C., Krasnow, A. D., Jacob, K., Bourne, L., & Mansfield, A. (2004). Effectiveness of cognitive-behavioral treatment for panic disorder versus treatment

as usual in a managed care setting. *Journal of Consulting and Clinical Psychology, 72*(4), 625-635. doi: http://dx.doi.org/10.1037/0022-006X.72.4.625

Addis, M. E., & Waltz, J. (2002). Implicit and untested assumptions about the role of psychotherapy treatment manuals in evidence-based mental health practice: Commentary. *Clinical Psychology: Science and Practice, 9*(4), 421-424. doi: http://dx.doi.org/10.1093/clipsy/9.4.421

Albano, A. M., Chorpita, B. F., & Barlow D. H. (1996). Anxiety disorders. In E. J. Mash & R. A. Barkley (Eds.), *Child Psychopathology* (pp. 196-241). New York: Guilford Press.

Alfano, C. (2012). Are children with "pure" generalized anxiety disorder impaired? A comparison with comorbid and healthy children. *Journal of Clinical Child and Adolescent Psychology, 41*(6), 739-745.

American Psychiatric Association (1994). *Diagnostic and statistical manual of mental disorders (4th edition).* Washington, D.C: Author.

American Psychiatric Association (2000). *Diagnostic and statistical manual of mental disorders (4th edition, text revision).* Washington, D.C: Author.

American Psychological Association. Presidential Task Force on Evidence-Based Practice (2006). Evidence-based practice in psychology. *American Psychologist*, *61*(4), 271-285. doi: http://dx.doi.org/10.1037/0003-066X.61.4.271

Andrade, A. R., Lambert, E. W., & Bickman, L. (2000). Dose effect in psychotherapy: Outcomes associated with negligible treatment. *Journal of the American Academy of Child & Adolescent Psychiatry, 39*(2), 161-168. doi: 10.1097/00004583-200002000-00014

Angold, A., Costello, E. J., & Erkanli, A. (1999). Comorbidity. *Journal of Child Psychology and Psychiatry, 40*(1), 57-87. doi: http://dx.doi.org/10.1111/1469-7610.00424

Baker-Ericzén, M. J., Hurlburt, M. S., Brookman-Frazee, L., Jenkins, M. M., & Hough, R. L. (2010). Comparing child, parent, and family characteristics in usual care and empirically supported treatment research samples for children with disruptive behavior disorders. *Journal of Emotional and Behavioral Disorders, 18*(2), 82-99. doi: http://dx.doi.org/10.1177/1063426609336956

Barrett, P. M. (1995). Group coping koala workbook. *Unpublished manuscript, School of Applied Psychology, Griffith University, Australia*.

*Barrett, P. M. (1998). Evaluation of cognitive-behavioral group treatments for childhood anxiety disorders. *Journal of Clinical Child Psychology, 27*(4), 459-468.

*Barrett, P. M., Dadds, M. R., & Rapee, R. M. (1996). Family treatment of childhood anxiety: A controlled trial. *Journal of Consulting and Clinical Psychology, 64*(2), 333-342. doi: http://dx.doi.org/10.1037/0022-006X.64.2.333

Barrett, P., Farrell, L., Dadds, M., & Boulter, N. (2005). Cognitive-behavioral family treatment of childhood obsessive-compulsive disorder: Long-term follow-up and predictors of outcome. *Journal of the American Academy of Child & Adolescent Psychiatry, 44*(10), 1005-1014. doi: http://dx.doi.org/10.1097/01.chi.0000172555.26349.94

Barrett, P., Healy-Farrell, L., & March, J. S. (2004). Cognitive-behavioral family treatment of childhood obsessive-compulsive disorder: A controlled trial. *Journal of*

the American Academy of Child & Adolescent Psychiatry, 43(1), 46-62. doi: http://dx.doi.org/10.1097/00004583-200401000-00014

Barrington, J., Prior, M., Richardson, M., & Allen, K. (2005). Effectiveness of CBT versus standard treatment for childhood anxiety disorders in a community clinic setting. *Behaviour Change, 22*(1), 29-43. doi: http://dx.doi.org/10.1375/bech.22.1.29.66786

Beck, A. T., & Steer, R. A. (1990). *BAI, Beck anxiety inventory: manual*. San Antonio, TX: Psychological Corporation.

Begg, C. B., & Mazumdar, M. (1994). Operating characteristics of a rank correlation test for publication bias. *Biometrics*, 1088-1101.

Beidas, R. S., Barmish, A. J., & Kendall, P. C. (2009). Training as usual: Can therapist behavior change after reading a manual and attending a brief workshop on cognitive behavioral therapy for youth anxiety? *The Behavior Therapist, 32*(5), 97-101.

Beidas, R. S., Benjamin, C. L., Puleo, C. M., Edmunds, J. M., & Kendall, P. C. (2010). Flexible applications of the coping cat program for anxious youth. *Cognitive and Behavioral Practice, 17*(2), 142-153. doi: http://dx.doi.org/10.1016/j.cbpra.2009.11.002

Beidas, R. S., & Kendall, P. C. (2010). Training therapists in evidence-based practice: A critical review of studies from a systems-contextual perspective. *Clinical Psychology: Science and Practice, 17*(1), 1-30. doi: http://dx.doi.org/10.1111/j.1468-2850.2009.01187.x

*Beidel, D. C., Turner, S. M., & Morris, T. L. (2000). Behavioral treatment of childhood social phobia. *Journal of Consulting and Clinical Psychology, 68*(6), 1072-1080. doi: http://dx.doi.org/10.1037/0022-006X.68.6.1072

Berman, S. L., Weems, C. F., Silverman, W. K., & Kurtines, W. M. (2000). Predictors of outcome in exposure-based cognitive and behavioral treatments for phobic and anxiety disorders in children. *Behavior Therapy, 31*(4), 713-731. doi: http://dx.doi.org/10.1016/S0005-7894(00)80040-4

Bickman, L. (1999). Practice makes perfect and other myths about mental health services. *American Psychologist, 54*(11), 965-978. doi: 10.1037/h0088206

Bickman, L. (2008). A measurement feedback system (MFS) is necessary to improve mental health outcomes. *Journal of the American Academy of Child & Adolescent Psychiatry, 47*(10), 1114-1119. doi: 10.1097/CHI.0b013e3181825af8

Bickman, L., Andrade, A. R., & Lambert, E. W. (2002). Dose response in child and adolescent mental health services. *Mental Health Services Research, 4*(2), 57-70. doi: 10.1023/A:1015210332175

Biostat (2006). Comprehensive Meta Analysis (2.0)[program]. Englewood, NJ: Biostat, Inc.

Bittner, A., Egger, H. L., Erkanli, A., Costello, E. J., Foley, D. L., & Angold, A. (2007). What do childhood anxiety disorders predict? *Journal of Child Psychology and Psychiatry, 48*(12), 1174-1183. doi: http://dx.doi.org/10.1111/j.1469-7610.2007.01812.x

Blais, M. A., Sinclair, S. J., Baity, M. R., Worth, J., Weiss, A. P., Ball, L. A., & Herman, J. (2012). Measuring outcomes in adult outpatient psychiatry. *Clinical Psychology & Psychotherapy, 19*(3), 203-213. doi: http://dx.doi.org/10.1002/cpp.749

Bodden, D. H. M., Bögels, S. M., Nauta, M. H., De Haan, E., RIingrose, J., Appelboom, C., . . . Appelboom-Geerts, Karen C. M. M. J. (2008). Child versus family cognitive-behavioral therapy in clinically anxious youth: An efficacy and partial effectiveness study. *Journal of the American Academy of Child & Adolescent Psychiatry, 47*(12), 1384-1394. doi: http://dx.doi.org/10.1097/CHI.0b013e318189148e

Bögels, S. M., & Siqueland, L. (2006). Family cognitive behavioral therapy for children and adolescents with clinical anxiety disorders. *Journal of the American Academy of Child & Adolescent Psychiatry, 45*(2), 134-141. doi: http://dx.doi.org/10.1097/01.chi.0000190467.01072.ee

Borkovec, T. D., & Costonguay, L. G. (1998). What is the scientific meaning of empirically supported therapy? *Journal of Consulting and Clinical Psychology, 66*(1), 136-142. doi: http://dx.doi.org/10.1037/0022-006X.66.1.136

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2011). *Introduction to meta-analysis*. Wiley.

Brady, E. U., & Kendall, P. C. (1992). Comorbidity of anxiety and depression in children and adolescents. *Psychological Bulletin, 111*(2), 244-255. doi: http://dx.doi.org/10.1037/0033-2909.111.2.244

Brannan, A. M., Heflinger, C. A., & Bickman, L. (1997). The caregiver strain questionnaire: Measuring the impact on the family of living with a child with serious

emotional disturbance. *Journal of Emotional and Behavioral Disorders, 5*(4), 212-222. doi: http://dx.doi.org/10.1177/106342669700500404

Brent, D. A., Kolko, D. J., Birmaher, B., Baugher, M., Bridge, J., Roth, C., & Holder, D. (1998). Predictors of treatment efficacy in a clinical trial of three psychosocial treatments for adolescent depression. *Journal of the American Academy of Child & Adolescent Psychiatry, 37*(9), 906-914. doi: http://dx.doi.org/10.1097/00004583-199809000-00010

Breton, J., Bergeron, L., Valla, J., Berthiaume, C., Gaudet, N., Lambert, J., . . . Lépine, S. (1999). Quebec child mental health survey: Prevalence of DSM-III—R mental health disorders. *Journal of Child Psychology and Psychiatry, 40*(3), 375-384. doi: http://dx.doi.org/10.1111/1469-7610.00455

Bullivant, J.R. (1996). Benchmarking in the UK National Health Service, *International Journal of Health Care Quality Assurance, 9(2*), 9 – 14. doi: http://dx.doi.org/10.1108/09526869610112707

Byles, J., Byrne, C., Boyle, M. H., & Offord, D. R. (1988). Ontario child health study: Reliability and validity of the general functioning subscale of the McMaster family assessment device. *Family Process, 27*(1), 97-104. doi: http://dx.doi.org/10.1111/j.1545-5300.1988.00097.x

CAFAS in Ontario (2010). *2010 Report: Level of Functioning Outcomes for Children and Youth Receiving Mental Health Treatment*. Toronto, ON: The Hospital for Sick Children.

Callahan, J. L., Almstrom, C. M., Swift, J. K., Borja, S. E., & Heath, C. J. (2009). Exploring the contribution of supervisors to intervention outcomes. *Training and*

Education in Professional Psychology, 3(2), 72-77. doi: http://dx.doi.org/10.1037/a0014294

Carroll, K. M., Martino, S., & Rounsaville, B. J. (2010). No train, no gain? *Clinical Psychology: Science and Practice, 17*(1), 36-40. doi: http://dx.doi.org/10.1111/j.1468-2850.2009.01190.x

Carroll, K. M., Nich, C., McLellan, A. T., McKay, J. R., & Rounsaville, B. J. (1999). 'Research'versus 'real-world'patients: representativeness of participants in clinical trials of treatments for cocaine dependence. *Drug and alcohol dependence*, *54*(2), 171-177.

*Cartwright-Hatton, S., McNally, D., Field, A. P., Rust, S., Laskey, B., Dixon, C., . . . Woodham, A. (2011). A new parenting-based group intervention for young anxious children: Results of a randomized controlled trial. *Journal of the American Academy of Child & Adolescent Psychiatry, 50*(3), 242-251. doi: http://dx.doi.org/10.1016/j.jaac.2010.12.015

Cartwright-Hatton, S., Roberts, C., Chitsabesan, P., Fothergill, C., & Harrington, R. (2004). Systematic review of the efficacy of cognitive behaviour therapies for childhood and adolescent anxiety disorders. *British Journal of Clinical Psychology, 43*(4), 421-436. doi: http://dx.doi.org/10.1348/0144665042388928

Caspi, A., Elder, G. H., & Bem, D. J. (1988). Moving away from the world: Life-course patterns of shy children. *Developmental Psychology, 24*(6), 824-831. doi: http://dx.doi.org/10.1037/0012-1649.24.6.824

Center for Mental Health Services. (2004). *The Comprehensive Community Mental Health Services for Children and Their Families Program: Evaluation findings—Annual report to Congress, 2004*. Atlanta, GA: Macro International Inc.

Chambless, D. L., & Hollon, S. D. (1998). Defining empirically supported therapies. *Journal of Consulting and Clinical Psychology, 66*(1), 7-18. doi: http://dx.doi.org/10.1037/0022-006X.66.1.7

Cho, M. K., & Bero, L. A. (1994). Instruments for assessing the quality of drug studies published in the medical literature. *JAMA: Journal of the American Medical Association, 272*(2), 101-104. doi:10.1001/jama.1994.03520020027007.

Chorpita, B. F., Bernstein, A., & Daleiden, E. L. (2011). Empirically guided coordination of multiple evidence-based treatments: An illustration of relevance mapping in children's mental health services. *Journal of Consulting and Clinical Psychology, 79*(4), 470-480. doi: http://dx.doi.org/10.1037/a0023982

Chorpita, B. F., & Daleiden, E. L. (2009). Mapping evidence-based treatments for children and adolescents: Application of the distillation and matching model to 615 treatments from 322 randomized trials. *Journal of Consulting and Clinical Psychology, 77*(3), 566-579. doi: 10.1037/a0014565

Chorpita, B. F., Daleiden, E. L., Ebesutani, C., Young, J., Becker, K. D., Nakamura, B. J., . . . Starace, N. (2011). Evidence-based treatments for children and adolescents: An updated review of indicators of efficacy and effectiveness. *Clinical Psychology: Science and Practice, 18*(2), 154-172. doi: http://dx.doi.org/10.1111/j.1468-2850.2011.01247.x

Chorpita, B. F., Daleiden, E. L., & Weisz, J. R. (2005). Identifying and selecting the common elements of evidence based interventions: A distillation and matching model. *Mental Health Services Research, 7*(1), 5-20. doi: 10.1007/s11020-005-1962-6

Clarke, G. N., Hawkins, W., Murphy, M., & Sheeber, L. B. (1995). Targeted prevention of unipolar depressive disorder in an at-risk sample of high school adolescents: A randomized trial of group cognitive intervention. *Journal of the American Academy of Child & Adolescent Psychiatry, 34*(3), 312-321. doi: http://dx.doi.org/10.1097/00004583-199503000-00016

ClinicalTrials.gov. Retrieved August 1 2012, from http://www.clinicaltrials.gov.

Cobham, V. E., Dadds, M. R., & Spence, S. H. (1998). The role of parental anxiety in the treatment of childhood anxiety. *Journal of Consulting and Clinical Psychology, 66*(6), 893-905. doi: 10.1037/0022-006X.66.6.893

Cohen J. (1988). *Statistical Power Analysis for the Behavioral Sciences.* 2nd ed. Hillsdale, NJ: Lawrence Erlbaum.

Cohen, J. A., Deblinger, E., Mannarino, A. P., & Steer, R. A. (2004). A multisite, randomized controlled trial for children with sexual abuse-related PTSD symptoms. *Journal of the American Academy of Child & Adolescent Psychiatry, 43*(4), 393-402. doi: http://dx.doi.org/10.1097/00004583-200404000-00005

Compton, S. N., Burns, B. J., Egger, H. L., & Robertson, E. (2002). Review of the evidence base for treatment of childhood psychopathology: Internalizing disorders. *Journal of Consulting and Clinical Psychology. Special Issue: Impact of Childhood*

Psychopathology Interventions on Subsequent Substance Abuse, 70(6), 1240-1266. doi: 10.1037/0022-006X.70.6.1240

Compton, S. N., March, J. S., Brent, D., Albano, A. M., Weersing, V. R., & Curry, J. (2004). Cognitive-behavioral psychotherapy for anxiety and depressive disorders in children and adolescents: An evidence-based medicine review. *Journal of the American Academy of Child & Adolescent Psychiatry, 43*(8), 930-959. doi: http://dx.doi.org/10.1097/01.chi.0000127589.57468.bf

Conners, C.K. (1973). Rating scales for use in drug studies with children. *Psychopharmacology Bulletin* [Special issue on children], 24-42.

Cook, R. D., & Weisberg, S. (1982). *Residuals and influence in regression* (Vol. 5). New York: Chapman and Hall.

Cook, C. R., Williams, K. R., Guerra, N. G., Kim, T. E., & Sadek, S. (2010). Predictors of bullying and victimization in childhood and adolescence: A meta-analytic investigation. *School Psychology Quarterly, 25*(2), 65-83. doi: http://dx.doi.org/10.1037/a0020149

Costello, E. J., Mustillo, S., Erkanli, A., Keeler, G., & Angold, A. (2003). Prevalence and development of psychiatric disorders in childhood and adolescence. *Archives of General Psychiatry, 60*(8), 837-844. doi: http://dx.doi.org/10.1001/archpsyc.60.8.837

Cougle, J. R. (2012). What makes a quality therapy? A consideration of parsimony, ease, and efficiency. *Behavior Therapy, 43*(3), 468-481. doi: http://dx.doi.org/10.1016/j.beth.2010.12.007

Craske, M. G. (1997). Fear and anxiety in children and adolescents. *Bulletin of the Menninger Clinic, 61*(2), A4-A36.

Crawford, A. M., & Manassis, K. (2001). Familial predictors of treatment outcome in childhood anxiety disorders. *Journal of the American Academy of Child & Adolescent Psychiatry, 40*(10), 1182-1189. doi: http://dx.doi.org/10.1097/00004583-200110000-00012

Crawley, S. A., Beidas, R. S., Benjamin, C. L., Martin, E., & Kendall, P. C. (2008). Treating socially phobic youth with CBT: Differential outcomes and treatment considerations. *Behavioural and Cognitive Psychotherapy, 36*(4), 379-389. doi: http://dx.doi.org/10.1017/S1352465808004542

Crepaz, N., Marshall, K. J., Aupont, L. W., Jacobs, E. D., Mizuno, Y., Kay, L. S., . . . O'Leary, A. (2009). The efficacy of HIV/STI behavioral interventions for African American females in the United States: A meta-analysis. *American Journal of Public Health, 99*(11), 2069-2078. doi: http://dx.doi.org/10.2105/AJPH.2008.139519

Cukrowicz, K. C., White, B. A., Reitzel, L. R., Burns, A. B., Driscoll, K. A., Kemper, T. S., & Joiner, T. E. (2005). Improved treatment outcome associated with the shift to empirically supported treatments in a graduate training clinic. *Professional Psychology: Research and Practice, 36*(3), 330-337. doi: http://dx.doi.org/10.1037/0735-7028.36.3.330

Current Controlled Trials. Retrieved August 1 2012, from http://www.controlled-trials.com/

Curtis, N. M., Ronan, K. R., Heiblum, N., & Crellin, K. (2009). Dissemination and effectiveness of multisystemic treatment in New Zealand: A benchmarking study.

Journal of Family Psychology, 23(2), 119-129. doi: http://dx.doi.org/10.1037/a0014974

DataStart, Inc. (1995-2011). StarStat: Significance Testing Calculator [computer software]. Retrieved August 1, 2012 from *http://www.surveystar.com/our_services/starstat.htm*.

David-Ferdon, C., & Kaslow, N. J. (2008). Evidence-based psychosocial treatments for child and adolescent depression. *Journal of Clinical Child and Adolescent Psychology.Special Issue: Evidence-Based Psychosocial Treatments for Children and Adolescents: A Ten Year Update, 37*(1), 62-62. doi: 10.1080/15374410701817865

Davidson, J., Stein, D. J., Rothbaum, B. O., Pedersen, R., Szumski, A., & Baldwin, D. S. (2012). Resilience as a predictor of treatment response in patients with posttraumatic stress disorder treated with venlafaxine extended release or placebo. *Journal of Psychopharmacology, 26*(6), 778-783. doi: http://dx.doi.org/10.1177/0269881111413821

Davis, T. E., May, A., & Whiting, S. E. (2011). Evidence-based treatment of anxiety and phobia in children and adolescents: Current status and effects on the emotional response. *Clinical Psychology Review, 31*(4), 592-602. doi: http://dx.doi.org/10.1016/j.cpr.2011.01.001

Deblinger, E., Mannarino, A. P., Cohen, J. A., Runyon, M. K., & Steer, R. A. (2011). Trauma-focused cognitive behavioral therapy for children: Impact of the trauma narrative and treatment length. *Depression and Anxiety, 28*(1), 67-75. doi: http://dx.doi.org/10.1002/da.20744

de Haan, E., Hoogduin, K. A. L., Buitelaar, J. K., & Keijsers, G. P. J. (1998). Behavior therapy versus clomipramine for the treatment of obsessive-compulsive disorder. *Journal of the American Academy of Child & Adolescent Psychiatry, 37*(10), 1022-1029. doi: http://dx.doi.org/10.1097/00004583-199810000-00011

Depp, C., & Lebowitz, B. D. (2007). Clinical trials: Bridging the gap between efficacy and effectiveness. *International Review of Psychiatry, 19*(5), 531-539. doi: http://dx.doi.org/10.1080/09540260701563320

Derzon, J. H. (2001). Antisocial behavior and the prediction of violence: A meta-analysis. *Psychology in the Schools, 38*(2), 93-106. doi: http://dx.doi.org/10.1002/pits.1002

Dobson, K. S., Hopkins, J. A., Fata, L., Scherrer, M., & Allan, L. C. (2010). The prevention of depression and anxiety in a sample of high-risk adolescents: A randomized controlled trial. *Canadian Journal of School Psychology, 25*(4), 291-310. doi: http://dx.doi.org/10.1177/0829573510386449

Doss, A. J., & Weisz, J. R. (2006). Syndrome co-occurrence and treatment outcomes in youth mental health clinics. *Journal of Consulting and Clinical Psychology, 74*(3), 416-425. doi: http://dx.doi.org/10.1037/0022-006X.74.3.416

Dumas, J. E., & Wahler, R. G. (1983). Predictors of treatment outcome in parent training: Mother insularity and socioeconomic disadvantage. *Behavioral Assessment*, *5(4),* 301-313.

Dunst, C. J., & Leet, H. E. (1987). Measuring the adequacy of resources in households with young children. *Child: Care, Health and Development, 13*(2), 111-125. doi: http://dx.doi.org/10.1111/j.1365-2214.1987.tb00528.x

Dunst, C. J., Leet, H. E., & Trivette, C. M. (1988). Family resources, personal well-being, and early intervention. *The Journal of Special Education, 22*(1), 108-116. doi: http://dx.doi.org/10.1177/002246698802200112

Durlak, J. A., Weissberg, R. P., & Pachan, M. (2010). A meta-analysis of after-school programs that seek to promote personal and social skills in children and adolescents. *American Journal of Community Psychology, 45*(3-4), 294-309. doi: http://dx.doi.org/10.1007/s10464-010-9300-6

Duval, S., & Tweedie, R. (2000). A nonparametric "trim and fill" method of accounting for publication bias in meta-analysis. *Journal of the American Statistical Association*, *95*(449), 89-98.

Duval, S., & Tweedie, R. (2000). Trim and Fill: A Simple Funnel-Plot–Based Method of Testing and Adjusting for Publication Bias in Meta-Analysis. *Biometrics*, *56*(2), 455-463.

Ebesutani, C., Bernstein, A., Nakamura, B. J., Chorpita, B. F., Higa-McMillan, C. K., & Weisz, J. R. (2010). Concurrent validity of the child behavior checklist dsm-oriented scales: correspondence with DSM diagnoses and comparison to syndrome scales. *Journal of psychopathology and behavioral assessment*, *32*(3), 373-384. http://dx.doi.org/10.1007/s10862-009-9174-9

Eddy, K. T., Dutra, L., Bradley, R., & Westen, D. (2004). A multidimensional meta-analysis of psychotherapy and pharmacotherapy for obsessive-compulsive disorder. *Clinical Psychology Review, 24*(8), 1011-1030. doi: http://dx.doi.org/10.1016/j.cpr.2004.08.004

Edelbrock, C., & Costello, A. J. (1988). Convergence between statistically derived behavior problem syndromes and child psychiatric diagnoses. *Journal of Abnormal Child Psychology, 16*(2), 219-231. doi: http://dx.doi.org/10.1007/BF00913597

Emerson, J. D., Burdick, E., Hoaglin, D. C., Mosteller, F., & Chalmers, T. C. (1990). An empirical study of the possible relation of treatment differences to quality scores in controlled randomized clinical trials. *Controlled Clinical Trials*, *11*(5), 339-352.

Epstein, M. H., Harniss, M. K., Pearson, N., & Ryser, G. (1999). The behavioral and emotional rating scale: Test-retest and inter-rater reliability. *Journal of Child and Family Studies, 8*(3), 319-327. doi: http://dx.doi.org/10.1023/A:1022067329751

Epstein, M. H. & Sharma, J. (1998). *Behavioral and Emotional Rating Scale: A strength-based approach to assessment*. Austin, TX: PRO-ED.

Epstein, N. B., Baldwin, L. M., & Bishop, D. S. (1983). The McMaster Family Assessment Device. *Journal of Marital and Family Therapy, 9*(2), 171-180. doi: http://dx.doi.org/10.1111/j.1752-0606.1983.tb01497.x

Essau, C. A., Conradt, J., & Petermann, F. (2000). Frequency, comorbidity, and psychosocial impairment of anxiety disorders in German adolescents. *Journal of anxiety disorders, 14*(3), 263-279. doi: http://dx.doi.org/10.1016/S0887-6185(99)00039-0

Ezpeleta, L., Keeler, G., Alaatin, E., Costello, E. J., & Angold, A. (2001). Epidemiology of psychiatric disability in childhood and adolescence. *Journal of Child Psychology and Psychiatry, 42*(7), 901-914. doi: http://dx.doi.org/10.1111/1469-7610.00786

Farrell, L. J., Schlup, B., & Boschen, M. J. (2010). Cognitive–behavioral treatment of childhood obsessive–compulsive disorder in community-based clinical practice:

Clinical significance and benchmarking against efficacy. *Behaviour Research and Therapy, 48*(5), 409-417. doi: http://dx.doi.org/10.1016/j.brat.2010.01.004

Fava, M., Evins, A. E., Dorer, D. J., & Schoenfeld, D. A. (2003). The problem of the placebo response in clinical trials for psychiatric disorders: Culprits, possible remedies, and a novel study design approach. *Psychotherapy and Psychosomatics, 72*(3), 115-127. doi: http://dx.doi.org/10.1159/000069738

Ferdinand, R. F. (2008). Validity of the CBCL/YSR DSM-IV scales anxiety problems and affective problems. *Journal of anxiety disorders, 22*(1), 126-134. doi: http://dx.doi.org/10.1016/j.janxdis.2007.01.008

Field, A. (2009). *Discovering statistics with SPSS, 3rd Edition.* London: Sage,

*Flannery-Schroeder, E. C., & Kendall, P. C. (2000). Group and individual cognitive–behavioral treatments for youth with anxiety disorders: A randomized clinical trial. *Cognitive Therapy and Research, 24*(3), 251-278. doi: 10.1023/A:1005500219286

Ford, T., Goodman, R., & Meltzer, H. (2003). The British Child and Adolescent Mental Health Survey 1999: The prevalence of DSM-IV disorders. *Journal of the American Academy of Child & Adolescent Psychiatry, 42*(10), 1203-1211. doi: http://dx.doi.org/10.1097/00004583-200310000-00011

Foster, E. M., Saunders, R. C., & Summerfelt, W. T. (1996). Predicting level of care in mental health services under a continuum of care. *Evaluation and Program Planning, 19*(2), 143-153. doi: http://dx.doi.org/10.1016/0149-7189(96)00005-5

Francis, G., & Holloway, J. (2007). What have we learned? Themes from the literature on best-practice benchmarking. *International Journal of Management Reviews, 9*(3), 171-189. doi: 10.1111/j.1468-2370.2007.00204.x

Franklin, M. E., Abramowitz, J. S., Kozak, M. J., Levitt, J. T., & Foa, E. B. (2000). Effectiveness of exposure and ritual prevention for obsessive-compulsive disorder: Randomized compared with nonrandomized samples. *Journal of Consulting and Clinical Psychology, 68*(4), 594-602. doi: http://dx.doi.org/10.1037/0022-006X.68.4.594

Franklin, M. E., Kozak, M. J., Cashman, L. A., Coles, M. E., Rheingold, A. A., & Foa, E. B. (1998). Cognitive-behavioral treatment of pediatric obsessive-compulsive disorder: An open clinical trial. *Journal of the American Academy of Child & Adolescent Psychiatry, 37*(4), 412-419. doi: http://dx.doi.org/10.1097/00004583-199804000-00019

Fristad, M. A. (1989). A comparison of the McMaster and circumplex family assessment instruments. *Journal of Marital and Family Therapy, 15*(3), 259-269. doi: http://dx.doi.org/10.1111/j.1752-0606.1989.tb00808.x

Garcia-Palacios, A., Hoffman, H., Carlin, A., Furness, T. A., & Botella, C. (2002). Virtual reality in the treatment of spider phobia: A controlled study. *Behaviour Research and Therapy, 40*(9), 983-993. doi: http://dx.doi.org/10.1016/S0005-7967(01)00068-7

Gardner, F., Connell, A., Trentacosta, C. J., Shaw, D. S., Dishion, T. J., & Wilson, M. N. (2009). Moderators of outcome in a brief family-centered intervention for preventing early problem behavior. *Journal of consulting and clinical psychology*, *77*(3), 543. doi: http://dx.doi.org/10.1037/a0015622

Gardner, F., Shaw, D. S., Dishion, T. J., Burton, J., & Supplee, L. (2007). Randomized prevention trial for early conduct problems: effects on proactive parenting and links to toddler disruptive behavior. *Journal of Family Psychology*, *21*(3), 398.

Garfield, S. L. (1986). Some comments on a revolution in the training of professional psychologists. *American Psychologist, 41*(10), 1175-1176. doi: http://dx.doi.org/10.1037/0003-066X.41.10.1175.b

Garland, A. F., Plemmons, D., & Koontz, L. (2006). Research-practice partnership in mental health: Lessons from participants. *Administration and Policy in Mental Health and Mental Health Services Research, 33*(5), 517-528. doi: http://dx.doi.org/10.1007/s10488-006-0062-2

Gaston, J. E., Abbott, M. J., Rapee, R. M., & Neary, S. A. (2006). Do empirically supported treatments generalize to private practice? A benchmark study of a cognitive-behavioural group treatment programme for social phobia. *British Journal of Clinical Psychology, 45*(1), 33-48. doi: 10.1348/014466505X35146

Ginsburg, G. S., Kendall, P. C., Sakolsky, D., Compton, S. N., Piacentini, J., Albano, A. M., . . . March, J. (2011). Remission after acute treatment in children and adolescents with anxiety disorders: Findings from the CAMS. *Journal of Consulting and Clinical Psychology, 79*(6), 806-813. doi: http://dx.doi.org/10.1037/a0025933

Goodman, W. K., Price, L. H., Rasmussen, S. A., & Mazure, C. (1989). The Yale-Brown obsessive compulsive scale: I. development, use, and reliability. *Archives of General Psychiatry, 46*(11), 1006-1011. doi: http://dx.doi.org/10.1001/archpsyc.1989.01810110048007

Greenland, S. (1994). Quality scores are useless and potentially misleading. *American Journal of Epidemiology, 140*, 300-301

Greenland, S., & O'Rourke, K. (2001). On the bias produced by quality scores in meta-analysis, and a hierarchical view of proposed solutions. *Biostatistics*, *2*(4), 463-471.

Gregory, A. M., Caspi, A., Moffitt, T. E., Koenen, K., Eley, T. C., & Poulton, R. (2007). Juvenile mental health histories of adults with anxiety disorders. *The American Journal of Psychiatry, 164*(2), 301-308. doi: http://dx.doi.org/10.1176/appi.ajp.164.2.301

Grimshaw, J. M., Shirran, L., Thomas, R., Mowatt, G., Fraser, C., Bero, L., . . . O'Brien, M. A. (2001). Changing provider behavior: An overview of systematic reviews of interventions. *Medical Care, 39*(8,Suppl2), II-2-II-45. doi: http://dx.doi.org/10.1097/00005650-200108002-00002

Hamerlynck, L. A. (2005). "When you pass the behavioral buck -- make it contingent" or reflection upon service in state government. *The Behavior Therapist, 28*(7), 158-159.

Hammerness, P., Geller, D., Petty, C., Lamb, A., Bristol, E., & Biederman, J. (2010). Does ADHD moderate the manifestation of anxiety disorders in children? *European Child & Adolescent Psychiatry, 19*(2), 107-112. doi: http://dx.doi.org/10.1007/s00787-009-0041-8

Han, S. S., & Weiss, B. (2005). Sustainability of teacher implementation of school-based mental health programs. *Journal of Abnormal Child Psychology, 33*(6), 665-679. doi: http://dx.doi.org/10.1007/s10802-005-7646-2

Hansen, N. B., Lambert, M. J., & Forman, E. M. (2002). The psychotherapy dose-response effect and its implications for treatment delivery services. *Clinical*

Psychology: Science and Practice, 9(3), 329-343. doi: http://dx.doi.org/10.1093/clipsy/9.3.329

Harmon, S. C., Lambert, M. J., Smart, D. M., Hawkins, E., Nielsen, S. L., Slade, K., & Lutz, W. (2007). Enhancing outcome for potential treatment failures: Therapist-client feedback and clinical support tools. *Psychotherapy Research, 17*(4), 379-392. doi: http://dx.doi.org/10.1080/10503300600702331

Harniss, M. K., Epstein, M. H., Ryser, G., & Pearson, N. (1999). The Behavioral and Emotional Rating Scale: Convergent validity. *Journal of Psychoeducational Assessment, 17*(1), 4-14. doi: http://dx.doi.org/10.1177/073428299901700101

Hedges, L. V. (1981). Distribution theory for glass's estimator of effect size and related estimators. *Journal of Educational Statistics, 6*(2), 107-128. doi: http://dx.doi.org/10.2307/1164588

Hedges, L. V., & Vevea, J. L. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods, 3*(4), 486-504. doi: http://dx.doi.org/10.1037/1082-989X.3.4.486

Heflinger, C. A., Northrup, D. A., Sonnichsen, S. E., & Brannan, A. M. (1998). Including a family focus in research on community-based services for children with serious emotional disturbance: Experiences from the Fort Bragg Evaluation Project. In M. E. Epstein, K. Kutash, & A. Duchnowski (Eds.), *Outcomes for children and youth with emotional and behavioral disorders and their families: Programs and evaluation best practices*. Austin, TX: PRO-ED Incorporated.

Herschell, A. D. (2010). Fidelity in the field: Developing infrastructure and fine-tuning measurement. *Clinical Psychology: Science and Practice, 17*(3), 253-257. doi: http://dx.doi.org/10.1111/j.1468-2850.2010.01216.x

*Heyne, D., King, N. J., Tonge, B. J., Rollings, S., Young, D., Pritchard, M., & Ollendick, T. H. (2002). Evaluation of child therapy and caregiver training in the treatment of school refusal. *Journal of the American Academy of Child & Adolescent Psychiatry, 41*(6), 687-695. doi: 10.1097/00004583-200206000-00008

Higgins, J. P. T., Altman, D. G., & Sterne, J. A. C. (Editors) (2011). Chapter 8: Assessing risk of bias in included studies. In J.P.T. Higgins & S. Green (Eds.), *Cochrane handbook for systematic reviews of interventions* (version 5.1.0). Retrieved from http://www.cochrane-handbook.org.

Higgins, J.P.T. & Green S. (Editors) (2006). Cochrane Handbook for Systematic Reviews of Interventions (version 4.2.6). Retrieved from http://www.cochrane.org/sites/default/files/uploads/Handbook4.2.6Sep2006.pdf

Hodges, K. (1990). *Child Assessment Schedule – Parent form* (3rd ed.) Ypsilanti: Eastern Michigan University.

Hodges, K. (2005). *Child and Adolescent Functional Assessment Scale*. New York, NY, US: Guilford Press.

Hodges, K., Doucette-Gates, A., & Kim, C. (2000). Predicting service utilization with the child and adolescent functional assessment scale in a sample of youths with serious emotional disturbance served by center for mental health services-funded demonstrations. *The Journal of Behavioral Health Services & Research, 27*(1), 47-59. doi: 10.1007/BF02287803

Hodges, K., Doucette-Gates, A., & Liao, Q. (1999). The relationship between the child and adolescent functional assessment scale (CAFAS) and indicators of functioning. *Journal of Child and Family Studies, 8*(1), 109-122. doi: 10.1023/A:1022902812761

Hodges, K., & Kim, C. (2000). Psychometric study of the child and adolescent functional assessment scale: Prediction of contact with the law and poor school attendance. *Journal of Abnormal Child Psychology, 28*(3), 287-297. doi: 10.1023/A:1005100521818

Hodges, K., & Wong, M. M. (1996). Psychometric characteristics of a multidimensional measure to assess impairment: The child and adolescent functional assessment scale. *Journal of Child and Family Studies, 5*(4), 445-467. doi: 10.1007/BF02233865

Hodges, K., & Wong, M. M. (1997). Use of the Child and Adolescent Functional Assessment Scale to predict service utilization and cost. *Journal of Mental Health Administration, 24*(3), 278-290. doi: 10.1007/BF02832662

Hodges, K., & Wotring, J. (2004). The role of monitoring outcomes in initiating implementation of evidence- based treatments at the state level. *Psychiatric Services, 55*(4), 396-400. doi: 10.1176/appi.ps.55.4.396

Hodges, K., Xue, Y., & Wotring, J. (2004). Outcomes for children with problematic behavior in school and at home served by public mental health. *Journal of Emotional and Behavioral Disorders, 12*(2), 109-119. doi: 10.1177/10634266040120020501

Holden, E. W., De Carolis, G., & Huff, B. (2002). Policy implications of the national evaluation of the comprehensive community mental health services for children and their families program. *Children's Services: Social Policy, Research, & Practice, 5*(1), 57-65. doi: http://dx.doi.org/10.1207/S15326918CS0501_5

Holden, E.W., Friedman, R.M., & Santiago, R.L. (2001). Overview of the National Evaluation of the Comprehensive Community Mental Health Services for Children and Their Families Program. *Journal of Emotional and Behavioral Disorders, 9*, 4-12.

Hudson, J. L., Rapee, R. M., Deveney, C., Schniering, C. A., Lyneham, H. J., & Bovopoulos, N. (2009). Cognitive-behavioral treatment versus an active control for children and adolescents with anxiety disorders: A randomized trial. *Journal of the American Academy of Child & Adolescent Psychiatry, 48*(5), 533-544. doi: http://dx.doi.org/10.1097/CHI.0b013e31819c2401

Hunsley, J. (2007). Addressing key challenges in evidence-based practice in psychology. *Professional Psychology: Research and Practice, 38*(2), 113-121. doi: 10.1037/0735-7028.38.2.113

Hunsley, J., & Lee, C. M. (2007). Research-informed benchmarks for psychological treatments: Efficacy studies, effectiveness studies, and beyond. *Professional Psychology: Research and Practice, 38*(1), 21-33. doi: http://dx.doi.org/10.1037/0735-7028.38.1.21

Hurlburt, M. S., Garland, A. F., Nguyen, K., & Brookman-Frazee, L. (2010). Child and family therapy process: Concordance of therapist and observational perspectives. *Administration and Policy in Mental Health and Mental Health Services Research, 37*(3), 230-244. doi: http://dx.doi.org/10.1007/s10488-009-0251-x

In-Albon, T., Kossowsky, J., & Schneider, S. (2010). Vigilance and avoidance of threat in the eye movements of children with separation anxiety disorder. *Journal of*

Abnormal Child Psychology, 38(2), 225-235. doi: http://dx.doi.org/10.1007/s10802-009-9359-4

In-Albon, T., & Schneider, S. (2006). Psychotherapy of childhood anxiety disorders: A meta-analysis. *Psychotherapy and Psychosomatics, 76*(1), 15-24. doi: 10.1159/000096361

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Medicine*, *2*(8), e124. doi:10.1371/journal.pmed.0020124.

Jacobson, N. S., & Christensen, A. (1996). Studying the effectiveness of psychotherapy: How well can clinical trials do the job? *American Psychologist, 51*(10), 1031-1039. doi: http://dx.doi.org/10.1037/0003-066X.51.10.1031

Jacobson, N. S., Roberts, L. J., Berns, S. B., & McGlinchey, J. B. (1999). Methods for defining and determining the clinical significance of treatment effects: Description, application, and alternatives. *Journal of Consulting and Clinical Psychology, 67*(3), 300-307. doi: http://dx.doi.org/10.1037/0022-006X.67.3.300

Jacobson, N. S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology, 59*(1), 12-19. doi: http://dx.doi.org/10.1037/0022-006X.59.1.12

Jensen, A. L., & Weisz, J. R. (2002). Assessing match and mismatch between practitioner-generated and standardized interview-generated diagnoses for clinic-referred children and adolescents. *Journal of Consulting and Clinical Psychology, 70*(1), 158-168. doi: http://dx.doi.org/10.1037/0022-006X.70.1.158

Jensen-Doss, A., & Weisz, J. R. (2008). Diagnostic agreement predicts treatment process and outcomes in youth mental health clinics. *Journal of Consulting and Clinical Psychology, 76*(5), 711-722. doi: 10.1037/0022-006X.76.5.711

Jüni, P., Witschi, A., Bloch, R., & Egger, M. (1999). The hazards of scoring the quality of clinical trials for meta-analysis. *JAMA: the Journal of the American Medical Association*, *282*(11), 1054-1060.

Kabacoff, R. I., Miller, I. W., Bishop, D. S., Epstein, N. B., & Keitner, G. I. (1990). A psychometric study of the McMaster Family Assessment Device in psychiatric, medical, and nonclinical samples. *Journal of Family Psychology, 3*(4), 431-439. doi: http://dx.doi.org/10.1037/h0080547

Karpenko, V., Owens, J. S., Evangelista, N. M., & Dodds, C. (2009). Clinically significant symptom change in children with attention-deficit/hyperactivity disorder: Does it correspond with reliable improvement in functioning? *Journal of Clinical Psychology, 65*(1), 76-93. doi: 10.1002/jclp.20549

Kazdin, A.E. & Bass, D. (1989). Power to detect differences between alternative treatments in comparative psychotherapy outcome research. *Journal of Consulting and Clinical Psychology, 57*, 138-147.

Kazdin, A. E., Esveldt-Dawson, K., French, N. H., & Unis, A. S. (1987). Effects of parent management training and problem-solving skills training combined in the treatment of antisocial child behavior. *Journal of the American Academy of Child & Adolescent Psychiatry, 26*(3), 416-424. doi: http://dx.doi.org/10.1097/00004583-198705000-00024

*Kendall, P. C. (1994). Treating anxiety disorders in children: Results of a randomized clinical trial. *Journal of Consulting and Clinical Psychology, 62*(1), 100-110. doi: 10.1037/0022-006X.62.1.100

Kendall, P. C., Brady, E. U., & Verduin, T. L. (2001). Comorbidity in childhood anxiety disorders and treatment outcome. *Journal of the American Academy of Child & Adolescent Psychiatry, 40*(7), 787-794. doi: 10.1097/00004583-200107000-00013

Kendall, P. C., Compton, S. N., Walkup, J. T., Birmaher, B., Albano, A. M., Sherrill, J., . . . Piacentini, J. (2010). Clinical characteristics of anxiety disordered youth. *Journal of anxiety disorders, 24*(3), 360-365. doi: http://dx.doi.org/10.1016/j.janxdis.2010.01.009

*Kendall, P. C., Flannery-Schroeder, E., Panichelli-Mindel, S. M., Southam-Gerow, M., Henin, A., & Warman, M. (1997). Therapy for youths with anxiety disorders: A second randomized clincal trial. *Journal of Consulting and Clinical Psychology, 65*(3), 366-380. doi: 10.1037/0022-006X.65.3.366

Kendall, P. C., Gosch, E., Furr, J. M., & Sood, E. (2008). Flexibility within fidelity. *Journal of the American Academy of Child & Adolescent Psychiatry, 47*(9), 987-993. doi: http://dx.doi.org/10.1097/CHI.0b013e31817eed2f

*Kendall, P. C., Hudson, J. L., Gosch, E., Flannery-Schroeder, E., & Suveg, C. (2008). Cognitive-behavioral therapy for anxiety disordered youth: A randomized clinical trial evaluating child and family modalities. *Journal of Consulting and Clinical Psychology, 76*(2), 282-282. doi: 10.1037/0022-006X.76.2.282

Kendall, P. C., Safford, S., Flannery-Schroeder, E., & Webb, A. (2004). Child anxiety treatment: Outcomes in adolescence and impact on substance use and depression at

7.4-year follow-up. *Journal of Consulting and Clinical Psychology, 72*(2), 276-287. doi: http://dx.doi.org/10.1037/0022-006X.72.2.276

Kendall, P. C., Settipani, C. A., & Cummings, C. M. (2012). No need to worry: The promising future of child anxiety research. *Journal of Clinical Child and Adolescent Psychology, 41*(1), 103-115. doi: http://dx.doi.org/10.1080/15374416.2012.632352

King, N. J., Heyne, D., & Ollendick, T. H. (2005). Cognitive-behavioral treatments for anxiety and phobic disorders in children and adolescents: A review. *Behavioral Disorders.Special Issue: Cognitive-Behavioral Interventions, 30*(3), 241-257.

Kilminster, S.M., & Jolly, B.C. (2000) Effective supervision in clinical practice settings: A literature review, *Medical Education, 34(10),* 827-840.

Kley, H., Heinrichs, N., Bender, C., & Tuschen-Caffier, B. (2012). Predictors of outcome in a cognitive-behavioral group program for children and adolescents with social anxiety disorder. *Journal of anxiety disorders, 26*(1), 79-87. doi: http://dx.doi.org/10.1016/j.janxdis.2011.09.002

Kobayashi, K. (2005). What limits the encoding effect of note-taking? A meta-analytic examination. *Contemporary Educational Psychology, 30*(2), 242-262. doi: http://dx.doi.org/10.1016/j.cedpsych.2004.10.001

Kovacs, M. (1992). *Children's Depression Inventory:[manual]*. Multi-Health Systems.

Kovacs, M., & Devlin, B. (1998). Internalizing disorders in childhood. *Journal of Child Psychology and Psychiatry, 39*(1), 47-63. doi: http://dx.doi.org/10.1017/S0021963097001765

Krol, Nicole P. C. M., De Bruyn, Eric E. J., Coolen, J. C., & van Aarle, Edward J. M. (2006). From CBCL to DSM: A comparison of two methods to screen for DSM-IV

diagnoses using CBCL data. *Journal of Clinical Child and Adolescent Psychology, 35*(1), 127-135. doi: http://dx.doi.org/10.1207/s15374424jccp3501_11

Kubik, M. Y., Lytle, L. A., Birnbaum, A. S., Murray, D. M., & Perry, C. L. (2003). Prevalence and correlates of depressive symptoms in young adolescents. *American Journal of Health Behavior, 27*(5), 546-553. doi: http://dx.doi.org/10.5993/AJHB.27.5.6

Lambert, M. (2007). Presidential address: What we have learned from a decade of research aimed at improving psychotherapy outcome in routine care. *Psychotherapy Research, 17*(1), 1-14. doi: http://dx.doi.org/10.1080/10503300601032506

Lambert, M. J., Hansen, N. B., & Finch, A. E. (2001). Patient-focused research: Using patient outcome data to enhance treatment effects. *Journal of Consulting and Clinical Psychology, 69*(2), 159-172. doi: http://dx.doi.org/10.1037/0022-006X.69.2.159

Lambert, M. J., Harmon, C., Slade, K., Whipple, J. L., & Hawkins, E. J. (2005). Providing feedback to psychotherapists on their patients' progress: Clinical results aod practice suggestions. *Journal of Clinical Psychology, 61*(2), 165-174. doi: http://dx.doi.org/10.1002/jclp.20113

Lambert, M. J., Hatfield, D. R., Vermeersch, D. A., Burlingame, G. M., Reisinger, C. W., & Brown, G. S. (2003). *Administration and scoring manual for the OQ-30.1*. East Setauket, NY: American Professional Credentialing Services.

Last, C. G., Hansen, C., & Franco, N. (1998). Cognitive-behavioral treatment of school phobia. *Journal of the American Academy of Child & Adolescent Psychiatry, 37*(4), 404-411. doi: http://dx.doi.org/10.1097/00004583-199804000-00018

Legerstee, J. S., Huizink, A. C., van Gastel, W., Liber, J. M., Treffers, P. D. A., Verhulst, F. C., & Utens, E. M. W. J. (2008). Maternal anxiety predicts favourable treatment outcomes in anxiety-disordered adolescents. *Acta Psychiatrica Scandinavica, 117*(4), 289-298. doi: http://dx.doi.org/10.1111/j.1600-0447.2008.01161.x

Lengua, L. J., Sadowski, C. A., Friedrich, W. N., & Fisher, J. (2001). Rationally and empirically derived dimensions of children's symptomatology: Expert ratings and confirmatory factor analyses of the CBCL. *Journal of Consulting and Clinical Psychology, 69*(4), 683-698. doi: 10.1037/0022-006X.69.4.683

Levy, K., Hunt, C., & Heriot, S. (2007). Treating comorbid anxiety and aggression in children. *Journal of the American Academy of Child & Adolescent Psychiatry, 46*(9), 1111-1118. doi: http://dx.doi.org/10.1097/chi.0b013e318074eb32

Lewczyk, C. M., Garland, A. F., Hurlburt, M. S., Gearity, J., & Hough, R. L. (2003). Comparing DISC-IV and clinician diagnoses among youths receiving public mental health services. *Journal of the American Academy of Child & Adolescent Psychiatry, 42*(3), 349-356. doi: http://dx.doi.org/10.1097/00004583-200303000-00016

Lewinsohn, P. M., Hops, H., Roberts, R. E., Seeley, J. R., & Andrews, J. A. (1993). Adolescent psychopathology: I. prevalence and incidence of depression and other DSM-III—R disorders in high school students. *Journal of Abnormal Psychology, 102*(1), 133-144. doi: http://dx.doi.org/10.1037/0021-843X.102.1.133

Lewinsohn, P. M., Zinbarg, R., Seeley, J. R., Lewinsohn, M., & Sack, W. H. (1997). Lifetime comorbidity among anxiety disorders and between anxiety disorders and other mental disorders in adolescents. *Journal of anxiety disorders, 11*(4), 377-394. doi: http://dx.doi.org/10.1016/S0887-6185(97)00017-0

Liber, J. M., Widenfelt, B. M., Leeden, A. J. M., Goedhart, A. W., Utens, E. M. W. J., & Treffers, P. D. A. (2010). The relation of severity and comorbidity to treatment outcome with cognitive behavioral therapy for childhood anxiety disorders. *Journal of Abnormal Child Psychology, 38*(5), 683-694. doi: http://dx.doi.org/10.1007/s10802-010-9394-1

Liber, J. M., Van Widenfelt, B. M., Utens, E. M. W. J., Ferdinand, R. F., Van, d. L., Van Gastel, W., & Treffers, P. D. A. (2008). No differences between group versus individual treatment of childhood anxiety disorders in a randomised clinical trial. *Journal of Child Psychology and Psychiatry, 49*(8), 886-893. doi: http://dx.doi.org/10.1111/j.1469-7610.2008.01877.

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

Lincoln, T. M., & Rief, W. (2004). How much do sample characteristics affect the effect size? An investigation of studies testing the treatment effects for social phobia. *Journal of anxiety disorders, 18*(4), 515-529. doi: http://dx.doi.org/10.1016/S0887-6185(03)00040-9

Linehan, M. M., Armstrong, H. E., Suarez, A., & Allmon, D. (1991). Cognitive-behavioral treatment of chronically parasuicidal borderline patients. *Archives of General Psychiatry, 48*(12), 1060-1064. doi: http://dx.doi.org/10.1001/archpsyc.1991.01810360024003

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

Lochman, J. E., Boxmeyer, C., Powell, N., Qu, L., Wells, K., & Windle, M. (2009). Dissemination of the coping power program: Importance of intensity of counselor training. *Journal of Consulting and Clinical Psychology, 77*(3), 397-409. doi: 10.1037/a0014514

Lorence, D. P., & Jameson, R. (2002). Adoption of information quality management practices in US healthcare organizations: A national assessment. *International Journal of Quality & Reliability Management*, *19*(6), 737-756.

Lueger, R. J., & Barkham, M. (2010). Using benchmarks and benchmarking to improve quality of practice and service. In M. Barkham, G. E. Hardy & J. Mellor-Clark (Eds.), *Practice-based evidence: A guide for psychological therapies* (pp. 223-256). Hoboken, NJ: Wiley-Blackwell.

*Lyneham, H. J., & Rapee, R. M. (2006). Evaluation of therapist-supported parent-implemented CBT for anxiety disorders in rural children. *Behaviour Research and Therapy, 44*(9), 1287-1300. doi: 10.1016/j.brat.2005.09.009

Malouff, J. M., Thorsteinsson, E. B., Rooke, S. E., Bhullar, N., & Schutte, N. S. (2008). Efficacy of cognitive behavioral therapy for chronic fatigue syndrome: A meta-analysis. *Clinical Psychology Review, 28*(5), 736-745. doi: http://dx.doi.org/10.1016/j.cpr.2007.10.004

Manteuffel, B., Stephens, R. L., & Santiago, R. (2002). Overview of the national evaluation of the Comprehensive Community Mental Health Services for Children and Their Families Program and summary of current findings. *Children's Services: Social Policy, Research, and Practice*, *5*(1), 3-20.

Manteuffel, B., Stephens, R. L., Sondheimer, D. L., & Fisher, S. K. (2008). Characteristics, service experiences, and outcomes of transition-aged youth in systems of care: programmatic and policy implications. *The Journal of Behavioral Health Services and Research*, *35*(4), 469-487.

March, J. S., Parker, J. D. A., Sullivan, K., & Stallings, P. (1997). The multidimensional anxiety scale for children (MASC): Factor structure, reliability, and validity. *Journal of the American Academy of Child & Adolescent Psychiatry, 36*(4), 554-565. doi: http://dx.doi.org/10.1097/00004583-199704000-00019

Martino, S., Ball, S. A., Nich, C., Frankforter, T. L., & Carroll, K. M. (2008). Community program therapist adherence and competence in motivational enhancement therapy. *Drug and Alcohol Dependence, 96*(1-2), 37-48. doi: http://dx.doi.org/10.1016/j.drugalcdep.2008.01.020

Martinsen, E. W., Olsen, T., Tønset, E., Nyland, K. E., & Aarre, T. F. (1998). Cognitive-behavioral group therapy for panic disorder in the general clinical setting: a naturalistic study with 1-year follow-up. *The Journal of Clinical Psychiatry*, *59*(8), 437.

Mazzucchelli, T. G., & Sanders, M. R. (2010). Facilitating practitioner flexibility within an empirically supported intervention: Lessons from a system of parenting support. *Clinical Psychology: Science and Practice, 17*(3), 238-252. doi: http://dx.doi.org/10.1111/j.1468-2850.2010.01215.x

McEvoy, P. M., & Nathan, P. (2007). Effectiveness of cognitive behavior therapy for diagnostically heterogeneous groups: A benchmarking study. *Journal of Consulting and Clinical Psychology, 75*(2), 344-350. doi: 10.1037/0022-006X.75.2.344

McEvoy, P. M., Nathan, P., Rapee, R. M., & Campbell, B. N. C. (2012). Cognitive behavioural group therapy for social phobia: Evidence of transportability to community clinics. *Behaviour Research and Therapy, 50*(4), 258-265. doi: http://dx.doi.org/10.1016/j.brat.2012.01.009

Mellor-Clark, J. (2006). Developing CORE performance indicators for benchmarking in NHS primary care psychological therapy and counselling services: An editorial introduction. *Counselling & Psychotherapy Research, 6*(1), 1-2. doi: 10.1080/14733140600581176

Merrill, K. A., Tolbert, V. E., & Wade, W. A. (2003). Effectiveness of cognitive therapy for depression in a community mental health center: A benchmarking study. *Journal of Consulting and Clinical Psychology, 71*(2), 404-409. doi: http://dx.doi.org/10.1037/0022-006X.71.2.404

Miller, I. W., Epstein, N. B., Bishop, D. S., & Keitner, G. I. (1985). The McMaster Family Assessment Device: Reliability and validity. *Journal of Marital and Family Therapy, 11*(4), 345-356. doi: http://dx.doi.org/10.1111/j.1752-0606.1985.tb00028.x

Minami, T., Davies, D. R., Tierney, S. C., Bettmann, J. E., McAward, S. M., Averill, L. A., . . . Wampold, B. E. (2009). Preliminary evidence on the effectiveness of psychological treatments delivered at a university counseling center. *Journal of Counseling Psychology, 56*(2), 309-320. doi: http://dx.doi.org/10.1037/a0015398

Minami, T., Serlin, R. C., Wampold, B. E., Kircher, J. C., & Brown, G. S. (2008). Using clinical trials to benchmark effects produced in clinical practice. *Quality & Quantity: International Journal of Methodology, 42*(4), 513-525. doi: http://dx.doi.org/10.1007/s11135-006-9057-z

Minami, T., Wampold, B. E., Serlin, R. C., Hamilton, E. G., Brown, G. S. & Kircher, J. C. (2008). Benchmarking the effectiveness of psychotherapy treatment for adult depression in a managed care environment: A preliminary study. *Journal of Consulting and Clinical Psychology, 76*(1), 116-124. doi: 10.1037/0022-006X.76.1.116

Minami, T., Wampold, B. E., Serlin, R. C., Kircher, J. C., & Brown, G. S. (2007). Benchmarks for psychotherapy efficacy in adult major depression. *Journal of Consulting and Clinical Psychology, 75*(2), 232-243. doi: http://dx.doi.org/10.1037/0022-006X.75.2.232

Moher, D., Pham, B., Jones, A., Cook, D. J., Jadad, A. R., Moher, M., ... & Klassen, T. P. (1998). Does quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analyses? *The Lancet*, *352*(9128), 609-613.

Morris, S. B. (2000). Distribution of the standardized mean change effect size for meta-analysis on repeated measures. *British Journal of Mathematical and Statistical Psychology*, *53*(1), 17-29.

Mufson, L., Dorta, K. P., Wickramaratne, P., Nomura, Y., Olfson, M., & Weissman, M. M. (2004). A randomized effectiveness trial of interpersonal psychotherapy for depressed adolescents. *Archives of General Psychiatry*, *61*(6), 577.

Muris, P., Meesters, C., & Gobel, M. (2002). Cognitive coping versus emotional disclosure in the treatment of anxious children: A pilot-study. *Cognitive Behaviour Therapy, 31*(2), 59-67. doi: http://dx.doi.org/10.1080/16506070252959490

Muris, P., Merckelbach, H., Holdrinet, I., & Sijsenaar, M. (1998). Treating phobic children: Effects of EMDR versus exposure. *Journal of Consulting and Clinical Psychology, 66*(1), 193-198. doi: 10.1037/0022-006X.66.1.193

Nakamura, B. J., Ebesutani, C., Bernstein, A., & Chorpita, B. F. (2009). A psychometric analysis of the child behavior checklist DSM-oriented scales. *Journal of Psychopathology and Behavioral Assessment, 31*(3), 178-189. doi: http://dx.doi.org/10.1007/s10862-008-9119-8

Nathan, P. E., & Gorman, J. M. (2002). Efficacy, effectiveness, and the clinical utility of psychotherapy research. In P. E. Nathan & J. M. Gorman (Eds.), *A guide to treatments that work* (2nd ed., 643–654). New York: Oxford University Press.

*Nauta, M. H., Scholing, A., Emmelkamp, P. M. G., & Minderaa, R. B. (2003). Cognitive-behavioral therapy for children with anxiety disorders in a clinical setting: No additional effect of a cognitive parent training. *Journal of the American Academy of Child & Adolescent Psychiatry, 42*(11), 1270-1278. doi: 10.1097/01.chi.0000085752.71002.93

Ng, R. M. K. (2005). Cognitive therapy supervision--A pilot study. *Hong Kong Journal of Psychiatry, 15*(4), 122-126.

Norton, P. J., & Price, E. C. (2007). A meta-analytic review of adult cognitive-behavioral treatment outcome across the anxiety disorders. *Journal of Nervous and Mental Disease, 195*(6), 521-531. doi: http://dx.doi.org/10.1097/01.nmd.0000253843.70149.9a

Oei, T. P. S., & Boschen, M. J. (2009). Clinical effectiveness of a cognitive behavioral group treatment program for anxiety disorders: A benchmarking study. *Journal of*

anxiety disorders, 23(7), 950-957. doi: http://dx.doi.org/10.1016/j.janxdis.2009.06.004

Ollendick, T. H., Jarrett, M. A., Grills-Taquechel, A. E., Hovey, L. D., & Wolff, J. C. (2008). Comorbidity as a predictor and moderator of treatment outcome in youth with anxiety, affective, attention deficit/hyperactivity disorder, and oppositional/conduct disorders. *Clinical Psychology Review, 28*(8), 1447-1471. doi: 10.1016/j.cpr.2008.09.003

Ollendick, T. H., & King, N. J. (1994). Diagnosis, assessment, and treatment of internalizing problems in children: The role of longitudinal data. *Journal of Consulting and Clinical Psychology, 62*(5), 918-927. doi: 10.1037/0022-006X.62.5.918

Ollendick, T. H., Öst, L., Reuterskiöld, L., & Costa, N. (2010). Comorbidity in youth with specific phobias: Impact of comorbidity on treatment outcome and the impact of treatment on comorbid disorders. *Behaviour Research and Therapy, 48*(9), 827-831. doi: http://dx.doi.org/10.1016/j.brat.2010.05.024

Parry, G., Roth, A. D., & Kerr, I. B. (2005). Brief and time-limited psychotherapy. In G. O. Gabbard, J. S. Beck & J. Holmes (Eds.), *Oxford textbook of psychotherapy*. (pp. 507-521). New York, NY, US: Oxford University Press.

Peduzzi, P., Concato, J., Kemper, E., Holford, T. R., & Feinstein, A. R. (1996). A simulation study of the number of events per variable in logistic regression analysis. *Journal of Clinical Epidemiology*, *49*(12), 1373-1379

Persons, J. B. (1997). Dissemination of effective methods: Behavior therapy's next challenge. *Behavior Therapy, 28*(3), 465-471. doi: http://dx.doi.org/10.1016/S0005-7894(97)80096-2

Persons, J. B., Bostrom, A., & Bertagbolli, A. (1999). Results of randomized controlled trials of cognitive therapy for depression generalize to private practice. *Cognitive Therapy and Research, 23*(5), 535-548. doi: http://dx.doi.org/10.1023/A:1018724505659

Pina, A. A., Silverman, W. K., Fuentes, R. M., Kurtines, W. M., & Weems, C. F. (2003). Exposure-based cognitive-behavioral treatment for phobic and anxiety disorders: Treatment effects and maintenance for Hispanic/Latino relative to European-American youths. *Journal of the American Academy of Child & Adolescent Psychiatry, 42*(10), 1179-1187. doi: http://dx.doi.org/10.1097/00004583-200310000-00008

Pine, D. S., Cohen, P., Gurley, D., Brook, J., & Ma, Y. (1998). The risk for early-adulthood anxiety and depressive disorders in adolescents with anxiety and depressive disorders. *Archives of General Psychiatry, 55*(1), 56-64. doi: http://dx.doi.org/10.1001/archpsyc.55.1.56

Prins, P. J., & Ollendick, T. H. (2003). Cognitive change and enhanced coping: Missing mediational links in cognitive behavior therapy with anxiety-disordered children. *Clinical Child and Family Psychology Review*, *6*(2), 87-105. http://dx.doi.org.ezproxy.lakeheadu.ca/10.1023/A:1023730526716

Quay, H. C., & Peterson, D. R. (1993). *The Revised Behavior Problem Checklist: Manual*. Odessa, FL: Psychological Assessment Resources.

Quist, R. M., & Matshazi, D. G. M. (2000). The Child and Adolescent Functional Assessment Scale (CAFAS): A dynamic predictor of juvenile recidivism. *Adolescence, 35*(137), 181-192.

Rapee, R. M. (2003). The influence of comorbidity on treatment outcome for children and adolescents with anxiety disorders. *Behaviour Research and Therapy, 41*(1), 105-112. doi: http://dx.doi.org/10.1016/S0005-7967(02)00049-9

*Rapee, R. M., Abbott, M. J., & Lyneham, H. J. (2006). Bibliotherapy for children with anxiety disorders using written materials for parents: A randomized controlled trial. *Journal of Consulting and Clinical Psychology, 74*(3), 436-444. doi: 10.1037/0022-006X.74.3.436

Rapee, R. M., Schniering, C. A., & Hudson, J. L. (2009). Anxiety disorders during childhood and adolescence: Origins and treatment. *Annual Review of Clinical Psychology, 5*, 311-341. doi: http://dx.doi.org/10.1146/annurev.clinpsy.032408.153628

Rettew, D. C., Lynch, A. D., Achenbach, T. M., Dumenci, L., & Ivanova, M. Y. (2009). Meta-analyses of agreement between diagnoses made from clinical evaluations and standardized diagnostic interviews. *International Journal of Methods in Psychiatric Research, 18*(3), 169-184. doi: http://dx.doi.org/10.1002/mpr.289

Reyno, S. M., & McGrath, P. J. (2006). Predictors of parent training efficacy for child externalizing behavior problem--a meta-analytic review. *Journal of Child Psychology and Psychiatry, 47*(1), 99-111. doi: http://dx.doi.org/10.1111/j.1469-7610.2005.01544.x

Reynolds, C. R., & Richmond, B. O. (1985). *Revised Children's Manifest Anxiety Scale*. Los Ángeles: Western Psychological Services.

Reynolds, S., Wilson, C., Austin, J., & Hooper, L. (2012). Effects of psychotherapy for anxiety in children and adolescents: a meta-analytic review. *Clinical Psychology Review, 32*(4), 251-262. doi:10.1016/j.cpr.2012.01.005

Richardson, L. P., Russo, J. E., Lozano, P., McCauley, E., & Katon, W. (2010). Factors associated with detection and receipt of treatment for youth with depression and anxiety disorders. *Academic Pediatrics, 10(1),* 36-40.

Rief, W., Nestoriuc, Y., Weiss, S., Welzel, E., Barsky, A. J., & Hofmann, S. G. (2009). Meta-analysis of the placebo response in antidepressant trials. *Journal of Affective Disorders, 118*(1-3), 1-8. doi: http://dx.doi.org/10.1016/j.jad.2009.01.029

Rosenblatt, A., & Rosenblatt, J. A. (2002). Assessing the effectiveness of care for youth with severe emotional disturbances: Is there agreement between popular outcome measures? *The Journal of Behavioral Health Services & Research, 29*(3), 259-273. doi: http://dx.doi.org/10.1007/BF02287367

Roza, S. J., Hofstra, M. B., van der Ende, J., & Verhulst, F. C. (2003). Stable prediction of mood and anxiety disorders based on behavioral and emotional problems in childhood: A 14-year follow-up during childhood, adolescence, and young adulthood. *The American Journal of Psychiatry, 160*(12), 2116-2121. doi: http://dx.doi.org/10.1176/appi.ajp.160.12.2116

Sackett, D. L. (1979). Bias in analytic research. *Journal of Chronic Diseases*, *32*(1-2), 51-63. doi:10.1016/0021-9681(79)90012-2

Sanderson, W. C., Raue, P. J., & Wetzler, S. (1998). The generalizability of cognitive behavior therapy for panic disorder. *Journal of Cognitive Psychotherapy, 12*(4), 323-330.

Schindler, A. C., Hiller, W., & Witthöft, M. (2011). Benchmarking of cognitive-behavioral therapy for depression in efficacy and effectiveness studies—How do exclusion criteria affect treatment outcome? *Psychotherapy Research, 21*(6), 644-657. doi: http://dx.doi.org/10.1080/10503307.2011.602750

Schoenwald, S. K., Chapman, J. E., Kelleher, K., Hoagwood, K. E., Landsverk, J., Stevens, J., . . . Rolls-Reutz, J. (2008). A survey of the infrastructure for children's mental health services: Implications for the implementation of empirically supported treatments (ESTs). *Administration and Policy in Mental Health and Mental Health Services Research, 35*(1-2), 84-97. doi: http://dx.doi.org/10.1007/s10488-007-0147-6

Schoenwald, S. K., & Hoagwood, K. (2001). Effectiveness, transportability, and dissemination of interventions: What matters when? *Psychiatric Services, 52*(9), 1190-1197. doi: http://dx.doi.org/10.1176/appi.ps.52.9.1190

Schoenwald, S. K., Sheidow, A. J., & Chapman, J. E. (2009). Clinical supervision in treatment transport: Effects on adherence and outcomes. *Journal of Consulting and Clinical Psychology, 77*(3), 410-421. doi: http://dx.doi.org/10.1037/a0013788

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*(4), 223-239. doi: http://dx.doi.org/10.1023/A:1026229102151

Seligman, L. D., Ollendick, T. H., Langley, A. K., & Baldacci, H. B. (2004). The utility of measures of child and adolescent anxiety: A meta-analytic review of the revised children's anxiety scale, the state-trait anxiety inventory for children, and the child behavior checklist. *Journal of Clinical Child and Adolescent Psychology, 33*(3), 557-565. doi: http://dx.doi.org/10.1207/s15374424jccp3303_13

Serlin, R. C., & Lapsley, D. K. (1985). Rationality in psychological research: The good-enough principle. *American Psychologist, 40*(1), 73-83. doi: http://dx.doi.org/10.1037/0003-066X.40.1.73

Serlin, R. C., & Lapsley, D. K. (1993). Rational appraisal of psychological research and the good-enough principle. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 199–228). Hillsdale, NJ: Erlbaum.

Shadish, W. R., & Baldwin, S. A. (2005). Effects of behavioral marital therapy: A meta-analysis of randomized controlled trials. *Journal of Consulting and Clinical Psychology, 73*(1), 6-14. doi: http://dx.doi.org/10.1037/0022-006X.73.1.6

Shadish, W. R., & Ragsdale, K. (1996). Random versus nonrandom assignment in controlled experiments: Do you get the same answer? *Journal of Consulting and Clinical Psychology, 64*(6), 1290-1305. doi: 10.1037/0022-006X.64.6.1290

Shaffer, D., Fisher, P., Lucas, C. P., Dulcan, M. K., & Schwab-Stone, M. E. (2000). NIMH diagnostic interview schedule for children version IV (NIMH DISC-IV): Description, differences from previous versions, and reliability of some common diagnoses. *Journal of the American Academy of Child & Adolescent Psychiatry, 39*(1), 28-38. doi: http://dx.doi.org/10.1097/00004583-200001000-00014

Shirk, S. R., Kaplinski, H., & Gudmundsen, G. (2009). School-based cognitive-behavioral therapy for adolescent depression: A benchmarking study. *Journal of Emotional and Behavioral Disorders, 17*(2), 106-117.

Sholomskas, D. E., Syracuse-Siewert, G., Rounsaville, B. J., Ball, S. A., Nuro, K. F., & Carroll, K. M. (2005). We don't train in vain: A dissemination trial of three strategies of training clinicians in cognitive-behavioral therapy. *Journal of Consulting and Clinical Psychology, 73*(1), 106-115. doi: http://dx.doi.org/10.1037/0022-006X.73.1.106

Shortt, A. L., Barrett, P. M., & Fox, T. L. (2001). Evaluating the FRIENDS program: A cognitive-behavioral group treatment for anxious children and their parents. *Journal of Clinical Child Psychology, 30*(4), 525-535.

Silverman, W. (1987). *Anxiety Disorders Interview Schedule for Children (ADIS).* State University of New York at Albany: Graywind Publications.

Silverman, W. K., & Albano, A. M. (1996). *Anxiety Disorders Interview Schedule for DSM-IV*. Oxford University Press.

*Silverman, W. K., Kurtines, W. M., Ginsburg, G. S., Weems, C. F., Lumpkin, P. W., & Carmichael, D. H. (1999a). Treating anxiety disorders in children with group cognitive-behavioral therapy: A randomized clinical trial. *Journal of Consulting and Clinical Psychology, 67*(6), 995-1003. doi: 10.1037/0022-006X.67.6.995

*Silverman, W. K., Kurtines, W. M., Ginsburg, G. S., Weems, C. F., Rabian, B., & Serafini, L. T. (1999b). Contingency management, self-control, and education support in the treatment of childhood phobic disorders: A randomized clinical trial.

Journal of Consulting and Clinical Psychology, 67(5), 675-687. doi: 10.1037/0022-006X.67.5.675

Silverman, W. K., & Nelles, W. B. (1988). The Anxiety Disorders Interview Schedule For Children. *Journal of the American Academy of Child & Adolescent Psychiatry, 27*(6), 772-778. doi: http://dx.doi.org/10.1097/00004583-198811000-00019

Silverman, W. K., & Ollendick, T. H. (2005). Evidence-based assessment of anxiety and its disorders in children and adolescents. *Journal of Clinical Child and Adolescent Psychology, 34*(3), 380-411. doi: http://dx.doi.org/10.1207/s15374424jccp3403_2

Silverman, W. K., Ortiz, C. D., Viswesvaran, C., Burns, B. J., Kolko, D. J., Putnam, F. W., & Amaya-Jackson, L. (2008). Evidence-based psychosocial treatments for children and adolescents exposed to traumatic events. *Journal of Clinical Child and Adolescent Psychology, 37*(1), 156-183. doi: http://dx.doi.org/10.1080/15374410701818293

Silverman, W. K., Pina, A. A., & Viswesvaran, C. (2008). Evidence-based psychosocial treatments for phobic and anxiety disorders in children and adolescents. *Journal of Clinical Child and Adolescent Psychology, 37*(1), 105-130. doi: http://dx.doi.org/10.1080/15374410701817907

Southam-Gerow, M. A., Chorpita, B. F., Miller, L. M., & Gleacher, A. A. (2008). Are children with anxiety disorders privately referred to a university clinic like those referred from the public mental health system? *Administration and Policy in Mental Health and Mental Health Services Research, 35*(3), 168-180. doi: http://dx.doi.org/10.1007/s10488-007-0154-7

Southam-Gerow, M. A., Kendall, P. C., & Weersing, V. R. (2001). Examining outcome variability: Correlates of treatment response in a child and adolescent anxiety clinic. *Journal of Clinical Child Psychology, 30*(3), 422-436. doi: http://dx.doi.org/10.1207/S15374424JCCP3003_13

Southam-Gerow, M. A., Rodríguez, A., Chorpita, B. F., & Daleiden, E. L. (2012). Dissemination and implementation of evidence based treatments for youth: Challenges and recommendations. *Professional Psychology: Research and Practice, 43*(5), 527-534. doi: http://dx.doi.org/10.1037/a0029101

*Southam-Gerow, M. A., Weisz, J. R., Chu, B. C., McLeod, B. D., Gordis, E. B., & Connor-Smith, J. K. (2010). Does cognitive behavioral therapy for youth anxiety outperform usual care in community clinics? An initial effectiveness test. *Journal of the American Academy of Child & Adolescent Psychiatry, 49*(10), 1043-1052. doi: http://dx.doi.org/10.1016/j.jaac.2010.06.009

Southam-Gerow, M. A., Weisz, J. R., & Kendall, P. C. (2003). Youth with anxiety disorders in research and service clinics: Examining client differences and similarities. *Journal of Clinical Child and Adolescent Psychology, 32*(3), 375-385. doi: http://dx.doi.org/10.1207/S15374424JCCP3203_06

Spence, S. H., Donovan, C., & Brechman-Toussaint, M. (2000). The treatment of childhood social phobia: The effectiveness of a social skills training-based, cognitive-behavioural intervention, with and without parental involvement. *Journal of Child Psychology and Psychiatry, 41*(6), 713-726. doi: 10.1111/1469-7610.00659

*Spence, S. H., Holmes, J. M., March, S., & Lipp, O. V. (2006). The feasibility and outcome of clinic plus internet delivery of cognitive-behavior therapy for childhood

anxiety. *Journal of Consulting and Clinical Psychology, 74*(3), 614-621. doi: http://dx.doi.org/10.1037/0022-006X.74.3.614

Sperry, L., Brill, P.L., Howard, K.I., & Grissom, G.R. (1996). *Treatment outcomes in psychotherapy and psychiatric interventions*. New York: Brunner/ Mazel.

Stephens, R. L., Holden, E. W., & Hernandez, M. (2004). System-of-care practice review scores as predictors of behavioral symptomatology and functional impairment. *Journal of Child and Family Studies, 13*(2), 179-191. doi: http://dx.doi.org/10.1023/B:JCFS.0000015706.77407.cb

Sterne, J. A., Gavaghan, D., & Egger, M. (2000). Publication and related bias in meta-analysis: power of statistical tests and prevalence in the literature. *Journal of Clinical Epidemiology*, *53*(11), 1119-1129.

Stirman, S. W., DeRubeis, R. J., Crits-Christoph, P., & Brody, P. E. (2003). Are samples in randomized controlled trials of psychotherapy representative of community outpatients? A new methodology and initial findings. *Journal of Consulting and Clinical Psychology, 71*(6), 963-972. doi: http://dx.doi.org/10.1037/0022-006X.71.6.963

Storch, E. A., Merlo, L. J., Larson, M. J., Geffken, G. R., Lehmkuhl, H. D., Jacob, M. L., . . . Goodman, W. K. (2008). Impact of comorbidity on cognitive-behavioral therapy response in pediatric obsessive-compulsive disorder. *Journal of the American Academy of Child & Adolescent Psychiatry, 47*(5), 583-592. doi: http://dx.doi.org/10.1097/CHI.0b013e31816774b1

Strauss, C. C., Frame, C. L., & Forehand, R. (1987). Psychosocial impairment associated with anxiety in children. *Journal of Clinical Child Psychology, 16*(3), 235-239. doi: http://dx.doi.org/10.1207/s15374424jccp1603_8

Stuart, G. L., Treat, T. A., & Wade, W. A. (2000). Effectiveness of an empirically based treatment for panic disorder delivered in a service clinic setting: 1-year follow-up. *Journal of Consulting and Clinical Psychology, 68*(3), 506-512. doi: http://dx.doi.org/10.1037/0022-006X.68.3.506

Swanson, J. M., Kraemer, H. C., Hinshaw, S. P., Arnold, L. E., Conners, C. K., Abikoff, H. B., . . . Wu, M. (2001). Clinical relevance of the primary findings of the MTA: Success rates based on severity of ADHD and ODD symptoms at the end of treatment. *Journal of the American Academy of Child & Adolescent Psychiatry, 40*(2), 168-179. Retrieved from http://ezproxy.lakeheadu.ca/login?url=http://search.proquest.com/docview/619655713?accountid=11956

Swets, J.A., & Pickett, R.M. (1982). *Evaluation of diagnostic systems*. New York: Academic Press.

Tabachnick, B. G., & Fidell, L. S. (2013). *Using Multivariate Statistics, 6h Edition*. Boston, MA: Pearson Education.

Task Force On Promotion And Dissemination Of Psychological Procedures. Division of Clinical Psychology, American Psychological Association. (1995). Training in and dissemination of empirically validated psychological treatments: Report and recommendations. *The Clinical Psychologist, 48(1),* 3-23.

Taylor, T. K., Schmidt, F., Pepler, D., & Hodgins, C. (1998). A comparison of eclectic treatment with Webster-Stratton's parents and children series in a children's mental health center: A randomized controlled trial. *Behavior Therapy, 29*(2), 221-240. doi: 10.1016/S0005-7894(98)80004-X

Trask, E. V., & Garland, A. F. (2012). Are children improving? Results from outcome measurement in a large mental health system. *Administration and Policy in Mental Health and Mental Health Services Research, 39*(3), 210-220. doi: http://dx.doi.org/10.1007/s10488-011-0353-0

Treadwell, K. R. H., Flannery-Schroeder, E. C., & Kendall, P. C. (1995). Ethnicity and gender in relation to adaptive functioning, diagnostic status, and treatment outcome in children from an anxiety clinic. *Journal of Anxiety Disorders, 9*(5), 373-384. doi: http://dx.doi.org/10.1016/0887-6185(95)00018-J

Trosa, S. & Williams, S. (1996). Benchmarking in public sector performance management. *Performance Measurement in Government.* OECD Occasional Papers No. 9. Paris: OECD.

Tuschen-Caffier, B., Pook, M., & Frank, M. (2001). Evaluation of manual-based cognitive-behavioral therapy for bulimia nervosa in a service setting. *Behaviour Research and Therapy, 39*(3), 299-308. doi: http://dx.doi.org/10.1016/S0005-7967(00)00004-8

Verdeli, H., Mufson, L., Lee, L., & Keith, J. A. (2006). Review of evidence-based psychotherapies for pediatric mood and anxiety disorders. *Current Psychiatry Reviews, 2*(3), 395-421. doi: http://dx.doi.org/10.2174/157340006778018102

Wade, W. A., Treat, T. A., & Stuart, G. L. (1998). Transporting an empirically supported treatment for panic disorder to a service clinic setting: A benchmarking strategy. *Journal of Consulting and Clinical Psychology, 66*(2), 231-239. doi: http://dx.doi.org/10.1037/0022-006X.66.2.231

Walrath, C. M., Mandell, D. S., & Leaf, P. J. (2001). Responses of children with different intake profiles to mental health treatment. *Psychiatric Services, 52*(2), 196-201. doi: http://dx.doi.org/10.1176/appi.ps.52.2.196

Warren, R., & Thomas, J. C. (2001). Cognitive–behavior therapy of obsessive–compulsive disorder in private practice: An effectiveness study. *Journal of Anxiety Disorders, 15*(4), 277-285. doi: http://dx.doi.org/10.1016/S0887-6185(01)00063-9

Weersing, V. R. (2005). Benchmarking the effectiveness of psychotherapy: Program evaluation as a component of evidence-based practice. *Journal of the American Academy of Child & Adolescent Psychiatry, 44*(10), 1058-1062. doi: 10.1097/01.chi.0000172682.71384.80

Weersing, V. R., Iyengar, S., Kolko, D. J., Birmaher, B., & Brent, D. A. (2006). Effectiveness of cognitive-behavioral therapy for adolescent depression: A benchmarking investigation. *Behavior Therapy, 37*(1), 36-48. doi: 10.1016/j.beth.2005.03.003

Weersing, V. R., & Weisz, J. R. (2002). Community clinic treatment of depressed youth: Benchmarking usual care against CBT clinical trials. *Journal of Consulting and Clinical Psychology, 70*(2), 299-310. doi: 10.1037/0022-006X.70.2.299

Weiss, B., Catron, T., Harris, V., & Phung, T. M. (1999). The effectiveness of traditional child psychotherapy. *Journal of Consulting and Clinical Psychology, 67*(1), 82-94. doi: 10.1037/0022-006X.67.1.82

Weisz, J. (2004). *Psychotherapy for children and adolescents: evidence based treatments and case examples*. Cambridge: Cambridge University Press.

Weisz, J. R., Chorpita, B. F., Palinkas, L. A., Schoenwald, S. K., Miranda, J., Bearman, S. K., . . . Gibbons, R. D. (2012). Testing standard and modular designs for psychotherapy treating depression, anxiety, and conduct problems in youth: A randomized effectiveness trial. *Archives of General Psychiatry, 69*(3), 274-282. doi: http://dx.doi.org/10.1001/archgenpsychiatry.2011.147

Weisz, J. R., Donenberg, G. R., Han, S. S., & Weiss, B. (1995). Bridging the gap between laboratory and clinic in child and adolescent psychotherapy. *Journal of Consulting and Clinical Psychology, 63*(5), 688-701. doi: 10.1037/0022-006X.63.5.688

Weisz, J. R., Doss, A. J., & Hawley, K. M. (2005). Youth psychotherapy outcome research: A review and critique of the evidence base. *Annual Review of Psychology, 56*, 337-363. doi: http://dx.doi.org/10.1146/annurev.psych.55.090902.141449

Weisz, J. R., & Gray, J. S. (2008). Evidence-based psychotherapy for children and adolescents: Data from the present and a model for the future. *Child and Adolescent Mental Health, 13*(2), 54-65. doi: http://dx.doi.org/10.1111/j.1475-3588.2007.00475.x

Weisz, J. R., Hawley, K. M., & Doss, A. J. (2004). Empirically tested psychotherapies for youth internalizing and externalizing problems and disorders. *Child and Adolescent*

Psychiatric Clinics of North America, 13(4), 729-815. doi: http://dx.doi.org/10.1016/j.chc.2004.05.006

Weisz, J. R., Jensen-Doss, A., & Hawley, K. M. (2006). Evidence-based youth psychotherapies versus usual clinical care: A meta-analysis of direct comparisons. *American Psychologist, 61*(7), 671-671. doi: 10.1037/0003-066X.61.7.671

Weisz, J. R., Sandler, I. N., Durlak, J. A., & Anton, B. S. (2005). Promoting and protecting youth mental health through evidence-based prevention and treatment. *American Psychologist, 60*(6), 628-648. doi: 10.1037/0003-066X.60.6.628

Weisz, J. R., Southam-Gerow, M. A., Gordis, E. B., Connor-Smith, J. K., Chu, B. C., Langer, D. A., . . . Weiss, B. (2009). Cognitive–behavioral therapy versus usual clinical care for youth depression: An initial test of transportability to community clinics and clinicians. *Journal of Consulting and Clinical Psychology, 77*(3), 383-396. doi: http://dx.doi.org/10.1037/a0013877

Weisz, J. R., Weiss, B., & Donenberg, G. R. (1992). The lab versus the clinic: Effects of child and adolescent psychotherapy. *American Psychologist, 47*(12), 1578-1585. doi: http://dx.doi.org/10.1037/0003-066X.47.12.1578

Westbrook, D., & Kirk, J. (2005). The clinical effectiveness of cognitive behaviour therapy: Outcome for a large sample of adults treated in routine practice. *Behaviour Research and Therapy, 43*(10), 1243-1261. doi: http://dx.doi.org/10.1016/j.brat.2004.09.006

Westen, D., & Morrison, K. (2001). A multidimensional meta-analysis of treatments for depression, panic, and generalized anxiety disorder: An empirical examination of the

status of empirically supported therapies. *Journal of Consulting and Clinical Psychology, 69*(6), 875-899. doi: http://dx.doi.org/10.1037/0022-006X.69.6.875

Westen, D., Novotny, C. M., & Thompson-Brenner, H. (2004). The empirical status of empirically supported psychotherapies: Assumptions, findings, and reporting in controlled clinical trials. *Psychological Bulletin, 130*(4), 631-663. doi: http://dx.doi.org/10.1037/0033-2909.130.4.631

Wise, E. A. (2004). Methods for analyzing psychotherapy outcomes: A review of clinical significance, reliable change, and recommendations for future directions. *Journal of Personality Assessment, 82*(1), 50-59. doi: http://dx.doi.org/10.1207/s15327752jpa8201_10

Wolfe, R., & Hanley, J. (2002). If we're so different, why do we keep overlapping? When 1 plus 1 doesn't make 2. *Canadian Medical Association Journal*, *166*(1), 65-66.

Wood, J. J., Piacentini, J. C., Southam-Gerow, M., Chu, B. C., & Sigman, M. (2006). Family cognitive behavioral therapy for child anxiety disorders. *Journal of the American Academy of Child & Adolescent Psychiatry, 45*(3), 314-321. doi: http://dx.doi.org/10.1097/01.chi.0000196425.88341.b0

Yoshikawa, T., Innes, J., Mitchel, F., & Tanaka, M. (1993). *Contemporary Cost Management*. London: Chapman and Hall.

Zayas, L. H., Drake, B., & Jonson-Reid, M. (2011). Overrating or dismissing the value of evidence-based practice: Consequences for clinical practice. *Clinical Social Work Journal, 39*(4), 400-405. doi: http://dx.doi.org/10.1007/s10615-010-0306-1

Appendix A

Lakehead
UNIVERSITY

Office of Research Services

Tel 807-343-8934
Fax 807-346-7749

February 17, 2010

Ms Carolyn Houlding
Doctoral student
Department of Psychology
Lakehead University
955 Oliver Rd
Thunder Bay ON P7B 5E1

Dear Ms Houlding:

Thank you for providing the Lakehead University Research Ethics Board the opportunity to review your dissertation entitled, "Benchmarking the effectiveness of community services for children with behavioural and emotional disorders".

Your intention to access a data set consisting of mental health outcomes pre-treatment and at 6 month intervals after the commencement of treatment for children receiving services from community mental health agencies has been reviewed by the Research Ethics Board. The Board has deemed that your project does not require ethical review.

On behalf of the Lakehead University Research Ethics Board, I wish you success with your research.

Sincerely,

Richard Maundrell
Chair, Research Ethics Board

/sw