2020

A meta-analysis investigating the correlation between treatment integrity and youth client outcomes

Ruben G. Martinez
Virginia Commonwealth University

A meta-analysis investigating the correlation between treatment integrity and youth client outcomes

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at Virginia Commonwealth University.

by

Ruben Gabriel Martinez

Bachelor of Arts, Indiana University, 2011

Master of Science, Virginia Commonwealth University, 2017

Director, Bryce D. McLeod, Ph.D.

Professor, Psychology Department

Virginia Commonwealth University

Richmond, VA

May, 2020

Acknowledgements

I would first like to acknowledge my parents, Romeo and Lourdes Martinez, who have supported me fervently in this and every other pursuit. Thank you for everything you have done and continue to do for me. To my siblings: Melissa, Romeo, and Roman, and all of my tias y tios, sobrinas y sobrinos: thank you for your support. I hope I have made you all proud.

I would also like to acknowledge my two primary mentors. First, Dr. Bryce D. McLeod, my graduate school mentor and advisor, who has patiently overseen my growth as a writer and scientist. Thank you for guiding me on my journey to finding my voice, and thank you for always looking out for my best interests.

Next, I would like to thank Dr. Cara C. Lewis for changing the trajectory of my life. I was lost when I graduated college, but your mentorship, care, and belief helped me believe in myself in a way that I never knew was possible. Thank you so much for the opportunity you gave me in 2011, and thank you and Eric for your continued mentorship, love, and support.

To my dissertation committee, Drs. Chow, Rybarczyk, Southam-Gerow, and Sullivan: I cannot adequately thank you. Your questions and feedback shaped this project and your support saw it through to the end.

To all of my former educators and clinical supervisors, and the faculty in the Psychology Department at VCU and the Semel Institute at UCLA: Thank you for shaping me into a scientist, clinician, and thinker. Your courses, supervision sessions, and knowledge were invaluable to my growth.

Finally, I would like to acknowledge my incredibly supportive friends who believed in me from the outset. Thank you all for the laughs, listening ears, trips, belays, and love when the going got tough.

Table of Contents

Chapter 4

Chapter 5

List of Tables

List of Figures

Abstract

A meta-analysis investigating the correlation between treatment integrity and youth client outcomes

By Ruben G. Martinez, M.S.

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at Virginia Commonwealth University.

Virginia Commonwealth University, 2020

Major Director: Bryce D. McLeod, Ph.D.
Professor, Psychology Department

Objective: The relation between treatment integrity and youth client outcomes in psychosocial treatment research has implications for therapist training, study design, and study interpretation. Despite work done in this area, this relation remains unclear. The aim of the current study was to meta-analyze and describe characteristics of investigations of integrity and outcome in youth.

Method and Analytic Plan: A total of $N = 30$ studies were identified. Conceptual and methodological factors were coded. Pearson's r represented the magnitude of the integrity-outcome correlation. Robust variance estimation was used to account for dependency of within-study effect sizes. Two models were run with identical methods, Model 1 did not impute/estimate effect sizes and Model 2 imputed/estimated missing effect sizes. Each model was built iteratively, including an unconditional model and moderator model. Tests of heterogeneity, publication bias, and sensitivity analyses were conducted.

Results: The mean integrity-outcome effect size in Model 1 and Model 2 were negative and statistically significant ($r = -.11$, $p < .0001$; $r = -.09$, $p < .0001$, respectively). Treatment integrity component and quality of treatment integrity procedures did not moderate the integrity-outcome correlation. Publication bias revealed some trim-and-fill analyses, indicating the possibility of publication bias. Outliers did not play a role in study findings.

Discussion: There appears to be a small correlation between treatment integrity and outcome. A small sample and inconsistent reporting practices made additional moderation analyses inadvisable. Potential explanations for findings were explored. Recommendations

are provided, including consistent reporting standards and sophisticated research paradigms for therapy process research.

Chapter one

Introduction

Treatment integrity, or the extent to which a psychosocial treatment is delivered as intended, is hypothesized to contribute to improvements in youth client clinical outcomes (i.e., changes meant to be affected by the delivery of psychosocial treatment; Hoagwood et al., 1996; McLeod et al., 2013; Miller & Binder, 2002). The hypothesized correlation between treatment integrity and client outcomes spans clinical psychology, implementation science, and education (Fixsen et al., 2019; Fryling et al., 2012; McHugh et al., 2009; Perepletchikova & Kazdin, 2005; Society for the Implementation Research and Collaboration, 2018). This hypothesis has received significant attention because it has implications for research and applied work, including methods for conducting treatment process research, training and supervision of therapists in research and community settings, and the interpretation and dissemination of study findings (Miller & Binder, 2002, Perepletchikova & Kazdin, 2005).

Treatment integrity, also known as treatment fidelity, consists of: (1) therapist adherence, or the extent to which a therapist uses interventions as prescribed by a treatment protocol, (2) therapist competence, or the extent to which those interventions are delivered with quality and skill, and (3) treatment differentiation, or the extent to which therapists use interventions that are not prescribed by the treatment protocol (McLeod et al., 2013; Perepletchikova & Kazdin, 2005). Treatment integrity has historically been used to indicate whether an experimental condition was delivered as intended in a clinical trial (i.e., internal validity; Sanetti, & Kratochwill, 2009; Southam-Gerow & McLeod, 2013; Waltz et al., 1993). While internal validity was the original purpose of treatment integrity, research in the mid-1980s brought a new way of thinking; what if treatment integrity could be used to predict or correlate with client outcomes (Bellg et al., 2004; Perepletchikova & Kazdin, 2005)?

The ability to predict or correlate patient's improvement with a metric of the "what" and "how" of treatment delivery would be invaluable (McHugh et al., 2009; McLeod et al., 2009; Perepletchikova & Kazdin, 2005). Multiple sources across fields suggest that treatment integrity levels may be related to changes in youth outcomes (hereafter referred to as the integrity - outcome relation; Carroll et al., 2007; Doss, 2004; McLeod et al., 2013; Miller & Binder, 2002; Perepletchikova & Kazdin, 2005; Schulte et al., 2009). This hypothesized relation is intuitive; if improvements in client outcomes are the end goal of treatment, and we believe that the components of that treatment affect change, then the observations of treatment delivery and later outcome change should correspond in some way (Perepletchikova & Kazdin, 2005).

Integrity-outcome relation studies have increased over the past 40 or so years (see Luborsky et al., 1985; Marziali, 1984; Piper et al., 1886 for early examples). This includes empirical studies on multiple treatment modalities and problem areas (e.g., Barber et al., 1996; Hogue et al., 2008; Meier et al., 2015; Holder et al., 2018). Two dedicated meta-analyses have also been conducted (Collyer et al., 2019; Webb et al., 2010). Despite the work done in this area, the role that treatment integrity plays in client outcomes is still unclear.

The first meta-analysis on this topic was performed by Webb and colleagues (2010). They investigated the overall correlation between: (1) observational measures of adherence and symptom outcome and (2) observational measures of competence and symptom outcome. They identified 32 adherence-outcome effect sizes and 17 competence-outcome effect sizes across 36 studies, including one with youth participants. The sample included multiple treatment modalities (e.g., interpersonal, cognitive behavioral) and target problems (e.g., depression, drug use). The authors calculated an "r-type" (Webb et al., 2010, p. 6) effect size, which estimates the overall correlation between adherence or competence and outcome (Rosenthal et al., 2006). They

found that neither the adherence-outcome ($r = .02, p > .05$) nor competence-outcome ($r = .07, p > .05$) effects were significantly different from zero, indicating no correlation between adherence or competence and outcome.

The Webb and colleagues study raised a number of questions about the integrity-outcome relation. Its findings fundamentally challenge the hypothesis that delivering a treatment with integrity should lead to better outcomes. This was an excellent early effort to summarize early integrity-outcome work, but the findings should be considered preliminary for a number of reasons. First, the available sample was limited due to their search procedures and the novelty of the research question. Second, their analysis focused exclusively on observational measurements of treatment integrity, which are considered the gold standard for treatment integrity measurement (Hogue et al., 1996; Mowbray et al., 2003). However, it is not uncommon for clinical trials and practice settings to obtain reports from parents, youth, or supervisors (e.g., Jacob et al., 2014; Kushner et al., 2013; Schoenwald et al., 2009; Sikkema et al., 2013). Finally, they handled within-study effect size dependency (i.e., the notion that effect sizes share variance with other effect sizes within the same study; Borenstein et al., 2009) by averaging within-study effect sizes and selecting a single effect size, a common approach that has implications for the precision of meta-analytic findings (Borenstein et al., 2010; Tipton et al., 2019a). This study was an important first step in meta-analyzing the integrity-outcome relation, but their findings are reflective of a relatively early cross-section of the relevant research in adults.

Multiple studies have investigated the integrity-outcome relation in youth since the completion of Webb et al.'s (2010) meta-analysis (Goense et al., 2016; Rapley & Loades, 2018). Collyer et al. (2019) undertook the first meta-analysis of the integrity-outcome relation in youth. Their sample included $N = 52$ studies of non-lay delivered treatments for youth up to 21 years of

age experiencing emotional or behavioral disturbances. Their review identified 29 adherence-outcome effects, 9 competence-outcome effects, and 5 composite-outcome effects. Composite effects included any combination of adherence and competence components. They found a marginal, significant adherence-outcome effect ($r = .10$, $p < .001$), no competence-outcome effect ($r = .03$, $p = .26$), and no composite-outcome effect ($r = .06$, $p = .36$). While the sample, search methods, and analytic plan differed in some ways from Webb et al. (2010), the results were commensurate in that little relation was found between adherence, competence, and outcome.

Collyer et al.'s (2019) search and analytic methods were similar to Webb et al. (2010), but some important differences bear noting. Both studies separated treatment integrity into components, meaning that they separately assessed the effect of each treatment integrity component on outcomes. Notably, the Collyer et al. (2019) article included the composite treatment integrity component, which was not included by Webb et al. (2010). Both studies used a correlational effect size; Collyer et al. (2019) calculated a Pearson's r (Sedgwick, 2012), while Webb et al. (2010) calculated an "r-type" effect, so it is unclear if other types of correlations that are sometimes represented with the r notation were included (Tate, 1954). Adding to the differences in effect size calculation methods, each study used differing procedures to handle within-study effect size dependency; this minimizes or omits within-study variance components that are potentially valuable in explaining meta-analytic findings (Borenstein et al., 2009; Tipton et al., 2019a). Webb et al. (2010) mostly averaged within-studies effects, while Collyer et al. (2019) averaged effect sizes for the outcome directly targeted by the treatment or isolated a single effect size if the study contained multiple effect sizes that did not represent primary treatment outcomes.

Prior to the Collyer et al. (2019) meta-analysis, the field's hypothesis that integrity and outcome were related in youth was based upon the Webb et al. (2010) meta-analysis, conceptual work (e.g., Perpletchikova, 2011), and inferences from individual studies (like Hogue et al., 2008). The Collyer meta was a critical step forward because it provided an empirical base to the youth integrity-outcome relation. However, a comparison of methods between Collyer et al. (2019) and Webb et al. (2010) raises some additional questions. First, researchers from the earliest days of treatment integrity have assumed that adherence and competence are distinct. There is clearly a conceptual distinction between the two; however, there is some contention about whether adherence and competence are statistically separate entities (Muse & McManus, 2013; Rapley & Loades, 2018). This is partly due to the fact that adherence and competence investigations typically find a medium to strong correlation between the two (e.g., Bjaastad et al., 2016; Bloomquist et al., 2013; Hogue et al., 2008; McLeod et al., 2019; Guterman et al., 2015; Muse & McManus, 2013). Both past meta-analyses have relied on this conceptual distinction to guide their work, but this raises the question of whether treatment integrity as a whole, rather than each component, is related to youth client outcomes. Second, both studies attempted to handle effect size dependency by either averaging effect sizes within studies or isolating individual effect sizes, and both studies found similar overall effects. Thus, it is unclear whether this approach limited the precision of meta-analysis, subsequently over- or under-estimating the integrity-outcome correlation (Hedges et al., 2010). Third, neither study went into great descriptive depth when reporting methodological and conceptual characteristics of this sample. Rapley and Loades (2018) conducted a systematic review on the integrity-outcome relation in cognitive-behavioral therapy (CBT), but these characteristics are not well known across other treatment types.

When examined as a whole, our understanding of the integrity-outcome relation is preliminary (Collyer et al., 2019; McHugh et al., 2009; Perepletchikova, 2011; Webb et al., 2010). First, we do not understand how treatment integrity as a whole is related to outcome, as both previous meta-analyses separated conceptually by treatment integrity component. We also do not understand if, and to what extent, conceptual and methodological factors related to treatment integrity (i.e., conceptualization and operationalization) and outcome (i.e., conceptualization and measurement factors) moderate or otherwise impact this relation. With regard to the first question, if there is no integrity-outcome relation, then integrity as it is assessed and used for this purpose may have little, if any, utility in psychosocial treatment research (Miller & Binder, 2002). On the other hand, if there is a relation between treatment integrity and outcome, but methodological and conceptual factors are obfuscating this relation, then the field is using valuable time and resources to conduct post-hoc investigations that do not advance understanding of whether treatment integrity data can be leveraged to inform what we understand about the treatment process (Perepletchikova & Kazdin, 2005).

A number of sources suggest that conceptual and methodological factors may play a role in the integrity-outcome relation (Doss, 2004; McLeod et al., 2013; Miller & Binder, 2002; Perepletchikova, 2011; Perepletchikova & Kazdin, 2005). The integrity-outcome relation may be affected and/or obscured by factors across a number of levels, including: study characteristics, treatment characteristics, the quality of the correlational design, and factors associated with conceptualization and operationalization of the independent and dependent variables (Perpletchikova, 2007; Perepletchikova & Kazdin, 2005). At the study and treatment levels, data related to treatment type and format, target problem, and study design (e.g., accounting for correlational confounds) may impact the integrity-outcome relation. At the independent and

dependent variable levels, the methods used to characterize and measure treatment integrity and outcome, such as the informant, quality of the instrumentation, and method of measurement, may play a role in the obtained and analyzed data. Understanding the extent to which these are reported, and if and how these factors impact or moderate the integrity-outcome relation may help: (1) to guide future researchers in creating summative work like systematic reviews and meta-analyses, and (2) inform evidence-based decisions about study design, treatment integrity and outcome measurement, and the interpretation of integrity-outcome study results.

The current study aimed to: (1) describe the magnitude and direction of the integrity-outcome correlation, (2) maximize the number of effect sizes within studies while handling the within-study effect size dependency, and (3) describe and analyze the impact of important methodological and conceptual factors on the integrity-outcome relation in youth. These goals were addressed in a number of ways: first, a comprehensive review of the literature was used to identify a wide array of youth-focused studies, including varying treatment modalities, methods of treatment integrity measurement, outcome types, and populations. Second, the research question asked whether treatment integrity components, rather than any single component, correlated with any youth client outcome, as categorized by Hoagwood et al. (1996). Third, potential moderating or impactful variables identified in past integrity-outcome work were coded and descriptively analyzed to understand: (1) factors that are not consistently reported and (2) sources of variation across a number of levels that may help to clarify the integrity-outcome relation (e.g., study, treatment integrity measurement; Perepletchikova & Kazdin, 2005; Webb et al., 2010). Finally, robust variance estimation, a novel statistical method for this research question, was used to maximize the number of effect sizes in analyses and account for the dependency of within-study effect sizes (Hedges et al., 2010). It is hoped that these methods will

provide a rich view of what we know about the integrity-outcome relation in youth and add to

the growing literature that aims to meaningfully assess, measure, and apply treatment integrity

data.

Chapter two

Literature Review

A Hypothesized Relation Between Treatment Integrity and Outcomes

Treatment integrity, or the extent to which a psychosocial treatment (hereafter referred to as treatment) is delivered as intended, is believed to contribute to improvements in client clinical outcomes (e.g., reductions in symptoms or remission of diagnosis; Miller & Binder, 2002; McLeod et al., 2013). This hypothesis is widely held; it is not uncommon to read funding mechanisms, conference proceedings, or journal articles related to improving treatment integrity with the explicit goal of increasing the efficacy or effectiveness of a psychosocial treatment (hereafter referred to as treatment; Fryling et al., 2012; McHugh et al., 2009; Perepletchikova & Kazdin, 2005; Society for Implementation Research and Collaboration, 2018). The hypothesis that treatment integrity influences client outcomes is important; this assumption plays a role in the way that studies are conducted, therapists are trained, resources are allocated, and findings are interpreted and disseminated (Miller & Binder, 2002). Understanding the nature and direction of the integrity-outcome relation, as well as methodological and conceptual factors that may affect this relation can inform study design, treatment integrity measurement, therapist training, and interpretation of clinical findings.

Treatment integrity, also known as treatment fidelity or implementation fidelity, represents the extent to which interventions prescribed by a treatment protocol are delivered as intended and with quality (McLeod et al., 2013; Perepletchikova & Kazdin, 2005, Proctor et al., 2011). Though various conceptualizations exist, treatment integrity is most commonly conceptualized as having three components: adherence, competence, and differentiation (Margison et al., 2000; McLeod et al., 2013; Perepletchikova & Kazdin, 2005). Adherence

generally refers to the extent to which a therapist delivers components of the treatment as prescribed by the treatment protocol (Perepletchikova & Kazdin, 2005). For instance, if a therapist is delivering cognitive-behavioral therapy (CBT) for anxiety, adherence would refer to the extent to which the therapist delivered behavioral relaxation skills, thought restructuring, and other components explicitly prescribed by the treatment protocol (Kendall & Hedtke, 2006). Competence refers to the quality with which the therapist delivers the specific components of a treatment (Perepletchikova & Kazdin, 2005). In the CBT example, competence refers to the degree to which the therapist delivered the components of the treatment protocol with skill, warmth, responsiveness, and developmental appropriateness (e.g., the material is understandable and engaging). Finally, differentiation, which is sometimes considered an index of how purely a therapist delivers a treatment, refers to the extent to which a therapist delivers techniques that are explicitly proscribed or not prescribed by the treatment protocol. In CBT, this would include psychoanalytic techniques (McLeod et al., 2015).

Treatment integrity has historically been used as an indicator of whether, or the extent to which, an experimental treatment was delivered in an experimental condition as intended (i.e., internal validity; Kazdin, 2003); if an experimental treatment was delivered as intended and the treatment garnered positive effects, then experimenters can be more confident in interpreting those positive effects as a function of the experimental treatment as opposed to other factors (e.g., client or therapist characteristics; Kazdin, 2003). Research on treatment integrity began in the early 1980s, appearing in the school psychology and clinical trial literature (Waltz et al., 1993). Since that time, research on treatment integrity has evolved from being solely descriptive (i.e., "an end in its own right"; Perepletchikova & Kazdin, 2005, p. 366) to inferential (Bellg et al., 2004).

Various sources have suggested that there is promise in understanding the integrity-outcome relation (McHugh et al., 2009; McLeod et al., 2009; Perepletchikova & Kazdin, 2005). Multiple conceptual papers and models of treatment delivery, the treatment process, and implementation science have suggested that treatment integrity levels may be related to client outcomes (Carroll et al., 2007; Doss, 2004; McLeod et al., 2013; Miller & Binder, 2002; Perepletchikova & Kazdin, 2005; Schulte et al., 2009). Indeed, an oft-used argument to optimize therapist training is to increase treatment integrity, because an increase in treatment integrity will optimize client outcomes (e.g., Society for Implementation Research Collaboration, 2018). This hypothesized relation makes sense; the way a treatment is delivered should be related to client outcomes. If changing client outcomes is the end goal of a treatment and we believe that the components of that treatment are eliciting changes in client outcomes, then delivering the prescribed treatment components as intended and with quality should lead to changes in client outcomes (Perepletchikova & Kazdin, 2005).

Though this assumption makes logical sense, the empirical data have yet to provide a definitive understanding of integrity-outcome relation. This lack of understanding is problematic because it indicates one of two things may be true: (1) there is little or no relation between treatment integrity and outcome, which challenges the fundamental assumptions made about this relation, suggesting a misunderstanding of how treatments work to affect change or, (2) treatment integrity and client outcomes are related, but conceptualization and methodology of treatment integrity, outcome, and the integrity-outcome relation do not capture critical aspects of variables that explain this relation. It is possible that there is truth to both points; regardless of which hypothesis is accurate, the progress of the field is hindered by these inconsistent findings.

If there is no integrity-outcome relation, then integrity assessed and used for the purpose of predicting client outcomes may have little utility in treatment outcome research (Miller & Binder, 2002). On the other hand, if there is an integrity-outcome relation, but methodological or conceptual characteristics of these studies obfuscate the relation, then researchers are using valuable time and resources to conduct post-hoc investigations that do not advance understanding of treatment integrity as it applies to client outcome (Perepletchikova & Kazdin, 2005). Various sources suggest that conceptual and methodological factors across multiple study levels (e.g., study and treatment level) may play a role in the treatment integrity-outcome relation (Miller & Binder, 2002; Perepletchikova and Kazdin, 2005; Schulte et al., 2009) These variables exist at the study- and treatment- level (e.g., target problem, study design, quality of correlational design; Feeley et al., 1999; Stiles & Shapiro, 1989, 1994), the independent variable level (i.e., treatment integrity conceptualization and operationalization), and the dependent variable level (i.e., client outcome conceptualization and operationalization). Understanding the rate at which these factors are reported, the variability in these factors, and assessing their impact on the integrity-outcome relation may help to guide researchers by: (1) understanding if and to what extent these factors are reported across studies, (2) describing methodological and conceptual qualities in the sample (e.g., knowing if outcomes other than symptoms are being measured), and (3) by understanding if confounds, measurement quality, and conceptualization and operationalization of the independent and dependent variable affect the integrity-outcome relation.

An Inconsistent Relation

The study of the treatment integrity and outcome began in the mid-1980s (see Marziali, 1984; Sachs, 1983 for early examples). In the past 40 years, over 100 studies spanning multiple

treatment modalities in youth and adults have been conducted (Collyer et al., 2019; Goense et al., 2016; Holder et al., 2018; Meier et al., 2015; Rapley & Loades, 2019; Webb et al., 2010). In addition, two meta-analyses have been published to investigate this relation (Collyer et al., 2019; Webb et al., 2010). The first meta-analysis by Webb et al. (2010) investigated the extent to which adherence to or competence in delivering individual treatments were related to client clinical outcomes. Their sample consisted of 36 studies; the authors identified 32 adherence-outcome effect sizes and 17 competence-outcome effect sizes. The sample included multiple treatment modalities (e.g., interpersonal, cognitive behavioral), target problems (e.g., depression, drug use), and primarily adult-focused studies. One youth-focused study met inclusion criteria. The authors calculated an "r-type effect size" (Webb et al., 2010; p. 6), which estimates the percentage of change in client outcomes that can be attributed to adherence or competence (Rosenthal et al., 2006). Neither the adherence-outcome ($r = .02$, $p > .05$) nor competence-outcome ($r = .07$, $p > .05$) relations were significantly different from zero, indicating that there was, on average, no relation between adherence and outcome or competence and outcome.

Importantly, the authors found that there was significant heterogeneity in the adherence and competence samples, suggesting that there was significant variability around the mean over-and-above what is expected by chance (i.e., a statistically significant Q statistic; Huedo-Medina et al., 2006; Webb et al., 2010). The authors tested two conceptual moderators: treatment type and target problem, neither of which showed an effect on the integrity-outcome relation. They also tested two methodological moderators. First, they assessed whether the study authors established temporal precedence, which indicates whether or not the authors statistically accounted for change in symptoms that happened before treatment integrity measurement (Judd & Kenny, 1981; Stiles, 1988); they again found no effect. Second, they assessed the effect of

studies that did and did not control for the alliance, finding no effect. This study was critical in describing the early state of the field and suggested that, with some caveats, treatment integrity may not be related to client outcomes.

The Webb and colleagues study provided a good starting point for understanding the integrity-outcome relation, but its findings are not definitive for a number of reasons. First, the scope of the review was narrow due to the state of the field at the time of the review. Well over 30 studies investigating the integrity-outcome relation have been published since the completion of their review, and even more with youth participants (e.g., Boswell et al., 2013; Campos-Melady et al., 2017; Goldman & Gregory, 2009; Holder et al., 2018; Lopez & Basco, 2015; Rapley & Loades, 2019; Ryum et al., 2010; Westra et al., 2011). Second, their analysis focused exclusively on observer-rated treatment integrity measures. This focus is understandable, as observational measurement is considered the gold standard for treatment integrity measurement (Hogue et al., 1996; Mowbray et al., 2003); however, self-report and reports from others (e.g., parents, clients) are used in clinical trials and practice settings (e.g., Jacob et al., 2014; Kushner et al., 2013; Schoenwald et al., 2009; Sikkema et al., 2013). The Webb et al. (2010) study was critical for presenting an initial knowledge base for the methodology of this work, but their results represent a fairly narrow cross-section of the integrity-outcome research.

It has been almost 10 years since Webb and colleagues completed their search (April 15, 2009; Webb et al., 2010). Since the time of completion, the interest in the treatment integrity and outcome relation has continued. Multiple studies have demonstrated an integrity-outcome relation (e.g., Ginzburg et al., 2012; Goldman & Gregory, 2009; Haug et al., 2016; Owen & Hilsenroth, 2014; Webb et al., 2012; Weck et al., 2015), and more studies have been conducted with youth participants (Rapley & Loades, 2018).

Collyer et al. (2019) undertook the first meta-analysis on this topic, arguing that a separate meta-analysis was needed for assessing the integrity-outcome in youth, as there could be fundamental differences between adult and youth treatment. Their sample included non-lay delivered treatments for youth up to 21 years of age experiencing emotional or behavioral disturbances. Collyer et al.'s review (2019) identified 29 adherence-outcome effects, 9 competence-outcome effects, and 5 composite adherence/competence-outcome effects across N = 53 studies. Their analyses indicated a marginal but significant adherence-outcome effect (r = .10, p < .001), no competence-outcome effect (r = .03, p = .26), and no composite adherence/competence-outcome effect (r = .06, p = .36). While the sample, search methods, and analytic plan differed in some significant ways from the Webb et al. (2010) meta, the results were commensurate with Webb et al. (2010), in that little, if any relation was found between treatment integrity components and outcome.

The Webb et al. (2010) and Collyer et al. (2019) studies' similarities and differences highlight their unique contributions to understanding the integrity-outcome relation. Collyer's search and analytic methods roughly followed Webb et al. (2010). First, the age of patients overlapped slightly, as Webb et al. (2010) included ages 18 and up, and the Collyer study included youth up to 21 years of age. Second, each study separated treatment integrity into respective components: Webb assessed adherence and competence, while Collyer assessed adherence, competence, and a composite treatment integrity term for studies that combined any components. Composite treatment integrity was defined by Collyer et al. (2019) to be any combination of the conceptualization of adherence and competence. Third, Collyer and colleagues calculated a Pearson's r as an effect size metric, while Webb and colleagues calculated an "r-type" effect size (Webb et al., 2010, p. 6), so it is unclear whether various types

of correlation that use the *r* notation were included (e.g., partial correlations, Spearman rank-ordered correlations; Tate, 1954). Regardless, each study included only one effect size for each component from each study. Fourth, each conducted a random effects meta-analysis on treatment integrity component-specific models (e.g., adherence only). Both studies handled effect sizes similarly; they either averaged effect sizes if all outcomes were matched to the target problem or isolated a representative effect size if some effect sizes were not matched to the specific target problem. In sum, similar search procedures, conceptual approaches to statistical analysis, and effect size calculation methods were used, but the populations were mostly distinct. Regardless of the similarities and differences in method and analysis, neither study found a consistently significant adherence-outcome or competence-outcome correlation.

With regard to characterizing the sample, Webb et al. (2010) included some important descriptive information, such as treatment type, target problem, and an assessment of correlational confounds (e.g., temporal precedence, Feeley et al., 1999), but did not attempt to characterize the sample in other ways. Collyer et al. (2019) focused mostly on the methodological qualities of studies, including psychometric reporting for treatment integrity instruments, correlational confounds, sample size and power, and rates of drop-out. They also roughly characterized treatment type (CBT, non-CBT, parenting only, family therapy), the treatment group (e.g., emotional disorders, autism spectrum disorder, substance use), and the informant of treatment integrity. One study by Rapley and Loades (2018) characterized some of these factors in youth-focused CBT. Otherwise, no studies have described, in this sample, the quality of treatment integrity procedures, the operationalization and conceptualization of treatment integrity or client outcomes, the timing of treatment integrity measurement, and

24

various other factors that may be related to the integrity-outcome relation in a sample including multiple treatment types.

Prior to the Collyer et al. (2019) meta-analysis, the field's understanding of the integrity-outcome relation in youth was based upon inferences drawn from the Webb et al. (2010) meta-analysis, individual studies conducted since that time, and a hypothesis that adherence and competence are related to clinical outcomes. The Collyer et al. (2019) meta was a critical step forward in expanding the integrity-outcome knowledge base, but it raises some questions. First, since the earliest days of treatment integrity research there has been an assumption that adherence and competence are conceptually distinct; indeed, both past meta-analyses followed this convention. There is, however, some contention about whether adherence and competence are statistically separate entities (Muse & McManus, 2013; Rapley & Loades, 2018), which raises the question of whether treatment integrity as a whole, rather than each component, is related to youth outcomes.

The second question brought about by the Collyer et al. (2019) review is regarding the handling of within-study effect size dependency. Within-study effect sizes have some covariation, as they rely on the same participants in the same paradigms, and so forth (Borenstein et al., 2009). Because traditional forms of meta-analytic statistical techniques (e.g., random effects meta-analysis; Tipton et al., 2019a, 2019b) do not account for this inter-correlation, effect sizes within studies are often averaged together or a single effect size within a study is chosen, which can have effects on the precision of meta-analysis (Hedges et al., 2010). New statistical methods, known as robust variance estimation, allow for the inclusion of all possible effect sizes and ultimately the handling of within-study effect size dependency (Hedges et al., 2010). The third question is related to the consistency and depth of reporting practices for methodological or

conceptual moderators. Consistently reported study information is needed to compare studies in meta-analysis through the use of moderator analyses (Borenstein et al., 2009; Lipsey, 2003), and there is still a relative lack of information regarding if and how important factors related to methodology and measurement are reported (e.g., the definition of treatment integrity components and operationalization of each component, types of clinical outcomes assessed).

A Way Forward

At this point, the field has established a preliminary understanding of the nature and magnitude of the correlation between treatment integrity components and clinical outcomes in adult and youth. What is still unknown is (1) the correlation between treatment integrity as a whole and outcome, (2) methodological characteristics of this literature related to treatment integrity and outcome, and (3) if novel statistical methods can help to disentangle or clarify the magnitude of the integrity-outcome relation. The previous meta-analyses provide a methodological starting point for conducting meta-analyses on the integrity-outcome relation. Ideally, a new meta-analysis would take what was learned from past evidence and improve upon those methods with new and innovative methods. Thus, the focus of a new meta-analysis can build on past methodology by: (1) maximizing the obtained sample through rigorous search and inclusion procedures, (2) attempting to maximize statistical power of the sample, giving the opportunity to explore the role of previously-studied and new moderators or other impactful factors in a larger sample, (3) assessing the correlation between treatment integrity as a whole and outcomes using novel statistical methods, and (4) attempting to identify and describe reporting practices and descriptive data related to moderators and other impactful variables that have been demonstrated or are hypothesized to affect the integrity-outcome relation.

The Current Study

The current study seeks to build on past reviews by performing a meta-analysis on the effects of treatment integrity on youth client outcomes. This study differed from both Webb et al. (2010) and Collyer et al. (2010) in a number of ways.

Gathering the Sample

Gathering the widest extent of literature in meta-analyses is critical for producing generalizable results and maximizing statistical power for conducting meta-analysis (Cohn & Becker, 2003; Valentine et al., 2010). To achieve the goal of a broad sample, this study used a comprehensive search strategy, incorporating broad search strings and inclusion criteria in an attempt to capture as much literature as possible. As such, three major changes were made to the Webb et al. (2010) search strategies. First, Webb and colleagues only searched PsycINFO. The current study included a search for relevant dissertations or theses, fugitive or grey literature (Conn et al., 2003), or in databases outside of PsycINFO (e.g., PubMed) that likely include studies of the integrity-outcome relation. Additionally, the search strings for the current study were adapted from past work by combining terms used in past meta-analyses and removing search terms specific to populations (e.g., child/adolescent/youth). The inclusion criteria used in the current study differed from Webb, in that Webb and colleagues focused exclusively upon observational coding of treatment integrity, and this study included self- and other-reported treatment integrity. The measurement of treatment integrity through self- and other-report is not uncommon (e.g., O'Malley et al., 1988; Schoenwald, Chapman, et al., 2009; Weck et al., 2015), so the possible effect of this type of measurement on the integrity-outcome relation warrants inclusion in any new summative work.

The search strategy and inclusion criteria also differed from the Collyer et al. (2019). Specifically, the search procedures and inclusion criteria were broader than those in Collyer et

al.'s (2019) review. Collyer and colleagues searched PsychINFO, Embase, and Medline. The current study includes more sources, including PsychINFO, PubMed, Web of Science, the ProQuest Dissertation database, relevant systematic review and meta-analyses, a hand search of relevant journals, and a review of the grey literature (Conn et al., 2003). Collyer and colleagues inclusion criteria included: (1) sample mean age up to 21 years, (2) a quantitative assessment of the integrity-outcome relation, (3) no interventions delivered by teachers, peers, paraprofessionals, or parents, (4) no studies with universal intervention, (5) no health or education outcomes, (6) studies must have been published in a peer-reviewed outlet, and (7) English language studies. These criteria differ from the current study, in that the current included mean sample age up to 18, allowed interventions delivered by educators and other professionals, and studies could include health outcomes as secondary, but not primary outcomes.

Including Representative Moderators and Impactful Variables

Of note, a distinction is made here between moderating variables and impactful variables; moderating variables were included in moderator analyses, while impactful variables were used in sensitivity analyses. Moderating variables are often used in meta-analysis to better explain a relation between an independent and dependent variable (Lipsey, 2003). These variables are typically entered into models in order to account for their potential impact on an effect size of interest. An example of this would be measurements of treatment integrity quality; a researcher may be interested in knowing if the quality of treatment integrity procedures influences the integrity-outcome relation, and thus would enter this into a statistical model including the effect sizes and outcomes.

On the other hand, sensitivity analyses are used to ensure that factors inherent to a study or group do not impact study results (Borenstein, et al., 2009). These factors are handled

differently than moderators (Borenstein et al., 2009; Tanner-Smith & Tipton, 2013). Specifically, these factors are typically used to cross-section samples to ensure that characteristics inherent to a sample are not obscuring the models. Some examples of this would be to run models with and without outliers or estimated effect sizes, which can both play a role in the precision of meta-analyses (Borenstein et al., 2009).

There is reason to believe that studies in this area differ along a number of dimensions that could be important in interpreting and understanding the integrity-outcome relation (Perepletchikova & Kazdin, 2005). Thus, efforts were made to identify and measure moderators that have been shown to be impactful in past studies and meta-analyses. Unfortunately, studies that investigate the integrity-outcome relation have infrequently assessed moderators or impactful variables of this relation, making the empirical selection of these factors difficult. For this reason, moderator and impactful factors measured were included in the descriptive effort based upon: (1) review of empirical moderators identified in this relation (from the Webb et al. (2010) and Collyer et al. (2019) study), and (2) a review of the empirical and conceptual literature (e.g., Perepletchikova & Kazdin, 2005). These factors are described in more detail below.

Study and Treatment-level Factors

Study and treatment-level factors vary between and within studies (e.g., sample characteristics) and may help to explain the inconsistent findings in treatment integrity-outcome studies (Lipsey, 2003). The experimental methods used to conduct clinical trials and investigations into treatment integrity and outcome have not stayed stagnant in the past 10 years; this work is being done in increasingly diverse settings with increasingly diverse methods as the field's understanding of investigating clinical interventions has evolved (e.g., Schulte et al.,

29

2009; Southam-Gerow & McLeod, 2013). As a result, the relation between treatment integrity and outcome is not restricted to one type of setting, treatment, target problem, or population. A new meta-analysis should account for and describe the variability in study- and treatment-level factors in the integrity-outcome relation, as it is possible that this relation may differ along one or more of these dimensions.

Study Design. Multiple factors in study design may influence study findings (Kazdin, 2003). These include factors such as recruitment practices (clinically referred, recruited for study), the treatment settings (home, school), random assignment, and method of assignment (Kazdin, 2003). These differences can play a role in the way that treatment integrity and outcomes are measured, as well as the outcomes of an intervention (Perepletchikova & Kazdin, 2005). It is possible that differences in study design may influence the integrity-outcome relation, but no studies to date have measured or described these relevant study-level characteristics. For this reason, this meta-analysis focused upon describing studies based upon these qualities and attempting to assess their impact on the integrity-outcome relation.

Target Problem. The efficacy and effectiveness of treatments may differ based upon whether the measured outcome was directly targeted by a specific treatment. For instance, Weisz et al. (2017) found that treatments for anxiety in youth were overall more efficacious than treatments of conduct problems, Attention Deficit Hyperactivity Disorder (ADHD), depression, and studies that assessed multiple problems. Some studies with adults have similarly shown that client outcomes within the same treatment type (e.g., CBT, mindfulness-based therapies) differ based upon the target problem (Hofmann et al., 2012; Khoury et al., 2013). Webb et al. (2010) assessed target problem as a moderator of the adherence-outcome and competence-outcome relation. They found a non-significant trend that indicated treatments for depression may

evidence higher treatment adherence-outcome relations ($r = .12$, $p = .08$, a small effect according to Cohen, 1992) and a significant trend indicating that target problem moderates the competence-outcome relation in studies of depression ($r = .28$, $p < .001$, a small effect; Cohen, 1992). These findings, along with the evidence that outcomes can differential based upon target problem even within the same treatment type, indicate that target problem may be a valuable moderator of the integrity-outcome relation.

Treatment Type. Psychological research has focused a great deal on assessing the effects of different treatment types, defined here as conceptual approaches to treatment (e.g., behavioral, cognitive-behavioral; Silverman & Hinshaw, 2008; Weisz et al., 2017). There is some evidence to suggest that treatment types are not equally effective across target problems (e.g., CBT for anxiety; Hofmann et al., 2012; Weisz et al., 2017). Webb et al. (2010) assessed the effect of treatment type; they found no effect of treatment type on the relation between adherence and outcome or competence and outcome, though the competence-outcome relation was trending toward significance. Collyer et al., (2019) also included an analysis of treatment type, finding that the adherence-outcome correlation was significant across treatment types with the exception of youth non-CBT interventions ($r = .01$, $p > .05$). They categorized treatments as either CBT, parent and youth non-CBT, family therapy, or parenting. This categorization mixes both format (individual, group, multi-system) and type of treatment (behavioral, cognitive-behavioral). It therefore would be helpful to better understand the types and formats of treatments being delivered in these studies, as each could independently contribute to the integrity-outcome relation.

Independent Variable-level Factors

Treatment integrity is a broad concept that can be conceptualized, operationalized, and measured in many ways. These dimensions depend primarily on how the authors attempt to use treatment integrity to answer their research question. The field has seen a divergence in the ways that treatment integrity is measured, and it is clear that methods used to measure treatment integrity are highly variable (Cox et al., 2019; Goense et al., 2014; Perepletchikova et al., 2007). This is likely the case because there is no clear consensus on the optimal way to do any of these things; conceptual work guides measurement efforts but few empirical investigations examine the extent to which obtained data vary along treatment integrity measurement dimensions. As such, the impact of these dimensions is important to consider in the integrity-outcome relation.

Conceptualization and Operationalization of Treatment Integrity. Treatment integrity has been conceptualized and operationalized in different ways. Conceptualizations include adherence only, competence only, differentiation only, and combined ratings of multiple treatment integrity components (e.g., combined adherence and competence; e.g., Barber et al., 1996; Podell et al., 2013). As evidenced in the Webb et al. (2010) analysis as well as the empirical literature conducted since then, the differential effects of treatment integrity components on client outcomes remains unclear. Unfortunately, little information is available about the effect of differentiation and combined ratings. Ideally, a new analysis would investigate whether the conceptualization of treatment integrity has an impact on the integrity-outcome relation.

After treatment integrity is conceptualized, authors must decide how they will operationalize treatment integrity. Some studies use global scores, which approximate the average treatment integrity component across the course of treatment (e.g., Hogue et al., 2008; Liber et al., 2010). Other studies sample from a particular phase of treatment, which provides an

estimation of how treatment integrity early or late in treatment, or at a specific session in treatment, affects client outcomes (e.g., Meier et al., 2015; Sasso et al., 2016). Other studies randomly select sessions from treatment stratified by therapist, session number, or another criterion (e.g., Hoffart, et al., 2005; Lopez & Basco, 2015). The most complex and resource-intensive way to operationalize treatment integrity is through multiple measurements to assess change over time (e.g., Farmer et al., 2017; Sasso et al., 2016; Schoenwald, Chapman, et al., 2009). Some early work suggests that using session-by-session scores is the most precise estimate of treatment integrity and maximizes the variability in the sample (Moncher & Prinz, 1991; Peterson et al., 1982, Waltz et al., 1993), but no work has been done to understand how this may affect the integrity-outcome relation specifically. If studies using session-by-session data tend to show more robust effects, researchers may wish to begin to utilize session-by-session data, as opposed to global scores, for treatment integrity-outcome studies.

Method and Reporter of Treatment Integrity Measurement. The method and reporter of treatment integrity data have been hypothesized to play a role in the accuracy of treatment integrity ratings and scores (Herschell et al., 2019). The observational method of coding treatment integrity through audio- or video-recordings has been considered the gold standard, but it is not uncommon for studies to assess treatment integrity through the use of therapist- or client-report. There is contradictory data regarding the accuracy of self- and other-reported treatment integrity data (Dart et al., 2020; Hogue et al., 2015; Sanetti & Kratochwill, 2011; Wickstrom, 1995; Wickstrom et al., 1998). There is also some evidence to suggest that treatment integrity reporters may have significantly discrepant ratings (e.g., Herschell et al., 2020). Webb et al. (2010) only included studies observational treatment integrity data, so both the method and reporter of treatment integrity were restricted. Collyer et al. (2019) included reporter, but not

33

method of treatment integrity measurement. Because treatment integrity measurement outside of observational methods are conducted, and because the field has not assessed the impact of these methods on the integrity-outcome relation, the potential impact of the treatment integrity measurement method warrants inclusion in the current study. Indeed, if specific types of treatment integrity measurement are shown to be related to some outcomes, but not others, this raises questions about the integrity-outcome research done to date.

Quality of Treatment Integrity Procedures. The quality of treatment integrity measurement procedures may be critical in understanding the integrity-outcome relation. The quality of treatment integrity measurement refers to the steps taken to ensure systematic and accurate measurement of treatment integrity. Indeed, if the methods used to measure and characterize treatment integrity are poor, it is difficult to imagine a scenario where gathered data will meaningfully characterize treatment integrity. Perepletchikova et al. (2007) created a system for assessing the rigor of treatment integrity measurement and performed a review using these criteria. This review provided some initial evidence that these methods are variable across study. They found that treatment integrity measurement procedures were poor in quality and infrequently performed. Goense et al. (2014) applied the system created for Perepletchikova et al.'s (2007) study to youth with externalizing behavior problems; they found much the same. The Perepletchikova et al. (2007) study was updated by Cox et al., (2019), who concluded that while the quality of the implementation of treatment integrity procedures has improved overall, the methods and quality of such efforts are still highly variable.

The methods used to assess and evaluate treatment integrity are particularly important in understanding the quality of obtained treatment integrity data (Perepletchikova et al., 2007). The system created by Perpletchikova et al. (2007), called the Implementation of Treatment Integrity

Procedures (ITIPS), evaluates studies on three criteria: treatment integrity establishment, assessment, and evaluation/reporting (Perepletchikova 2006a, 2006b, 2006c). Treatment integrity establishment refers to the extent to which researchers provide specific rationale and methods for measuring treatment integrity. Treatment integrity assessment refers to the extent to which researchers measured treatment integrity with direct (i.e., observational) or indirect (i.e., self-report) methods and the extent to which researchers used tools with psychometric evidence to measure treatment integrity. Treatment integrity evaluation/reporting refers to the extent to which researchers gathered treatment integrity data across cases, therapists, and situations, controlled for reactivity (i.e., whether therapists knew or did not know when or on what factors they were being assessed), and report these data in a clear and meaningful way. Together, these data paint a fairly detailed picture of how much confidence can be placed in the collected and analyzed treatment integrity data.

Goense et al. (2016) conducted a meta-analysis that took a somewhat different approach to exploring the integrity-outcome relation in randomized controlled trials for antisocial behavior (Goense et al., 2016). They first assessed the effects of each treatment using Cohen's *d* (Cohen, 1988). After calculating Cohen's *d*, they assessed the moderating effect of "high" or "low" levels of establishing, assessing, and evaluating/reporting treatment integrity procedures on the effect of the treatment. It is important to note that these categories of high and low do not correspond to how adherent or competent a therapist was; rather, the categories describe the quality of the procedures used to establish, assess, and evaluate/report treatment integrity. The authors also did not provide clear criteria regarding how studies were categorized as "low" or "high" treatment integrity.

A total of 17 studies were included in Goense et al. (2016). Overall, they found that the treatments in their sample were efficacious for treating antisocial behavior (Cohen's $d = .30$, $p < .05$; a small-to-medium effect; Cohen, 1988). This indicated that, on average, the children in the experimental groups of these RCTs were more likely to experience positive outcomes than those in the control group. They identified a moderating effect of the quality of treatment integrity procedures on the effect size between treatment and comparison groups, such that "high" treatment integrity studies obtained effects of Cohen's $d = .63$ ($p < .001$; medium-to-large effect; Cohen, 1988). On the other hand, when treatments did not have a positive effect, the quality of treatment integrity procedures tended to be low, as indicated by Cohen's $d = .14$ (a non-significant effect). They also found that effect sizes differed significantly depending on treatment type and modestly on the basis of treatment duration. When including all of this information into their final model, the authors claim that they found an effect of "high" treatment integrity, such that "high" treatment integrity explained the differences between treatment and control group over and above intervention characteristics (e.g., treatment type or duration). The authors go so far as to say that these results "imply that delivering interventions with high treatment integrity to youth with antisocial behavior is vital" (Goense et al., 2016, p. 106).

This study is not asking the same question as Webb et al. (2010), Collyer et al. (2019) or the current study. The focus and question were fundamentally different; Goense and colleagues were not explicitly testing whether treatment integrity is correlated to outcomes; rather, they asked whether treatment integrity procedures happened to be higher in studies where treatment was effective for a treatment group. However, this study does connect treatment integrity and outcome in a novel way. Procedures to ensure the quality of treatment integrity measurement are not consistently conducted or reported (Cox et al., 2019; Goense et al., 2014, Perepletchikova et

al., 2007). In addition, there is preliminary evidence that the quality of treatment integrity

measurement is in some way related to client outcomes (Goense et al., 2016). These two points

indicate that the quality of treatment integrity procedures warrants inclusion in the current study.

Assessing the impact of rigor of treatment integrity measurement on this relation may help to

clarify the role of treatment integrity measurement procedures in the integrity-outcome relation.

Dependent Variable-level Factors

Client outcomes can be conceptualized, operationalized, and measured in different ways.

The evidence-based assessment movement (Achenbach, 2005; Hunsley & Mash, 2005, 2007)

emphasized the importance of using consistent and scientific methods to assess client outcomes,

referring to the state of the science without scientific measurement as "building a magnificent

house with no foundation" (Achenbach, 2005, p. 547). Similar to treatment integrity, there are

many ways to conceptualize, operationalize, and measure client outcomes. These approaches are

variable across studies, and different approaches are represented in the study of treatment

integrity and outcome (Rapley & Loades, 2018). Accounting for the varying methods of outcome

measurement may aid in understanding whether the integrity-outcome relation is affected by the

type of outcome that is measured.

Conceptualization and Operationalization of Outcomes. Client clinical outcomes can

be conceptualized and operationalized in a number of ways. Hoagwood et al. (1996) outline five

domains of client outcomes, including: symptoms and diagnoses, functioning, consumer

perspectives, environments, and systems. The majority of treatment integrity-outcome studies to

date have assessed client symptoms, but assessment of client functioning and changes in

environment are employed (e.g., Brown et al., 2013; Hilsenroth et al., 2003; Kuyken &

Tsivrikos, 2009; Podell et al., 2013; Schoenwald, Sheidow, et al., 2009; Serralta et al., 2010).

Neither the Webb et al. (2010) or Collyer et al. (2019) analyses characterized outcome type. An empirical investigation into whether integrity affects domains of outcomes differentially may spur along measurement of other meaningful outcomes in these studies.

Outcome Measurement Method and Reporter. The outcome reporter may play a role in the accuracy of outcome ratings. There is sufficient evidence to suggest that parent- and child-report of the same information can be discrepant (Alfano et al., 2015; De Los Reyes et al., 2015; De Los Reyes & Ohannessian, 2016; Lagattuta et al., 2012). This was not addressed in Webb et al.'s (2010) or Collyer et al.'s (2019) review. It is possible that certain outcome methods or reporters are related to integrity but others are not. Understanding the impact of outcome method and reporter has important implication in design choices for researchers conducting integrity-outcome studies**.**

Design Characteristics in Correlational Studies

Webb et al. (2010) assessed for the effects of the alliance and the establishment of temporal precedence, two design characteristics that impact the establishment of causality in a correlational design (Barber et al., 2007; Feeley et al., 1999; Judd & Kenny, 1981; McLeod et al., 2013). Collyer et al. (2019) also measured this, but did not include these design choices in any moderator analyses.

Alliance. Multiple studies have identified the alliance (defined as the affective and collaborative bond in the youth-therapist relationship; Elvins & Green, 2008; McLeod, 2011) as an important third variable that can affect the integrity-outcome relation (Barber et al., 2007; McLeod et al., 2011; Rapley & Loades, 2018). Not assessing and controlling for the alliance can increase the possibility of finding a relation between treatment integrity and outcome when there is no relation (Webb et al., 2010). A number of the reviewed studies assessed or controlled for

the alliance (Liber et al., 2010; Owen & Hilsenroth, 2014; Webb et al., 2012), while others did not (Becker et al., 2012, Boswell et al., 2013). More consistent measurement of alliance is a step forward in the field; despite this, assessments of the alliance are not universally done, nor are they often included in integrity-outcome investigations (Webb et al., 2010).

Webb and colleagues found no moderating role of the alliance in the adherence-outcome relation, but found that competence-outcome effects were significantly smaller in studies that controlled for the alliance. Their samples were small, with 11 of 15 adherence-outcome studies and nine of 15 competence-outcome studies controlling for the alliance. The Collyer et al. (2019) meta showed that only three studies included in their review assessed for the alliance. There is sufficient empirical evidence that suggests that the alliance may impact both treatment integrity and outcome (e.g., Andrews et al., 2016; Hogue et al., 2008; Laws et al., 2017; Martin et al., 2000; Weck et al., 2015). The prominent role of the alliance in past integrity-outcome literature and preliminary findings from Webb et al. (2010) warrant re-assessment of the moderating effect of the alliance on the integrity-outcome relation in a youth sample.

Temporal Precedence. Establishing temporal precedence is an important consideration when attempting to establish the integrity-outcome relation (Feeley et al., 1999; Judd & Kenny, 1981; Kazdin, 2007). The two requirements of establishing temporal precedence are as follows (Feeley et al., 1999; Judd & Kenny, 1981). First, treatment integrity should be measured temporally before the outcome variable that it is meant to predict. Second, studies should account for any outcome change that occurred prior to the measurement of treatment integrity. If the requirements for temporal precedence are not met, it is difficult to say with confidence that (a) the predicting variable is related to the dependent variable and (b) third variables do not explain the relation between intervention(s) X and client outcome Y (Judd & Kenny, 1981). Webb and

colleagues assessed for the effect of establishing temporal precedence, finding that it had no effect on the adherence-outcome and competence-outcome relation. It is important to note that their results may have been preliminary due to the size of their sample. Only eight of 15 adherence-outcome studies and four of 11 competence-outcome studies established temporal precedence. Only 43% of Collyer et al.'s (2019) sample controlled for baseline symptom severity or used an index of change, indicating variability in the establishment of temporal precedence. Establishing temporal precedence is clearly important to integrity-outcome literature (Feeley et al., 1999). Meta-analytic efforts should continue to identify the extent to which this is typically done in integrity-outcome studies in youth, as this may provide a much-needed reminder about the importance, and relative lack of, establishing temporal precedence.

Summary and Hypotheses

The nature and magnitude of the relation between treatment integrity and outcome in youth is still preliminary. A number of factors may impact or moderate the integrity-outcome relation, but the rate and consistency with which these are reported is still unclear. The current study aimed to: (1) describe the magnitude and direction of the correlation between treatment integrity as a whole and outcome in a unique sample of youth receiving treatment with novel statistical methods, and (2) attempt to identify and describe and analyze the impact of study-, treatment-, independent variable, and dependent variable factors that may impact the integrity-outcome relation in youth.

This was done in multiple ways: first, the sample represented a wide array of youth-focused studies, including varying treatment modalities, methods of treatment integrity measurement, outcome types, and populations. Second, the research question was conceptualized to understand the effect that treatment integrity, rather than adherence or competence, related to

any clinically relevant outcome, as categorized by Hoagwood et al. (1996). Third, potential moderating or impactful variables identified in past meta-analyses and reviews of the literature were coded and descriptively analyzed in order to further understand what is needed to clarify the integrity-outcome relation (e.g., Perepletchikova & Kazdin, 2005; Webb et al., 2010). Finally, robust variance estimation, a novel statistical method that handles the dependency of within-study effect sizes, was used to maximize the number of effect sizes by accounting for the dependency of within-study effect sizes (Hedges et al., 2010). It is hoped that these methods provide a summative lens through which to view the integrity-outcome relation, allowing for more meaningful measurement and application of treatment integrity and outcome data to inform treatment outcome and process research.

Hypotheses 1a-b

Hypothesis one states that the mean weighted integrity-outcome effect size (Pearson's r) would be statistically significant.

Hypotheses 2a-c, d-f

Hypothesis two states that the treatment integrity component type (adherence, competence, composite) would moderate the integrity-outcome relation.

Hypotheses 3a-b

Hypothesis three states that the quality of treatment integrity procedures would moderate the integrity-outcome relation.

Chapter three

Method

Data Sources, Search Strings, and Study Selection

Data Sources

Studies were extracted from five sources. First, PubMed, PsycINFO, and Web of
Sciences Social Sciences index were searched in order to capture the published, peer-reviewed
literature. Second, PsycEXTRA was searched in order to identify relevant grey literature (i.e.,
presentations, unpublished literature; Conn et al., 2003). Third, reference sections from relevant
meta analyses and systematic reviews were used for rolling reviews of the literature. Fourth, the
ProQuest Dissertation and Theses index was used to identify relevant dissertations. Finally, a
title and abstract review was conducted for articles published between January 2019 and March
2020 in Behavior Therapy, Journal of Consulting and Clinical Psychology (JCCP), and Journal
of Clinical Child & Adolescent Psychology (JCCAP).

Selecting Representative Search Terms

Search terms are ideally identified with reference to the independent variable, dependent
variable, and specification of population of interest (Lipsey & Wilson, 2001). Search strings
were adapted from Webb et al. (2010) in consultation with a librarian at Virginia Commonwealth
University. The final search terms consisted of the following and were identical across search
engines: therapist OR psychotherapist OR clinician OR practitioner AND fidelity OR "treatment
integrity" OR "integrity" OR adher* OR competen* OR differentiation AND outcome. No
restrictions were applied to publishing year. Because this dissertation was originally to include
adults, no restriction was applied to population terms; rather, adult-focused studies were
excluded later in the process.

Grey Literature

Upon review, PsycEXTRA (American Psychological Association, 2020) does not have a typical Boolean search feature, meaning that the combination of search terms in a string was not possible. In order to remedy this, the words "adherence," "competence," "therapist adherence," "therapist competence," "treatment integrity," and "fidelity," were used to search the following categories: "Behavior Therapy & Behavior Modification," "Cognitive Therapy," "General Psychology," "Health and Mental Health Services," "Health and Mental Health Treatment and Prevention," and "Psychotherapy and Psychotherapeutic Counseling."

ProQuest Dissertation Search

The ProQuest Thesis and Dissertation database was searched with the search terms provided above using the following limiting criteria: English Language, dissertations only, abstract search only, full text not required, and "Psychotherapy" as index.

Rolling Review

The rolling review included a hand search of reference sections in identified meta-analyses and systematic reviews (Collyer et al., 2019; Goense et al., 2014; Goense et al., 2016; Rapley & Loades, 2018; Webb et al., 2010).

Inclusion Criteria

This study used six inclusion criteria. First, the studies must have been English-language, consistent with Webb et al. (2010) and Collyer et al. (2019). Second, the studies must have broadly defined adherence as the extent to which the therapist delivered a given treatment as intended, competence as the quality of treatment delivery, or differentiation as the extent to which a therapist integrates components of other distinct treatments (defined as such in McLeod et al., 2013). Studies that combined treatment integrity components into one measurement were also collected. This was done to ensure that the collected studies were comparable on the

conceptual basis of the treatment integrity components that they assessed. These criteria are similar to those in Webb et al. (2010), with the primary differences being the inclusion of differentiation and inclusion of combined treatment integrity components. Collyer et al. (2019) collected combined treatment integrity components, but did not have specified definitions for adherence or competence. Third, studies included a quantitative assessment of adherence and/or competence or differentiation and client symptom outcome and statistically assessed the relation between the treatment integrity component during treatment and outcome following treatment. This was necessary in order to obtain effect sizes for meta-analysis and is consistent with the Webb et al. (2010) and Collyer et al. (2019) approaches. Fourth, studies included more than $n = 5$ youth participants (mean age 18 years or younger) selected and treated for psychopathology. This diverged from Webb et al. (2010), as they primarily assessed adult literature, and also from Collyer et al. (2019), as they included a mean age of up to 21 years old. This criterion also explicitly excludes single-case designs. Fifth, studies must have delivered in-person, as opposed to internet, telephone, or self-delivered treatment (e.g., bibliotherapy). This differed from Webb et al. (2010), as they focused primarily on individual, in-person treatment, excluding family or group therapies. The fifth criteria also differed from Collyer et al. (2019), as they specifically excluded interventions delivered by teachers, lay professionals, or caregivers. Sixth, studies were excluded if the child was not involved in either the assessment or treatment components (i.e., parenting-only interventions with no youth contact were excluded). This criterion was included because only one such study was identified during this search, and the study did not require that the child be present for any part of the assessment or treatment.

Abstrackr

Abstrackr (Wallace et al., 2012) is an online application that uses machine learning to organize and identify relevant abstracts from literature reviews. Abstrackr takes into account users' decisions for inclusion/exclusion and specified keywords to assign probabilities for inclusion/exclusion of unreviewed abstracts. Abstrackr was used to search and identify relevant abstracts in the initial title and abstract review phase for the database search (PubMed, Web of Science, PsycINFO).

Data Extraction, Coding, and Reliability

Steps of Effect Size Coding and Reliability

Step 1. Effect sizes (Pearson's *r*) were first searched for in an article. The effect size of interest was the bivariate correlation between a component of treatment integrity (adherence, competence, or differentiation) and outcome, so studies that conducted multiple comparisons contained multiple effect sizes. When available, Pearson's *r*, or a convertible effect size (e.g., Cohen's *d*, odds ratios; Borenstein et al., 2009) was gathered for all calculated correlations between the treatment integrity component and outcome. These data were readily available in the text or tables of *n* = 10 studies. Parent studies, supplemental tables, and online archives were gathered and investigated for each study without available data, though no studies included supplemental tables with bivariate correlations that could not be found in the text or tables of the article.

Step 2. For studies that did not report Pearson's *r*, corresponding authors were emailed with a request containing (1) the purpose of the project, (2) the name and details of the relevant study, (3) the statistics (bivariate correlations; Pearson's *r*) and names of variables/instruments that were being requested. The e-mail also offered for corresponding authors to provide de-

identified datasets, visuals (e.g., tables), or any other means by which to calculate Pearson's r. If the author's emails were not current, efforts were made to either: (1) identify their most recent e-mail via an internet search or (2) contact the authors via ResearchGate.net, a website that allows authors to post their published work. In sum, 16 authors were contacted about 18 studies, data were obtained for $n = 6$ studies, and data were unable to be obtained for $n = 13$ studies, leaving a total of $n = 13$ studies without a Pearson's r or convertible effect size (43% of sample), and $n = 17$ studies (57% of sample) with a Pearson's r.

Step 3. All provided effect size statistics aside from Pearson's r (Cohen's d, odds ratios) were converted into Pearson's r using an online calculator (Lenhard & Lenhard, 2016). A total of 13 effect sizes were converted using this method, 9 were odds ratios and 4 were Cohen's d.

Step 4. Ensuring that effect sizes across studies is comparable is critical to interpreting meta-analyses. In addition, the sign (+ or -) corresponding to Pearson's r is dependent on the scoring and conceptualization of included variables. For instance, a positive correlation of adherence and symptom outcomes may mean that as treatment integrity increased, symptoms did also, but it can also mean the opposite depending on how the data are conceptualized and scored. Thus, if taken directly from studies without changing the sign, it is possible that a positive correlation from one study and a negative correlation from another conceptually meant that the client improved. Given that so many types of outcomes were included, identifying the appropriate sign and adapting Pearson's r to a consistent sign was critical for interpretation of the data. Thus, outcomes were treated such that an improvement in the outcome measure reflected a negative correlation, such that higher treatment integrity indicated fewer overall problems. This approach was chosen because the majority of outcomes in this sample are symptom measures,

which are typically scored such that higher scores correspond to higher numbers or severity of symptoms (worse outcomes).

Step 5. The remaining analyses fell into two categories. The first category included analyses that provided a verbal description of significance or non-significance and were interpretable in the context of bivariate correlations. For instance, a simple linear regression model including only one treatment integrity component and one outcome was included in this category. The second category included analyses that were not interpretable in the context of a bivariate correlation, meaning that they presented results in a manner where the relation between one predictor and one outcome were not reported in a clear way, including complex structural equation models or hierarchical linear regression models where multiple treatment integrity components were reported, or only the interaction of treatment integrity components with other factors (alliance, for instance) were reported.

Step 6. For studies in the interpretable category, two methods were used to estimate effect sizes. If studies reported that the effect was not statistically significant, $r = 0$ was assumed. If studies reported that the effect was statistically significant, a critical value of Pearson's r (i.e., the minimum required correlation to be statistically significant given a sample size) was calculated. A function in RStudio was used to estimate the smallest possible statistically significant Pearson's r by identifying the critical value, or value used to identify the minimum sample size needed to obtain a significant effect in power analyses (Babcock, B., 2015; Howitt & Cramer, 2014; Page-Gould, E., 2015). For instance, a statistically significant result of an analysis with $n = 40$ (df = 38) participants corresponded to $r = .31$. For quality assurance, the results were cross-referenced with established critical values tables and were found to be consistent (Howitt

& Cramer, 2014). The code for these calculations is available upon request from the author.

Studies that fell into the uninterpretable category were excluded from analysis ($n = 5$; 18%).

Coding Moderators and Other Variables of Interest

Moderator extraction and coding were based upon the data provided in the study and followed the recommendations of Lipsey and Wilson (2001). A detailed codebook was created for all moderator and descriptive variables, which were either categorical or continuous in nature. The codebook is available upon request from the study author. This study coded all moderators discussed in previous sections, including those at the levels of the study, treatment, independent variable, dependent variable, and factors that influence causality in a correlational framework.

Study-level Codes. Each study was coded for the setting factors used in the Weisz et al. (2017) meta-analysis. Studies were coded for: (1) study region (inside/outside of North America), (2) setting (university, community clinic, hospital), (3) participant demographics (income, age, sex), (4) participant recruitment method (clinically referred, recruited for the study, or mandated treatment), (5) target problem and method of diagnostic assessment, and (6) participant assignment method (randomization and method of randomization). One study-level codesheet was completed for each study.

Strength of Correlational Design. Studies were also coded based upon whether the study established temporal precedence (yes/partially/not at all), and whether the study measured, or controlled for the alliance (yes/measured but did not use/did not measure at all).

Treatment-level Codes. Treatment-level moderators included a description of each intervention, the demographics of therapist participants, and the quality of treatment integrity

procedures for each treatment. One codesheet was completed for each treatment delivered in the study.

Treatment Type. It is possible to categorize treatment in many ways (See McLeod et al., 2015 and Weisz et al., 2017 for some examples). The following categories were chosen to represent distinct theory-based approaches to affecting change in treatment. These categories roughly follow those found in the Therapy Process Observational Coding System – Revised Strategies scale (McLeod et al., 2015), which includes five theory-based subscales: cognitive, behavioral, cognitive-behavioral, psychodynamic, and client-centered. The current approach to coding treatment type employs all five categories and includes a multi-system approach. This category was added to represent modalities with a clear multi-system, or ecological treatment type that employ methods from cognitive, behavioral, cognitive-behavioral, and client-centered approaches in a variety of formats (individual, group, school, and family; Henggeler, 1999). Thus, the following categories were used to categorize treatment type: (a) cognitive only, (b) behavioral only, (c) cognitive-behavioral, (d) psychodynamic, (e) client-centered, and (f) multi-system.

Treatment Format. The treatment format was assessed by coding whether the primary method of intervention was delivered at the level of individual, group, family, or multiple system levels. The multi-system or ecological code was used to classify treatments that are delivered across 3 or more levels of the ecology. This code was adopted because therapies such as Multisystemic Therapy (MST; Henggeler, 1999) deliver interventions across many levels of the youth's ecology, and thus do not fit neatly into any of the other above categories. Additionally, the extent to which contact with the target individual, parents, family, and school was coded.

Treatment dose was measured by coding total number of hours, sessions, weeks, or hours of treatment.

Therapist Demographics. Therapist demographics represented basic information related to the therapist, including: therapist age, level of educational attainment, trainee status, and years of experience.

Quality of Treatment Integrity Procedures. The Implementation of Treatment Integrity Procedures (ITIPS; Perepletchikova 2006a, 2006b, 2006c) was originally created in order to assess the quality of implementation of treatment integrity procedures in randomized controlled trials in an effort to provide a state of the science. Since that time, the ITIPS has been applied to reviews of various interventions with both adults and youth in systematic reviews and a meta-analysis (Bhar & Beck, 2009; Goense et al., 2014; Goense et al., 2016). The ITIPS consists of three quantitative subscales (establishing, assessing, and evaluating/reporting treatment integrity) and a summed scale score, which represents the overall quality of the implementation of treatment integrity procedures. Items on the ITIPS range from one to four, with higher item scores indicating higher quality of that specific aspect of treatment integrity. Full scale scores range from 22 (a score of one on all items, indicating a complete lack of treatment integrity measurement procedures) to 88 (a score of four on all items, indicating extensive implementation of procedures used to ensure and assess treatment integrity). The cutoffs provided by Perepletchikova (2006a) for this scale are as follows: "Inadequate" if 22 – 44, "Approaching Adequacy" if between 45 and 66, and "Adequate" if above 66. Little validity evidence has been presented for the ITIPS and it has demonstrated variable inter-rater reliability and internal consistency throughout studies (Bhar & Beck, 2009; Goense et al., 2014; Goense et al., 2016). In order to maintain consistency with the Goense and colleague's (2016) analysis, only the full-

scale ITIPS score was used, as it is an indicator of overall quality. ITIPS were gathered for each individual treatment within a study, so some studies had multiple ITIPS measurements.

Independent Variable-level Codes

Each study was coded for the conceptualization and operationalization of treatment integrity, the reporter of treatment integrity measurement, and the method of treatment integrity measurement. If multiple, separate treatment integrity components were studied, multiple codesheets were completed.

Conceptualization and Operationalization of Treatment Integrity. Each study was coded for (1) the way that treatment integrity was conceptualized (e.g., adherence only, competence only, differentiation only, or combined), (2) relevant scoring information (e.g., presence/absence, extensiveness), and (3) whether they measured treatment integrity: (a) only once, (b) at the beginning and end of treatment, (c) multiple times over treatment, or (d) on a session-by-session basis.

Method of Treatment Integrity Measurement. The method of treatment integrity, which refers to the method used to obtain the data that speak to treatment integrity levels, was assessed by coding: (1) the source of treatment integrity assessment (e.g., independent observer, therapist, client, parent), (2) the type of report used (e.g., survey/checklist, observational coding), and (3) whether the instrument was "homegrown" (e.g., created specifically for the study; Martinez et al., 2014) and had any established psychometric evidence.

Dependent Variable-level Moderators

Each study was coded with regard to the outcome domain outcome, reporter of outcome measurement, and method of outcome measurement. If multiple outcomes were studied, multiple

codesheets were created for each outcome. Each category is expanded upon in the following section.

Conceptualization and Operationalize of Outcomes. Outcome conceptualization categories included: symptoms and diagnoses, functioning, consumer perspectives, environments, and systems (Hoagwood et al., 1996). Studies were coded for whether they measured client outcomes: (1) only once, (2) at the beginning and end of treatment, (3) multiple times over treatment, or (4) on a session-by-session basis. Finally, outcome tools were coded for whether or not the outcome assessment represents the target of treatment (e.g., if treatment was focused on anxiety and outcome was meant to assess anxiety specifically versus a general assessment of psychopathology).

Outcome Measurement Method and Reporter. This study coded for: (1) outcome reporter (therapist, youth, parent, teacher), and (2) the type of report (questionnaire, clinical interview, or objective data counts).

Coder Training Procedures and Reliability

All data were coded by the author, a 30-year-old Latino man. Reliability samples were coded by one 27-year-old Caucasian woman. Training procedures for all coding documents proceeded in three steps. First, each coder read the codebook for the current study and the ITIPS manuals (Perepletchikova, 2006a; 2006b; 2006c) and jointly discussed any questions, problems, or inconsistencies in understanding for each item. Second, the coders met and co-rated a small sample of pilot articles (*n* = 5 of the total sample) on all variables, discussing discrepancies in coding and reaching consensus on all items. These particular articles were chosen because they resembled articles found in the current sample but did not meet inclusion criteria. Third, coders independently rated another small sample (10% of the total sample) articles independently.

These data were analyzed for discrepancies by the author. After training, a separate 20% ($n = 6$) of articles in the final sample were double-coded using the ITIPS and methodological coding system created for this study in order to assess reliability.

Different methods of reliability were used to assess categorical and continuous variables. All reliability information can be found in Tables 2 – 6. The gold standard for assessing inter-rater reliability in coding categorical variables is Cohen's Kappa (κ; Cohen, 1960; Landis & Koch, 1977). The κ statistic provides an overall agreement between two raters, ranging from κ = -1 to +1. Much like other correlation coefficients, positive coefficients indicate more agreement between coders, where negative coefficients indicate less agreement between coders. The most recent anchors developed for interpreting Cohen's κ are as follows: -.10 – .20 indicates "No Agreement," .21 – .39 indicates "Minimal" agreement, .40 – .59 indicates "Weak" agreement, .60 - .79 indicates "Moderate" agreement, .80 – .90 indicates "Strong" agreement, and above .90 indicates "Almost Perfect" agreement (McHugh, 2012). Cohen's κ for the methodological coding system ranged from .11 – 1.00 (M = .90, SD = .20); one item (whether the client received incentives for treatment) reflected "No Agreement," two items (scoring of adherence, treatment setting) reflected "Weak" agreement, one item reflected "Moderate" agreement, six items reflected "Strong" agreement, and 20 items reflected "Almost Perfect" agreement. One item (family contact) was not represented in the reliability sample, and thus κ was unable to be calculated.

Continuous variables, including the obtained effect sizes and moderator variables, were assessed using Intraclass Correlation Coefficients (ICCs; Shrout et al., 1987). The model ICC (2,2) was used for the current study because the same two coders coded the selected reliability sample article. The model dictates reliability coefficients and is based upon a two-way random

effects model. Cicchetti (1994) specified cutoffs for ICCs of "Poor" (below .40), "Fair" (.40 -

.59), "Good" (.60 - .74), and "Excellent" (.75 and above) agreement. ICCs for the ITIPS ranged

from .62 – 1.00 (M = .86, SD = .13), indicating that reliability on items were in the "Good" to

"Excellent" range. ICCs for the methodological coding system ranged from ranged from .13 – 1,

with one item being "Poor", two items being "Excellent," and one item having no variance.

Data Analytic Plan

Software

All study data were entered and managed using REDCap (Harris et al., 2009; Harris et

al., 2019) for organizational purposes and extracted as SPSS 26 (IBM Corp, 2019) datasets. The

descriptive analyses were conducted via SPSS 26. The effect size analyses were initially

conducted in R (R Core Team, 2020), and RStudio (RStudio Team, 2020) as the graphical user

interface. For quality assurance, each model was also run in Stata (StataCorp, 2020). The

following packages were used in R for all meta-analyses: *foreign* (R Core team, 2020)*, metafor*

(Viechtbauer, 2010)*, meta* (Balduzzi et al., 2019)*, robumeta* (Fisher et al., 2017)*, grid* (R Core

team, 2020)*, psych* (Revelle, 2019)*,* and *metaviz* (Kossmeier et al., 2020). In Stata, the *robumeta*

(Hedges et al., 2010) and *meta bias* (StataCorp, 2020) packages was used for analyses.

Data Synthesis

Conceptual Separation of Models

Separating Effect Sizes by Treatment Integrity Component. Adherence and

competence are typically thought of as distinct components of treatment integrity (McLeod et al.

2013, Perepletchikova, 2011). However, statistical comparisons of adherence and competence

frequently find a medium to large correlation ($r > .30$ per Cohen, 1992) across empirically

supported treatments and samples (Bjaastad et al., 2016; Bloomquist et al., 2013; Hogue et al.,

2008; McLeod et al., 2019; Gutermann et al., 2015; Muse & McManus, 2013). This speaks to the question of whether adherence and competence truly are distinct concepts, and if so, whether current measurement strategies successfully disentangle them (Muse & McManus, 2013; Rapley & Loades, 2018). Separating analyses by treatment integrity component has a number of conceptual and analytic advantages, in that being able to assess the unique contribution of each component to outcome is conceptually important and may make interpretation of analyses easier. However, this approach was taken by both Webb et al. (2010) and Collyer et al. (2019) and has been the traditional way of conceptualizing treatment integrity since the earliest days of treatment integrity research (Waltz et al., 1993). On the other hand, not separating analyses by treatment integrity component allows for more statistical power, as it increases the number of studies in one model and allows for the inclusion of composite treatment integrity effects, which otherwise would have been removed due to the low number of composite-outcome effect sizes identified in this study. This approach also asks a novel question: "to what extent is treatment integrity, regardless of component, correlated with youth clinical outcomes?" Given that the argument can be made that adherence and competence are consistently statistically related, and given that the current research question is related to the correlation between treatment integrity and outcome as a whole, initial models were run irrespective of treatment integrity component. Later, the treatment integrity component was entered into models to assess for any moderating effects of each conceptualization on the integrity-outcome correlation.

Estimating Missing Effect Sizes

Despite the use of effect size estimation approaches (e.g., considering nonsignificant effects $r = 0$) in past meta-analyses (Webb et al., 2010; Collyer et al., 2019; Weisz et al., 2017), there is some question as to the validity of such practices (Pigott & Polanin, 2020). Little

research has been done on the topic, but it is possible that estimating effect sizes may bias the study results either by underestimating the magnitude of some statistically significant effects or overestimating others (Pigott & Polanin, 2020). A common approach to handling missing effect sizes is to exclude studies with missing effect size information (Tipton et al., 2019b). However, that approach limits power by reducing the overall number of studies available for analyses (Hedges et al., 2010). A conservative approach consistent with past meta-analyses was adopted for the current study. It is possible that effect size estimation would have a notable effect on the statistical models. Thus, two sets of models were run with identical methods. The first set of models included only studies for which either: (1) Pearson's *r* was readily available or supplied by an author, or (2) a readily convertible effect size of a different metric (e.g., Cohen's *d*, odds ratios) was available. The second set of models included all studies in the first set and all studies with estimated effect sizes, as obtained with the methods described above in the "Steps of Effect Size Coding and Reliability" section.

Outliers

Outliers in meta-analyses can limit the precision of findings (Baker & Jackson, 2008). A commonly accepted method of assessing outliers is by creating a funnel plot to visually inspect the distribution of studies (Baker & Jackson, 2008). Funnel plots provide a visual representation of studies with small sample sizes and large effects. If outliers were identified through a visual check of the funnel plot (i.e., they fall outside the confidence interval lines of the funnel plot), sensitivity analyses were run with and without these studies in order to identify if, or the extent to which outliers affected the integrity-outcome correlation.

Synthesis of Effect Sizes

This study followed an iterative process for deciding which analytic method was most appropriate for the observed data. The overall correlation (Pearson's *r*) between treatment integrity and outcome was assessed through the use of correlated effects robust variance estimation (RVE; Hedges et al., 2010; Tanner-Smith & Tipton, 2013). This approach was used in order to account for the correlation of multiple effect size estimates within studies. Tanner-Smith and Tipton (2013) estimate that this approach can be used with as few as 10 studies but recommend more than 40, so RVE was determined to be an acceptable analytic approach for the current study sample.

Step 1: Determining Data Structure. The first step in conducting RVE with meta-analytic data is determining whether data are more appropriate for correlated effects or hierarchical effects analysis (Tanner-Smith & Tipton, 2013). Correlated effects data are typically used when the same outcome is present in multiple effect sizes, while hierarchical effects models are used when multiple independent studies are nested within one study or when multiple studies are nested within one research group (Tanner-Smith & Tipton, 2013). Because these effect sizes were all related to the same construct of clinical outcome, correlated effects models were used. Thus, the effect sizes within studies were considered level one, and the studies themselves at level two.

Step 2: Weighting Studies. In meta-analyses, studies with larger sample sizes are often given more weight than studies with small sample sizes; this is done because higher statistical power (i.e., a larger sample) likely leads to more accurate results (Kazdin, 2003). In order to gain a more accurate effect size estimate that accounts for sample size, studies are typically weighted by the inverse of their variance. While RVE does not explicitly require any specific weighting

procedure, Hedges et al. (2010) recommend weighting using an approximation of inverse

variance, calculated as follows in Equation 1:

$$w_{ij} = \frac{1}{\{(V_{.j} + \tau^2)[1 + (k_j - 1)\rho]\}} \tag{1}$$

In this equation w_{ij} refers to the weight of an individual study, $V_{.j} + \tau^2$ refers to the addition of

the average between and within-study variance, k_j refers to the number of effect sizes in the

given study, and ρ refers to the correlation of within-study effect sizes (Tanner-Smith & Tipton,

2013). This weighting scheme is automatically done in the *robumeta* packages in both R and

Stata when the user specifies the correlated effects weighting scheme (Tanner-Smith & Tipton,

2013). An online calculator (Wilson, n.d.) was used to transform all correlations to a

standardized Fisher's *z* in this step (Corey et al., 1998). All Fisher's *z* values were converted

back to Pearson's *r* to aid in interpretability of results. The standard error for each value of

Fisher's *z* was calculated using Equation 2 found in Cohen and Cohen (2003):

$$se_r = \sqrt{\frac{1 - r^2}{n - 2}} \tag{2}$$

Step 3: Heterogeneity Analysis. Assessments of heterogeneity in meta-analysis estimate

the amount of variability in the sample, which is sometimes viewed as a measure of how

appropriate the sample is for moderator analyses (Huedo-Medina et al., 2006; Tanner-Smith &

Tipton, 2013). Both models were assessed for heterogeneity in a uniform way. First, the *Q*

statistic was calculated, which provides an overall estimate of whether or not there is significant

heterogeneity in the sample. A statistically significant *Q* ($p < .05$) indicates significant

heterogeneity in the sample (Huedo-Medina et al., 2006). Neither R nor Stata provided a

corresponding significance level for *Q*. Because the *Q* statistic estimates a χ^2 distribution, the

statistical significance of *Q* was assessed by comparing the observed value of *Q* with k - 1

degrees of freedom χ^2 critical values, where k was the number of studies in the model (Hedges, 1994; Howitt & Cramer, 2014).

  Step 4: Model Building. An unconditional model was estimated to assess the magnitude of the correlation between treatment integrity and clinical outcome. A correlated effects weighting approach was used with a default, estimated $\rho = .80$ (Tanner-Smith & Tipton, 2013). The value of ρ is estimated because the covariance structure of within-study effect sizes is typically unknown, but assumed to be high (Hedges et al., 2010; Tanner-Smith & Tipton, 2013). A value of $\rho = .80$ means that the assumptive covariance of effect sizes within studies is about .80. In order to determine that the default/assumed ρ value did not impact the unconditional model results in a significant way, a sensitivity analysis was conducted. The method chosen in this study was to estimate the unconditional model with values of $\rho = 0$ to $.90$ in increments of .10, as advised by multiple sources (Hedges et al., 2010; Tanner-Smith & Tipton, 2013). When sensitivity analyses produce a stable coefficient, standard error, and τ^2 across differing levels of ρ, the assumptive within-study effect size correlation of $\rho = .80$ is acceptable (Tipton, 2015).

  The i^2 statistic is also commonly estimated and interpreted as a measure of heterogeneity, as it provides a ratio of how much proportion of heterogeneity in the sample is true between-study heterogeneity, as opposed to heterogeneity expected by chance (Huedo-Medina et al., 2006). Thus, i^2 was estimated and interpreted with the following anchors for true between-study heterogeneity: "low" is represented by $i^2 = 25$, "medium" is represented by $i^2 = 50$, and "high" is represented by $i^2 = 75\%$ (Higgins & Thompson, 2002). Finally, τ^2, an estimate of between-study effect size variability, was estimated and reported without assessment of statistical significance, as suggested by Tanner-Smith and Tipton (2013). Of note, $\omega 2$ is a between-studies, within-

cluster variance component in hierarchically-weighted models; because a hierarchical weighting scheme was not used, the value of ω2 was not reported (Hedges et al., 2010)

Step 5: Moderator Preparation and Analyses. According to Tanner-Smith and Tipton (2013), moderators in RVE can be assigned to level one (the effect size within the study) or level two (the study level). In order to do this, a between- and within-study mean was calculated for each moderator, the purpose of which is to identify if the variance in the moderator is primarily between or within studies. Between-study means are calculated by averaging the mean for each moderator across studies. Within-study means were calculated by subtracting the mean between-study moderator values from the between-study mean. If between-study moderator means were relatively consistent across studies but evidenced significant variability within-study, they were included at level one. If between-study means evidence variability then they were included at level two.

Moderators were then added into unconditional models. Due to the low number of studies and power, only factors that have previously evidenced a significant relation between integrity and outcome were included, including the treatment integrity component and ITIPS sum scale. The ITIPS sum scale was considered a between-study moderator due to the fact that the majority of variance was between, rather than within-studies. This is due to the nature of the ITIPS, in that it is only variable within a study if the study contains more than one treatment condition. Models were built by combining all moderators into one model for all studies. In order to prepare these data for analysis, the between-group ITIPS sum score was calculated, and the treatment integrity component (adherence, competence, or composite) was dummy coded.

Simulations of RVE with smaller samples indicate an increased risk of Type 1 error (i.e., an incorrect rejection of the null hypothesis; Tipton, 2015). For this reason, Tipton (2015)

created a procedure for small-study corrections in RVE that can be applied in both R and Stata. They suggest that small-study corrections even be applied to RVE models with larger samples. These corrections specify that if degrees of freedom for any moderator in the model are below *df* = 4, then the observed statistical significance may be double or higher than what is provided (Tipton, 2015). For instance, if *df* = 3 and statistical significance is *p* = .05, it is likely that the significance is closer to *p* = .1. Thus, these small-sample corrections were applied to both sets of models in order to take a conservative approach to analyses.

Step 6: Sensitivity Analyses. In order to ensure that characteristics of the sample did not account for any observed effects, sensitivity analyses were performed. First, the analyses were run with and without outliers identified by the funnel plot. Second, multiple studies included outcomes that were not primary outcomes of the treatment type (e.g., measurement of internalizing symptoms in a family therapy for behavior problems), so one model was run only with studies that were primary outcomes.

Step 7: Assessing Publication Bias. Studies with null results are less likely to be published (Polanin et al., 2016). Thus, it was necessary to ensure that publication bias did not influence the observed findings. This was conducted in four steps: first, effect sizes were averaged within studies, as no current publication bias tests are able to assess for publication bias within the framework of within-study effect size dependencies used in RVE (Peng et al., 2020). Second, a funnel plot of the studies in the sample was created and visually analyzed. Third, Egger's test of asymmetry was used to provide an estimate of bias in the sample (Egger et al., 1997). A significant (p < .1) result of Egger's test suggests potential publication bias (Egger et al., 1997). Finally, a trim-and-fill analysis was performed. Generally, a lack of trim and filled effect sizes is typically seen as an indicator that the effect was not influenced by publication bias

(Duval & Tweedie, 2000). On the other hand, trim-and-filled values are sometimes viewed as an indication of publication bias. Results of each assessment are reported below.

Literature Search

Summary of Records Obtained Through Search Procedures

The search of all three online databases returned a total of 9,088 records for review. The search of relevant meta-analyses and systematic reviews returned a total of 101 records pulled from four meta-analyses and systematic reviews (Collyer et al., 2019; Goense et al., 2016; Rapley and Loades, 2018, Webb et al., 2010). The review of the six categories of PsycEXTRA revealed a total of nine relevant abstracts. The hand-search of Behavior Therapy, JCCP, and JCCAP indices returned no relevant abstracts. A total of 1,123 dissertations were collected. In sum, a total of 1,233 studies were identified through sources other than academic search engines.

Summary of Search and Prisma Flowchart

In total, 10,321 studies were identified across all search procedures. After removal of irrelevant, adult, and duplicate studies and combination of relevant studies from all sources, a total of 8,681 articles were assessed for relevance through a title and abstract review, at which point 8,422 records were removed. A total of 259 records were assessed for eligibility through a full-text review, and 229 of these were excluded for not meeting inclusion criteria. Thus, $N = 30$ studies were included in the descriptive analysis of studies. Of those, $n = 25$ contained viable effect size data or garnered a response from the author that contained interpretable or convertible effect size data, leaving a total of $n = 25$ studies in the meta-analysis. Figure 1 includes the PRISMA flowchart detailing exclusion reasons through the search.

Characteristics of Identified Studies

Publishing dates ranged from 1999 – 2019 across 19 journals, including one dissertation. Studies were most frequently published in JCCAP ($n = 4$; 7%), JCCP ($n = 4$; 7%), and Behavior

Therapy (n = 3; 10%). All studies included a mean age of youth at or below 18 years of age. The most common target problems were oppositional/conduct/behavior problems (n = 14; 47%), anxiety problems (n = 4; 13%), or substance use problems (n = 8; 27%). Full diagnostic criteria were used and not used at almost equal rates (n = 14; 47% and n = 13; 43%, respectively). When diagnostic criteria were used, a reliable method of diagnosis was commonly implemented (n = 11; 37%).

Study-level Methodological Characteristics

Studies were conducted both within and outside of North America (n = 19; 63% and n = 10, 33.3% respectively). On average, study populations were comprised of 66.9% males with a mean age ranging from seven to 16 years of age (M = 14.06 years; SD = 2.43). Very little consistency was found in reporting of race and ethnicity, such that some studies reported on participants at the study level, others at the treatment level, and others not at all. The mean percentage of White participants was thus calculated, and was 49.48%. Participants were recruited in a number of ways, but were most commonly clinically-referred or treatment seeking (n = 13; 43%) or recruited specifically for the study (n = 8; 27%). Studies were mostly conducted in non-university outpatient hospital or clinics (n = 12; 40%) or multiple settings (e.g., home and school, school and clinic; n = 10; 33%). Most often, participants were randomly assigned (n = 20; 67%) with randomization approaches that stratify by covariates (Bugni et al., 2018; n = 9; 30%). A total of n = 11 (37%) studies did not randomize participants. See Table 2 for a detailed summary of study-level descriptors.

Treatment-level Characteristics

The most commonly represented treatment types were cognitive-behavioral (n = 14; 38%), behavioral therapy, multi-system (including multisystemic and behavioral family

treatments; $n = 22$; 60%) or client-centered ($n = 1$; 3%). These treatments were typically

delivered in individual ($n = 10$; 27%) and multi-system ($n = 20$; 54%) formats, meaning that

treatment was delivered at multiple levels of the system (e.g., school, family, individual). All

treatments included significant contact with the target individual, over half included significant

parent ($n = 23$; 77%) or family ($n = 19$; 63%) component. Therapist's mean age was 41.19 years

($SD = 5.08$), and no studies included trainees as the majority treatment providers. See Table 3 for

a complete list of treatment-level characteristics and list of included treatments.

Measurement of Treatment Integrity

The vast majority of represented treatment integrity measurement components were

focused on adherence ($n = 38$; 72%) or competence ($n = 12$; 23%). Notably, no treatment

differentiation-outcome studies were identified. When adherence was measured, it was typically

conceptualized equally as adherence to discrete interventions ($n = 18$; 34%) or adherence to the

overall principles or goals of a treatment ($n = 18$; 34%). Methods used to score adherence tools

were inconsistent, with the majority of scoring strategies being in the "other" category ($n = 18$;

34%). When competence was measured, it was most commonly conceptualized as technical, or

"domain-limited" competence, defined as skillfulness at delivering discreet interventions (Barber

et al., 2007; $n = 10$; 19%). The majority of treatment integrity collection was done through

observational coding ($n = 26$; 49%) with a minority of other-reported measurements ($n = 10$;

19%, including caregiver report). The most common source was an independent observer ($n = 25$; 47%) or a caregiver ($n = 10$; 19%). The majority of instruments were not "home-grown" ($n = 32$; 60%), meaning that more than half of instruments used had established psychometric

properties from past research. Finally, treatment integrity assessment was typically done either

multiple times in treatment ($n = 34$; 64%), or at each session ($n = 13$; 25%). See Table 4 for a complete summary of treatment integrity data.

Measurement of Outcomes

The vast majority of outcomes were conceptualized as symptoms/diagnosis ($n = 54$; 57%), functioning ($n = 25$; 27%) and changes in the environment ($n = 11$; 12%). The majority of outcomes were matched to the target problem ($n = 77$; 81%) and reported by the youth ($n = 34$; 36%) or caregiver ($n = 35$; 37%), primarily through questionnaires ($n = 79$; 83%). See Table 5 for a summary of outcome measurement and a list of instruments used in studies.

Confounds

The majority of studies did not include a measurement of alliance ($n = 21$; 72%), a small number measured the alliance but did not incorporate these measurements into analyses ($n = 2$; 7%), and only $n = 6$ (21%) incorporated alliance measurements into analyses. Additionally, temporal precedence was typically only partially established, ($n = 21$; 72%) usually in that the treatment integrity measurement occurred temporally prior to the outcome measurement of interest. In rare instances ($n = 7$; 24%), temporal precedence was completely established such that outcome change up to the point of the first treatment integrity measurement was accounted for. See Table 6 for a complete summary of alliance incorporation, establishment of temporal precedence, and list of confounding variables used across studies.

Implementation of Treatment Integrity Procedures

Full-scale ITIPS scores from the $N = 30$ studies generated 38 total ITIPS scores, ranging from 48 – 80 ($M = 61.85$, $SD = 9.67$), indicating that treatment integrity procedures for all treatments within the sample were implemented in the *Approaching Adequacy* ($n = 26$; 68%) to *Adequate* ($n = 12$; 32%) range.

Effect Size Description

Model 1 – Effect Sizes with No Estimation

The anchors for effect size interpretation are consistent with Cohen (1992), such that a "small" effect is $r = .10 - .29$, a "medium" effect is $.30 - .49$, and a "large" effect is .50 or greater. A negative correlation reflects improved outcomes, such that a higher treatment integrity score corresponds to a lower overall outcome score. A positive correlation reflects worse outcomes, such that higher levels of treatment integrity corresponded with a higher overall outcome score.

Across articles for which effect sizes could be obtained without using effect size estimation ($n = 17$), there were a total of 127 effect sizes. Of these, 94 were adherence-outcome, 30 competence-outcome, and 3 composite integrity-outcome effect sizes. Adherence-outcome effects ranged from $r = -.52 - .20$ ($M = -.09$, $SD = .13$), competence-outcome effects ranged from $r = -.30 - .19$ ($M = -.05$, $SD = .11$), and composite integrity-outcome effects ranged from $r = -.16 - .10$ ($M = -.12$, $SD = .03$). In sum, 127 effect sizes were gathered for Model 1, ranging from $r = -.52 - .20$ ($M = -.08$, $SD = .13$). The mean number of effect sizes obtained from one study was $M = 6.96$ and ranged from $1 - 24$ effect sizes per article.

Model 2 – Effect Sizes with Estimation

A total of 47 additional effect sizes were estimated. Of these, 34 effect sizes were estimated with $r = 0$, and 13 were estimated with Pearson's r ranging from $-.37 - .08$ ($M = -.07$, $SD = .17$). Across all studies ($n = 25$), a total of 128 adherence-outcome, 39 competence-outcome, and 7 composite integrity-outcome effect sizes were collected. All Model 2 adherence-outcome effects ranged from $r = -.52 - .20$ ($M = -.07$, $SD = .13$), competence-outcome effects ranged from $r = -.30 - .19$ ($M = -.04$, $SD = .10$), and composite integrity-outcome effects ranged

from $r = -.17 - 0$ ($M = -.08$, $SD = .07$). In sum, 174 effect sizes were gathered, ranging from $r = -.52 - .20$ ($M = -.06$, $SD = .12$). The mean number of effect sizes obtained from one study was $M = 6.96$ and ranged from $1 - 24$ effect sizes.

Data Synthesis

Synthesis of Effect Sizes

Model 1. Studies that reported Pearson's r or other effect size metric (e.g., Cohen's d or odds ratio).

Outliers. The funnel plot can be seen in Figure 2. No studies were identified as outliers.

Figure 2.
Funnel Plot – Outlier Assessment of Model 1
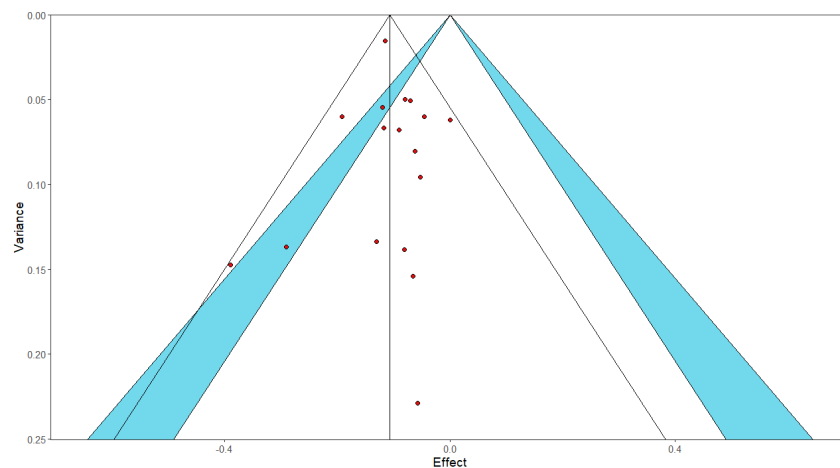
Heterogeneity. There was not significant heterogeneity in the sample, (Q (16) = 2.67, $p >$ .05, critical value = 26.30; Howitt & Cramer, 2014). In addition, the observed between-study variability could have been explained by chance alone, as indicated by $i^2 = 0$. Finally, there was virtually no between-study variance to be explained ($\tau^2 = -.06$).

Unconditional Model. The overall weighted correlation ($r = -.11$) indicated that the correlation between treatment integrity and outcome was negative and not statistically

significant, (df = 8.9, p < .0001, 95% CI = -.14 − -.07). This indicates that there was a "small,"
statistically significant correlation between treatment integrity and outcome, such that higher
treatment integrity correlated to better overall outcomes. The degrees of freedom (df > 4)
indicated there was adequate power to perform this analysis (Tipton, 2015). Sensitivity analyses
yielded stable coefficients, standard errors, and τ^2, indicating that assuming highly correlated (ρ
= .80) effect sizes within studies was an acceptable approach.

Moderator Model. After moderators were entered, the overall weighted correlation was
negative and not statistically significant (p > .05). Whether the effect size was focused on
adherence (p > .05), competence (p > .05), or a composite (p > .05) score did not moderate the
correlation between treatment integrity and client clinical outcomes. The quality with which the
treatment integrity procedures were implemented in the overall trial also did not moderate the
correlation between treatment integrity and client clinical outcomes (p > .05). It should be noted
that, despite using a small-sample correction, the degrees of freedom for the competence term
were fewer than four, indicating that the significance value (p) may be double what is provided
above (Tipton, 2015). Sensitivity analyses for this model yielded stable coefficients, standard
errors, and τ^2.

Sensitivity Analyses. Analyses were also run only with effects from outcomes that were
matched to problem area. The mean weighted effect size (r = -.12) was stable and statistically
significant (df = 8.95, p = < .05, 95% CI = -.15 − -.08). The moderator models were consistent
with the main models, in that neither treatment integrity components nor treatment integrity
quality moderated the integrity-outcome relation. All models were stable to varying levels of
estimated within-study effect size correlations.

Publication Bias. Upon visual inspection, the funnel plot was mostly symmetrical around the mean, with roughly half of the effects on each side of the mean line. Egger's test of asymmetry was not significant (bias estimate $\beta = .05$, $p > .10$), which suggests a lack of publication bias. Next, a trim-and-full funnel plot was created and is presented as Figure 3. The trim-and-fill analysis yielded two trim-and-filled effects, filled in black in the figure below, indicating the possibility of publication bias.

Figure 3.
Funnel Plot – Trim-and-Fill Analysis for Model 1
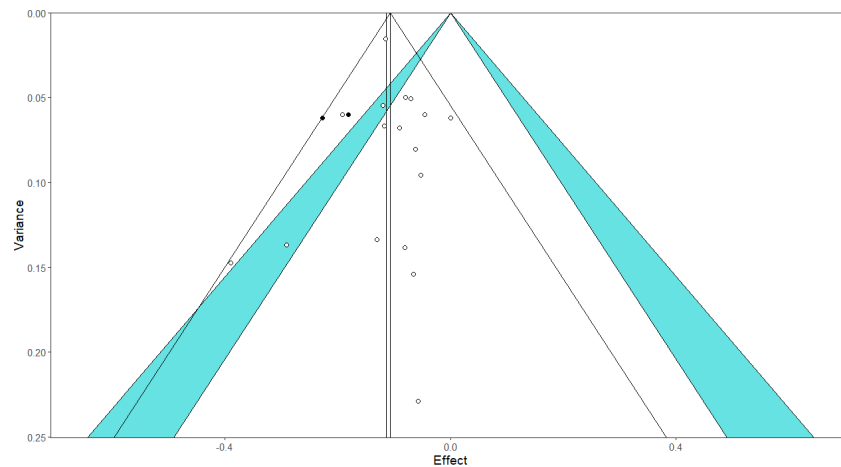
Model 2. All studies with all available effect sizes.

Outliers. The funnel plot can be seen in Figure 4. Data from a total of 10 effect sizes from $n = 4$ studies were noted to be completely outside the confidence intervals. These effects were included in model building but excluded for sensitivity analyses.

Figure 4.
Funnel Plot – Outlier Assessment of Model 2
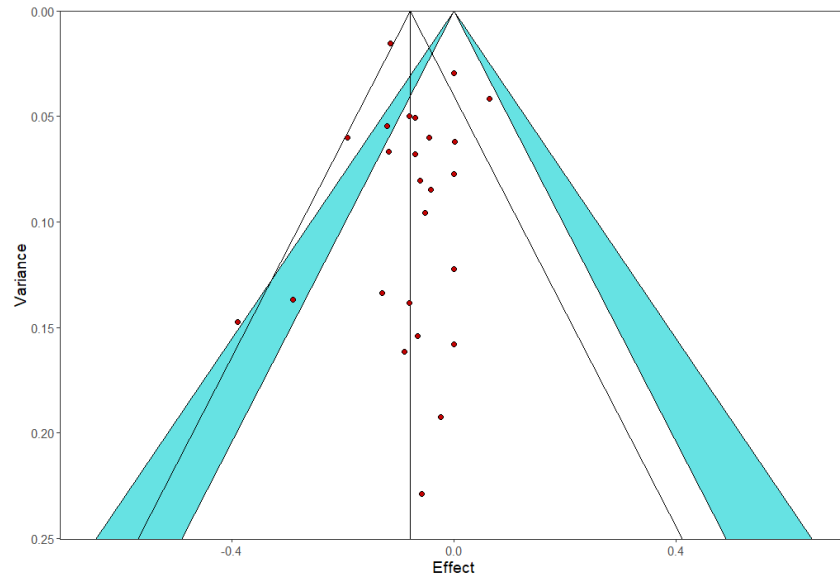
Heterogeneity. There was not significant heterogeneity in the sample, (Q (24) = 3.38, $p >$ .05, critical value = 36.42; Howitt & Cramer, 2014). In addition, the observed between-study variability could have been explained by chance alone, as indicated by $i^2 = 0$. Finally, there was virtually no true between-study variance ($\tau^2 = $ -.06).

Unconditional Model. The overall weighted correlation ($r = $ -.09) was negative and statistically significant ($df = 13.91$, $p = < .0001$, 95% CI = -.12 − -.06), such that the correlation between treatment integrity and outcome indicated that higher treatment integrity was correlated with bettering of outcomes. It is important to note that, while statistically significant, this mean weighted effect size does not meet the threshold of a "small" effect, per Cohen (1992). The degrees of freedom ($df > 4$) indicated there was adequate power to perform this analysis (Tipton, 2015). Sensitivity analyses yielded stable coefficients, standard errors, and τ^2.

Moderator Model. After moderators were entered, the mean weighted correlation was negative and not statistically significant. Whether the effect size was focused on adherence ($p >$

.05), competence ($p > .05$), or a composite ($p > .05$) score did not moderate the correlation between treatment integrity and client clinical outcomes. The quality with which the treatment integrity procedures were implemented also did not moderate the correlation between treatment integrity and client clinical outcomes ($p > .05$). In this model, the degrees of freedom for the ITIPS term was greater than four, indicating that the significance values for the adherence, competence, and composite terms may be double what was reported above (Tipton, 2015). Sensitivity analyses for this model yielded stable coefficients, standard errors, and τ^2.

Sensitivity Analyses. Sensitivity analyses were run without outliers. The unconditional model estimated a statistically significant, slightly higher mean weighted effect of $r = -.09$, such that higher treatment integrity corresponded with better outcomes ($df = 11.65$, $p = < .05$, 95% CI $= -.12 - -.06$). The results from the unconditional models exhibited no differences from the set of models that contained outliers; each model was robust to varying levels of estimated within-study effect size correlations and no moderators significantly impacted the integrity-outcome relation. Analyses were also run only with effects from outcomes that were matched to problem area. The mean weighted effect size ($r = -.09$) was stable and statistically significant ($df = 13.48$, $p = < .05$, 95% CI $= -.12 - -.06$). The moderator models were consistent with the main models, in that treatment integrity components and treatment integrity quality did not moderate the integrity-outcome relation. All models were stable to varying levels of estimated within-study effect size correlations.

Publication Bias. Upon visual inspection, the funnel plot was skewed to the right of the mean; given that the effect sizes were skewed such that more lower magnitude effect sizes were represented, it is unlikely that this asymmetry reflects the results of publication bias. Egger's test of asymmetry was significant (bias estimate $\beta = -.83$, $p = .07$), which suggests potential

publication bias. Next, a trim-and-full funnel plot was created and is presented as Figure 5. The

trim-and-fill analysis yielded five trim-and-filled effects, filled in black in the figure below,
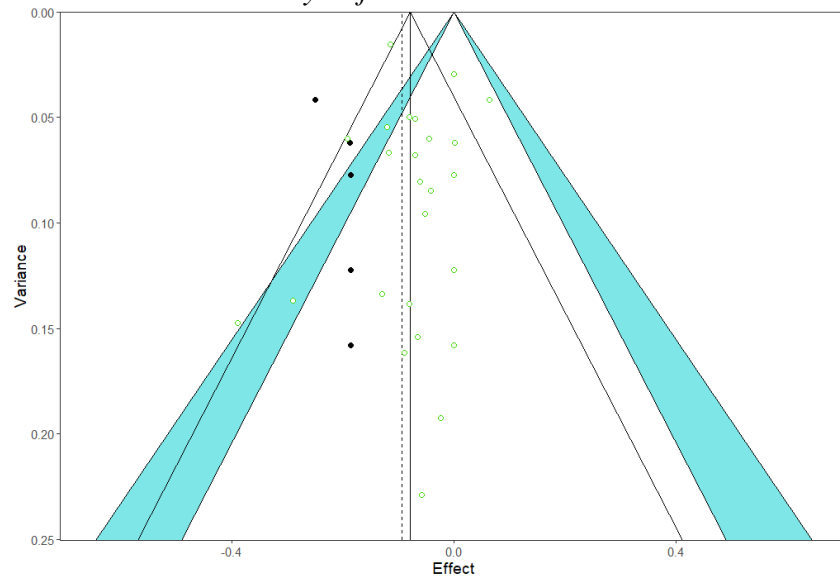
indicating the possibility of publication bias.

Figure 5.
Funnel Plot – Trim-and-Fill Analysis for Model 2

Discussion

Summary of Meta-analysis

This meta-analysis investigated the correlation between treatment integrity and youth client clinical outcomes. A total of $N = 30$ studies were included in qualitative analyses, and $n = 25$ of these produced 174 effect sizes that were included in the quantitative analyses. Two models were used for effect size synthesis. In Model 1, only *bona fide* effect sizes were included. Hypothesis 1a was upheld: Model 1 demonstrated a "small" and statistically significant correlation between treatment integrity and outcome, such that higher treatment integrity was associated with improvements in outcomes. Hypothesis 1b was also upheld: Model 2 was more inclusive, with estimated nonsignificant and significant effect sizes, and demonstrated a statistically significant integrity-outcome correlation, but the mean weighted correlation in this model did not meet the threshold for a "small" effect (Cohen, 1992). The entirety of hypotheses 2 and 3 were not upheld. Moderator analyses from both models indicated that the type of treatment integrity component did not moderate the integrity-outcome correlation. There was also no moderating effect of the quality of treatment integrity procedures on the integrity-outcome correlation. Outliers did not play a role in findings, but there was some evidence of publication bias in Model 2. Overall, these models are consistent with past meta-analyses. If there is a correlation between integrity and outcomes, it does not appear to be robust (Collyer et al., 2019; Webb et al., 2010).

The unconditional models were similar in a number of ways. First, there was little to no true between-study variance in effect sizes, suggesting that any observed variation in effects across studies was spurious (Borenstein et al., 2010). Second, both models evidenced a significant negative correlation between treatment integrity and outcome, such that higher treatment integrity corresponded to better clinical outcomes. Fourth, all models were robust to

varying levels of within-study effect sizes correlations. Finally, the models that incorporated moderators did not have adequate power to reliably detect any moderating effects of treatment integrity components or the quality of treatment integrity procedures.

However, the models also had some significant differences. Model 2 showed evidence of publication bias while Model 1 did not. This could be reflective of the addition of estimated effect sizes or indicative of problems with the search procedure (Borenstein et al., 2009). The magnitude of the statistically significant unconditional models was different such that Model 1 met the threshold for a "small" effect, while Model 2 did not. This difference was likely brought about by the addition of the non-significant effect sizes. It is especially striking in the light that this meta-analysis included an estimation of non-zero, statistically significant effect sizes, which were ignored in past meta-analyses. Without these, it is likely the case that the observed magnitude of the integrity-outcome relation would have decreased even more from Model 1 to Model 2. If this is true, then the practice of estimating nonsignificant effect sizes in past meta-analysis likely played a role in the small or null results found in Collyer et al. (2019) and Webb et al. (2010).

Summary of Descriptive Analysis

A major goal of this dissertation was to study the moderating effects of various methodological and conceptual moderators on the integrity-outcome correlation. Unfortunately, the small sample gathered, and even smaller sample of studies with usable effect size information, made additional moderator analyses inadvisable, as the set of models described above did not have sufficient power for reliably detecting any statistical associations (Tipton, 2015). In addition, the inconsistent reporting of moderators and a lack of statistical information for effect size calculation made efforts to effectively meta-analyze this sample difficult. Thus,

efforts were made to provide descriptive data in order to: (1) characterize the sample and (2) identify gaps for future research to address.

The lack of reported information was most striking when calculating effect sizes. The statistical methods and reported statistics were highly variable, including various types of correlations (Pearson's *r*, partial correlations, point-biserial correlations), structural equation modeling, linear and logistic regressions, and nonparametric tests (e.g., McNemar's test; Eiraldi et al., 2008). Only a few studies had readily available correlation tables despite the correlational nature of integrity-outcome investigations, which is common (Tipton et al., 2019b). Fewer effect sizes in studies means smaller samples, less statistical power, and ultimately less precision (Borenstein et al., 2010). The lack of consistently reported information is important for at least two reasons. First, moderator analyses cannot be done if there are not enough studies that report data on that moderator. Second, without reporting of the effect size of interest, meta-analysts must rely on missing effects analyses (e.g., single and multiple imputation, listwise deletion) that can limit the precision of meta-analytic findings (Pigott, 2001; Tipton et al., 2019a, 2019b). If these factors cannot be determined, then statistically solid meta-analytic efforts will continue to fall short of making stronger conclusions about the integrity-outcome relation.

Significant variability was observed across all levels of methodological coding, including at the study-, treatment-, treatment integrity-, and outcome-level. This variability existed both in study characteristics and the reporting of these characteristics. Missing data for methodological codes were common. In particular, client and therapist demographics were often missing (defined as > 50% of studies), including client/family income, information on socioeconomic status, descriptors of race/ethnicity, and therapist years of experience. Some treatment-level information was also inconsistently defined or missing, including time in treatment, average

number of sessions, or average weeks in treatment. Without these data in future meta-analyses, moderator analyses will continue to be based off samples with lackluster power or necessarily imputed missing information (Tipton, 2019a). More consistent reporting of these factors may help to clarify some significant questions, like whether the integrity-outcome relation matters more for some types of treatment, characteristics of therapists or clients, or other treatment factors (e.g., time in treatment; Borenstein et al., 2010; Lipsey, 2003).

What Do These Findings Mean in Context?

The findings from this meta-analysis all fit within the greater scope of past meta-analyses. The three meta-analyses have shown marginal or nonexistent correlations between treatment integrity and outcome. All three meta-analyses conceptually and methodologically diverge from the meta-analytic work done by Goense et al. (2016). Goense et al. (2016) estimated the moderating effect of the ITIPS on the effect size of the comparison of a treatment and control group, essentially asking whether studies with higher quality treatment integrity procedures demonstrated a greater difference between the treatment and control group post-treatment. Webb et al. (2010) and Collyer et al. (2010) asked the question of whether specific integrity components were related to outcome. This study took a step back from these two studies and asked whether treatment integrity as a whole is related to outcome. Thus, while it is important to acknowledge the contribution of Goense et al. (2016), it is difficult to compare that meta-analysis with the other three. The consistency of findings from the current study lend some confidence to the Webb et al. (2010) and Collyer et al. (2010) meta-analyses.

Interpreting the meta-analyses (Collyer et al., 2019; Webb et al., 2010) together raises some interesting points related to the use of treatment integrity data in understanding how treatment affects change. First, regardless of whether treatment integrity has a small or null correlation with youth client outcomes, it does not appear to explain a great deal of change in

outcomes. Thus, we should reconsider arguments that treatment integrity needs to be assured for the explicit goal of improving client outcomes, rather than as an indicator of internal validity or an outcome of implementation. Treatment integrity, or fidelity as it is often referred to in implementation science, has a number of other useful applications (Proctor et al., 2011; Schoenwald et al., 2011). There is great potential in using treatment integrity data to help researchers and community partners better understand the way that therapists adopt, adapt, and sustain treatments after an implementation process has occurred (Stirman et al., 2012). This has implications for training, supervision, treatment, and the allocation of resources (Shelton et al., 2018). These new lines of research are exciting and impactful, as we are on the verge of asking a broader set of questions than whether the treatment process relates to outcomes, which is: after training and implementation, how do therapists use implemented treatments to best suit the needs of their unique and dynamic settings and patients, and how can we use those data to maximize the benefits of treatment (Barrera et al., 2017).

What is Needed to Advance?

Research resources may be better spent on a broader application of treatment integrity to the treatment process. The treatment process is defined here as the activities and processes involved in the delivery and receipt of treatment (e.g., Doss, 2004; McLeod et al., 2013). These include client change mechanisms and processes, which are intermediary factors hypothesized to be responsible for client's outcome improvement (Doss, 2004; Elliott, 2010). Treatment change processes and mechanisms have received less attention than treatment integrity or the alliance (Elliott, 2010). An example of the impact of change processes and mechanisms is as follows. The delivery of a prescribed therapeutic intervention may not lead directly to change. Rather, the delivery of the intervention (e.g., cognitive restructuring) spurs along the alteration of a client

change mechanism (e.g., changes in negative thoughts) through a client's repeated engagement

in a change process (e.g., repeated use of a thought restructuring exercise), which in turn leads to

changes in client outcomes (e.g., reduction in distress associated with worries).

This aligns with thinking put forth by Perepletchikova (2011, p. 149), who suggested that

other factors may be correlated with outcomes and that treatment integrity is simply a "proxy"

for these types of mechanisms. The importance of mechanisms of change was echoed by a 2014

call from the National Institutes of Mental Health (NIMH) to conduct treatment research with a

focus on identifying and targeting mechanisms of change (Insel & Gogtay, 2014). Given this

paradigm, the study of treatment integrity and outcome alone may be too narrow a scope.

Paradigms that include or account for client change mechanisms and processes, or other more

intermediate outcomes may help to broaden this scope and provide a more comprehensive

picture of the treatment process.

Due to the larger scope, paradigms like this would likely necessitate integrity,

mechanism, and outcome measurement that is pragmatic, feasible, and sustainable (Stanick et al.,

2018). Promising efforts in implementation science assess the use of computer-automated

treatment integrity measurement (see Atkins et al., 2014; Xiao et al., 2015 for examples).

Another low-resource method for approximating treatment integrity is the collection of

worksheets used in treatment (e.g., Stirman et al., 2018). There are also efforts in the

Motivational Interviewing literature to identify client change talk and other client behaviors

through computer-automated processes (Tanana et al., 2016). Other novel techniques, such as

ecological momentary assessment (Shiffman et al., 2008), may prove invaluable in assessing and

tracking client change processes on a moment-to-moment basis. Given that we live in a world

where our data are constantly being collected, it is not hard to image a time where data related to

treatment delivery, client change mechanisms and processes, and outcomes are collected regularly, through subjective and objective measures, with low intrusion and burden. Such processes would allow for researchers to examine change across a number of levels of analyses and answer complex questions about the relation between integrity, client change processes and mechanisms, and outcomes, further clarifying how treatment affects clinical change.

Lack of consistent reporting of methods and correlations are common difficulties when conducting meta-analysis (Pigott & Polanin, 2020; Tipton et al., 2019a). The field should seek some harmony in reporting methods, descriptive data, and statistical information that allows for accurate effect size computation. Inconsistent reporting has been a larger problem in the field; reproducibility is limited, cross-study comparisons are difficult, and we are working hard toward the same end goal, often to find those efforts difficult to compare with others' studies (Open Science Collaboration, 2015). This was also the case in the alliance literature until meta-analyses called for more standardized reporting and continuity of methods (e.g., McLeod, 2011).

In accordance with the effort to standardize reporting in integrity-outcome studies, some recommendations are provided. First, it is critical to report means, standard deviations, and sample sizes for instruments used in analyses at each time point. Second, measures sections or results should contain some language regarding anchors (e.g., what does a high score on this measure mean?), as this allows for meta-analysts to more precisely match the sign to the research question. Third, studies should provide correlation matrices for Pearson's *r* or explicitly provide the type of correlation coefficient that was calculated (e.g., Pearson's *r*, point biserial correlations). Fourth, studies should report effect sizes of any metric (Cohen's *d,* Pearson's *r*, odds ratios) when possible. A more detailed guide on effect sizes can be found in Borenstein et al. (2010). Fifth, studies should consistently report on therapist and client demographics,

including age, race, and ethnicity at the treatment and study level. For parsimony, it may be helpful to adopt the United States Census (United States Census Bureau, 2020) categories for all demographic assessment, as these categories separate race (White American, Black or African American, American Indians or Alaska Native, Asian American, Native Hawaiian or other Pacific Islander) and ethnicity (Latino/Hispanic or not), which was infrequently done by integrity-outcome studies.

The extent of the data requested above may seem unwieldy to anyone who has published in an academic journal. The most obvious place for these data is a supplemental archive housed by the article publisher. However, it has become more common to place descriptive or statistical data that do not fit into manuscripts into online supplements using free tools such as Google Docs (docs.google.com; e.g., Cox et al., 2019). This strategy would prove useful for efforts to gather data for future meta-analyses and systematic reviews. The implementation of such reporting standards and practices is critical for moving forward with a better understanding of the relation between the treatment process and outcome (Pigott & Polanin, 2020).

Alternative Explanations: The Responsiveness Critique

The responsiveness critique must be addressed in integrity-outcome research as a potential explanation of null findings (Stiles & Shapiro, 1989, 1994). This critique posits that the dose-response model of medicine may not adequately account for change that occurs in treatment. In other words, the magnitude of response to an intervention may not reflect the magnitude of dosage of that intervention. This idea has been discussed in some conceptual papers (e.g., Perepletchikova, 2011; Stiles, 1988) but no empirical work has addressed this critique. One potential solution to better understanding the role of responsiveness would be to create a metric of therapist responsiveness that reflects: (1) client requirements for

responsiveness (e.g., literacy), (2) deviation from the therapist's plan, and (3) client's resistance, all of which are important to a therapist's responsiveness (Stiles & Shapiro, 1989, 1994). The responsiveness metric could then be integrated into the context of integrity-outcome relations in any number of ways. Regardless, with no prior empirical work on the subject, this critique is difficult to rule out as an explanation for findings.

Study Limitations

This study was not without factors that limit the interpretability of findings. First, the use of effect size estimation strategies may limit the precision of meta-analysis and also may have had a downstream effect on statistical models (Pigott & Polanin, 2020). Second, this study was limited to youth literature. Perhaps the inclusion of adult literature would provide the power to assess the integrity-outcome relation in more depth. Third, not all studies were double coded, and given that some items in the methodological coding evidenced poor reliability, it is possible that coded factors could have been biased or inaccurate (Pigott & Polanin, 2020). Fourth, the power of this study was too small to reliably conduct many moderator analyses. Finally, there was some indication of publication bias in Model 2, so it is possible that search procedures or publication bias played a role in those findings. These limitations raise questions about the results of the current study, and add additional context for interpreting the meta-analytic and descriptive findings.

Conclusion

More work needs to be done before another meta-analytic effort is taken. Due to the limitations and gaps highlighted in this study, the meta-analyses done so far are limited and inconclusive. However, the consistency of a small or null correlation between treatment integrity and outcomes is striking. The field has pushed treatment integrity into an inferential space to

better understand how the delivery of manualized treatment affects change, and perhaps that space is appropriate. But it is also possible that the dose-response model of medicine may not adequately capture the complexity of treatment and client outcomes, requiring new experimental paradigms. Thus, in order to advance, the field would benefit greatly from consistent reporting standards, the use of appropriate meta-analytic methods, the development and testing of novel or established change mechanisms and processes, and an overall broader scope of understanding how the delivery of psychosocial treatments affects change.

References

Achenbach, T. M. (2005). Advancing assessment of children and adolescents: Commentary on evidence-based assessment of child and adolescent disorders. *Journal of Clinical Child and Adolescent Psychology*, *34*(3), 541-547. https://doi.org/10.1207/s15374424jccp3403_9

Alfano, C. A., Patriquin, M. A., & De Los Reyes, A. (2015). Subjective–objective sleep comparisons and discrepancies among clinically-anxious and healthy children. *Journal of Abnormal Child Psychology*, *43*, 1343-1353. https://doi.org/10.1007/s10802-015-0018-7

[1]Amaya-Jackson, L., Hagele, D., Sideris, J., Potter, D., Briggs, E. C., Keen, L., ... & Socolar, R. (2018). Pilot to policy: statewide dissemination and implementation of evidence-based treatment for traumatized youth. *BMC Health Services Research*, *18*(1), 589. https://doi.org/10.1186/s12913-018-3395-0

American Psychological Association. (2020). PsycEXTRA®. https://www.apa.org/pubs/databases/psycextra/

Andrews, M., Baker, A. L., Halpin, S. A., Lewin, T. J., Richmond, R., Kay-Lambkin, F. J., Filia, S., Castle, D., Williams, J. M., Clark, V., & Callister, R. (2016). Early therapeutic alliance, treatment retention, and 12-month outcomes in a healthy lifestyles intervention for people with psychotic disorders. *The Journal of Nervous and Mental Disease*, *204*(12), 894-902. https://doi.org/10.1097/NMD.0000000000000585

Atkins, D. C., Steyvers, M., Imel, Z. E., & Smyth, P. (2014). Scaling up the evaluation of psychotherapy: evaluating motivational interviewing fidelity via statistical text classification. *Implementation Science*, *9*(1), 49. https://doi.org/10.1186/1748-5908-9-49

Babcock, B. (2015). *What is the formula to calculate the critical value of correlation?*

ResearchGate.

https://www.researchgate.net/post/What_is_the_formula_to_calculate_the_critical_value_of

_correlation

Baker, R., & Jackson, D. (2008). A new approach to outliers in meta-analysis. *Health Care*

Management Science, *11*, 121-131. https://doi.org/10.1007/s10729-007-9041-8

Balduzzi, S., Rücker, G., & Schwarzer, G. (2019). How to perform a meta-analysis with R: a

practical tutorial. *Evidence-Based Mental Health*, *22*(4), 153-160.

http://dx.doi.org/10.1136/ebmental-2019-300117

Barber, J. P., Crits-Christoph, P., & Luborsky, L. (1996). Effects of therapist adherence and

competence on patient outcome in brief dynamic therapy. *Journal of Consulting and*

Clinical Psychology, *64*(3), 619-622. https://doi.org/10.1037/0022-006X.64.3.619

Barber, J. P., Sharpless, B. A., Klostermann, S., & McCarthy, K. S. (2007). Assessing

intervention competence and its relation to therapy outcome: A selected review derived

from the outcome literature. *Professional Psychology: Research and Practice*, 38(5), 493–

500. https://doi.org/10.1037/0735-7028.38.5.493

Barrera, M., Berkel, C., & Castro, F. G. (2017). Directions for the advancement of culturally

adapted preventive interventions: Local adaptations, engagement, and

sustainability. *Prevention Science*, *18*(6), 640-648. https://doi.org/10.1007/s11121-016-

0705-9

Becker, E. M., Becker, K. D., & Ginsburg, G. S. (2012). Modular cognitive behavioral therapy for youth with anxiety disorders: A closer look at the use of specific modules and their relation to treatment process and response. *School Mental Health*, *4*(4), 243-253. https://doi.org/10.1007/s12310-012-9080-2

Bellg, A. J., Borrelli, B., Resnick, B., Hecht, J., Minicucci, D. S., Ory, M., Ogedegbe, G., Orwig, D., Ernst, D., Czajkowski, S., & Treatment Fidelity Workgroup of the NIH Behavior Change Consortium. (2004). Enhancing treatment fidelity in health behavior change studies: Best practices and recommendations from the NIH Behavior Change Consortium. *Health Psychology, 23*(5), 443–451. https://doi.org/10.1037/0278-6133.23.5.443

[29]Berzins, P. (2008). *Therapeutic alliance as a predictor of psychotherapy process and outcome: The role of expert versus novice raters.* [Doctoral dissertation, Fordham University]. ETD Collection for Fordham University.

Bhar, S. S., & Beck, A. T. (2009). Treatment integrity of studies that compare short-term psychodynamic psychotherapy with cognitive-behavior therapy. *Clinical Psychology: Science and Practice*, *16*(3), 370-378. https://doi.org/10.1111/j.1468-2850.2009.01176.x

Bjaastad, J. F., Haugland, B. S. M., Fjermestad, K. W., Torsheim, T., Havik, O. E., Heiervang, E. R., & Öst, L. G. (2016). Competence and Adherence Scale for Cognitive Behavioral Therapy (CAS-CBT) for anxiety disorders in youth: Psychometric properties. *Psychological Assessment*, *28*(8), 908-916. https://doi.org/10.1037/pas0000230

[2]Bjaastad, J. F., Wergeland, G. J. H., Haugland, B. S. M., Gjestad, R., Havik, O. E., Heiervang, E. R., & Öst, L. (2018). Do clinical experience, formal cognitive behavioural therapy

training, adherence, and competence predict outcome in cognitive behavioural therapy for anxiety disorders in youth? *Clinical Psychology & Psychotherapy, 25*(6), 865-877. https://doi.org/10.1002/cpp.2321

[3]Bloomquist, M. L., August, G. J., Lee, S. S., Lee, C. Y. S., Realmuto, G. M., & Klimes-Dougan, B. (2013). Going-to-scale with the Early Risers conduct problems prevention program: Use of a comprehensive implementation support (CIS) system to optimize fidelity, participation and child outcomes. *Evaluation and Program Planning*, *38*, 19-27. https://doi.org/10.1016/j.evalprogplan.2012.11.001

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

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2010). A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research Synthesis Methods*, *1*(2), 97-111. https://doi.org/10.1002/jrsm.12

Boswell, J. F., Gallagher, M. W., Sauer-Zavala, S. E., Bullis, J., Gorman, J. M., Shear, M. K., Woods, S., & Barlow, D. H. (2013). Patient characteristics and variability in adherence and competence in cognitive-behavioral therapy for panic disorder. *Journal of Consulting and Clinical Psychology*, *81*(3), 443-454. https://doi.org/10.1037/a0031437

[4]Boyer, B., MacKay, K. J., McLeod, B. D., & van der Oord, S. (2018). Comparing Alliance in two cognitive-behavioural therapies for adolescents with ADHD using a randomized controlled trial. *Behavior Therapy*, *49*(5), 781-795. https://doi.org/10.1016/j.beth.2018.01.003

Brown, L. A., Craske, M. G., Glenn, D. E., Stein, M. B., Sullivan, G., Sherbourne, C., Bystritsky, A., Welch, S. S., Campbell-Sills, L., Lang, A., Roy-Byrne, P., & Rose, R. D. (2013). CBT competence in novice therapists improves anxiety outcomes. *Depression and Anxiety*, *30*(2), 97-115. https://doi.org/10.1002/da.22027

Bugni, F. A., Canay, I. A., & Shaikh, A. M. (2018). Inference under covariate-adaptive randomization. *Journal of the American Statistical Association*, *113*(524), 1784-1796. https://doi.org/10.1080/01621459.2017.1375934

Campos-Melady, M., Smith, J. E., Meyers, R. J., Godley, S. H., & Godley, M. D. (2017). The effect of therapists' adherence and competence in delivering the adolescent community reinforcement approach on client outcomes. *Psychology of Addictive Behaviors*, *31*(1), 117. https://doi.org/10.1037/adb0000216

Carroll, C., Patterson, M., Wood, S., Booth, A., Rick, J., & Balain, S. (2007). A conceptual framework for implementation fidelity. *Implementation Science*, *2*(1), 40. https://doi.org/10.1186/1748-5908-2-40

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, *6*(4), 284. https://doi.org/10.1037/1040-3590.6.4.284

Cohen, J. (1960). Kappa: Coefficient of concordance. *Educational Psychological Measurement*, *20*(37).

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* L. Erlbaum Associates.

Cohen, J. (1992). A power primer. Psychological Bulletin, 112(1), 155-159. https://doi.org/10.1037/0033-2909.112.1.155

Cohen, J., & Cohen, J. (Eds.). (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed). L. Erlbaum Associates.

Collyer, H., Eisler, I., & Woolgar, M. (2019). Systematic literature review and meta-analysis of the relationship between adherence, competence and outcome in psychotherapy for children and adolescents. *European Child & Adolescent Psychiatry*, *29*(4), 1-15. https://doi.org/10.1007/s00787-018-1265-2

Conn, V. S., Valentine, J. C., Cooper, H. M., & Rantz, M. J. (2003). Grey literature in meta-analyses. *Nursing Research*, *52*(4), 256-261. https://doi.org/10.1097/00006199-200307000-00008

Corey, D. M., Dunlap, W. P., & Burke, M. J. (1998). Averaging correlations: Expected values and bias in combined Pearson rs and Fisher's z transformations. *The Journal of General Psychology*, *125*(3), 245-261. https://doi.org/10.1080/00221309809595548

Cox, J. R., Martinez, R. G., Southam-Gerow, M.A. (2019). Treatment Integrity in Psychotherapy Research and Implications for Implementation Science and the Delivery of Quality Mental Health Services. *Journal of Consulting and Clinical Psychology, 87*(3), 221-233. https://doi.org/10.1037/ccp0000370

Dart, E. H., Collier-Meek, M. A., Chambers, C., & Murphy, A. (2020). Multi-informant

assessment of treatment integrity in the classroom. *Psychology in the Schools, 57*(5).

https://doi.org/10.1002/pits.22351

De Los Reyes, A., Augenstein, T. M., Wang, M., Thomas, S. A., Drabick, D. A. G., Burgers, D.

E., & Rabinowitz, J. (2015). The validity of the multi-informant approach to assessing child

and adolescent mental health. *Psychological Bulletin, 141*(4), 858-900.

http://dx.doi.org/10.1037/a0038498

De Los Reyes, A., & Ohannessian, C. M. (2016). Introduction to the special issue: Discrepancies

in adolescent–parent perceptions of the family and adolescent adjustment. *Journal of Youth*

and Adolescence, *45*(10), 1957-1972. https://doi.org/10.1007/s10964-016-0533-z

Doss, B. D. (2004). Changing the way we study change in psychotherapy. *Clinical Psychology:*

Science and Practice, 11(4), 368-386. https://doi.org/10.1093/clipsy.bph094

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. https://doi.org/

10.1111/j.0006-341x.2000.00455.x

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by

a simple, graphical test. *BMJ*, *315*, 629-634. https://doi.org/10.1136/bmj.315.7109.629

[5]Eiraldi, R., Mautone, J. A., Khanna, M. S., Power, T. J., Orapallo, A., Cacia, J., Schwartz, B. S.,

McCurdy, B., Keiffer, J., Paidipati, C., Kanine, R., Abraham, M., Tulio, S., Swift, L.,

Bressler, S. N., Cabello, B., Jawad, A. F. (2018). Group CBT for externalizing disorders in

urban schools: Effect of training strategy on treatment fidelity and child outcomes. *Behavior Therapy*, *49*(4), 538-550. https://doi.org/10.1016/j.beth.2018.01.001

Elliott, R. (2010). Psychotherapy change process research: Realizing the promise. *Psychotherapy Research*, *20*(2), 123-135. https://doi.org/10.1080/10503300903470743

Elvins, R., & Green, J. (2008). The conceptualization and measurement of therapeutic alliance: An empirical review. Clinical Psychology Review, 28(7), 1167−1187. https://doi.org/10.1016/j.cpr.2008.04.002

Farmer, C. C., Mitchell, K. S., Parker-Guilbert, K., & Galovski, T. E. (2017). Fidelity to the cognitive processing therapy protocol: evaluation of critical elements. *Behavior Therapy*, *48*(2), 195-206. https://doi.org/10.1016/j.beth.2016.02.009

Feeley, M., DeRubeis, R. J., & Gelfand, L. A. (1999). The temporal relation of adherence and alliance to symptom change in cognitive therapy for depression. *Journal of Consulting and Clinical Psychology*, *67*(4), 578-582. https://doi.org/10.1037//0022-006x.67.4.578

Fisher, Z., Tipton, E., Zhipeng, H. (2017). *robumeta*: Robust Variance Meta-Regression. R package version 2.0. https://CRAN.R-project.org/package=robumeta

Fixsen, D. L., Van Dyke, M., & Blase, K. A. (2019). Implementation science: Fidelity predictions and outcomes. Chapel Hill, NC: Active Implementation Research Network. www.activeimplementation.org/resources

Fryling, M. J., Wallace, M. D., & Yassine, J. N. (2012). Impact of treatment integrity on

    intervention effectiveness. *Journal of Applied Behavior Analysis*, *45*(2), 449-453.

    https://doi.org/10.1901/jaba.2012.45-449

[6]Garner, B. R., Godley, S. H., Funk, R. R., Dennis, M. L., Smith, J. E., & Godley, M. D. (2009).

    Exposure to adolescent community reinforcement approach treatment procedures as a

    mediator of the relationship between adolescent substance abuse treatment retention and

    outcome. *Journal of Substance Abuse Treatment, 36*(3), 252-264.

    https://doi.org/10.1016/j.jsat.2008.06.007

[7]Gillespie, M. L., Huey Jr, S. J., & Cunningham, P. B. (2017). Predictive validity of an observer-

    rated adherence protocol for multisystemic therapy with juvenile drug offenders. *Journal of*

    Substance Abuse Treatment, 76, 1-10. https://doi.org/10.1016/j.jsat.2017.01.001

[8]Gillham, J. E., Hamilton, J., Freres, D. R., Patton, K., & Gallop, R. (2006). Preventing

    depression among early adolescents in the primary care setting: A randomized controlled

    study of the Penn Resiliency Program. *Journal of Abnormal Child Psychology, 34*(2), 195-

    211.

[9]Ginsburg, G. S., Becker, K. D., Drazdowski, T. K., & Tein, J. Y. (2012). Treating anxiety

    disorders in inner city schools: Results from a pilot randomized controlled trial comparing

    CBT and usual care. *Child & Youth Care Forum, 41*(1), 1-19.

    https://doi.org/10.1007/s10566-011-9156-4

Ginzburg, D. M., Bohn, C., Höfling, V., Weck, F., Clark, D. M., & Stangier, U. (2012).

    Treatment specific competence predicts outcome in cognitive therapy for social anxiety

disorder. *Behaviour Research and Therapy*, *50*(12), 747-752.

https://doi.org/10.1016/j.brat.2012.09.001

Goense, P. B., Assink, M., Stams, G. J., Boendermaker, L., & Hoeve, M. (2016). Making 'what

works' work: A meta-analytic study of the effect of treatment integrity on outcomes of

evidence-based interventions for juveniles with antisocial behavior. *Aggression and Violent*

Behavior, *31*, 106-115. https://doi.org/10.1016/j.avb.2016.08.003

Goense, P., Boendermaker, L., van Yperen, T., Stams, G. J., & van Laar, J. (2014).

Implementation of treatment integrity procedures. *Zeitschrift für Psychologie*. *222*, 12-21.

https://doi.org/10.1027/2151-2604/a000161

Goldman, G. A., & Gregory, R. J. (2009). Preliminary relationships between adherence and

outcome in dynamic deconstructive psychotherapy. *Psychotherapy: Theory, Research,*

Practice, Training, *46*(4), 480. https://doi.org/10.1037/a0017947

[10]Graham, C., Carr, A., Rooney, B., Sexton, T., & Wilson Satterfield, L. R. (2014). Evaluation of

functional family therapy in an Irish context. *Journal of Family Therapy*, *36*(1), 20-38.

https://doi.org/10.1111/1467-6427.12028

Graves, T. A., Tabri, N., Thompson-Brenner, H., Franko, D. L., Eddy, K. T., Bourion-Bedes, S.,

Brown, A., Constantino, M. J., Fluckiger, C., Forsberg, S., Isserlin, L., Couturier, J.,

Karlsson, G. P., Mander, J., Teufel, M., Mitchell, J. E., Crosby, R. D., Prestano, C., Satir, D.

A., ... & Thomas, J. J. (2017). A meta-analysis of the relation between therapeutic alliance

and treatment outcome in eating disorders. *International Journal of Eating Disorders 55*(4),

323-340. https://doi.org/10.1002/eat.22672

Gutermann, J., Schreiber, F., Matulis, S., Stangier, U., Rosner, R., & Steil, R. (2015).

    Therapeutic adherence and competence scales for Developmentally Adapted Cognitive

    Processing Therapy for adolescents with PTSD. *European Journal of*

    Psychotraumatology, 6, 26632. https://doi.org/10.3402/ejpt.v6.26632

Harris, P. A., Taylor, R., Minor, B. L., Elliott, V., Fernandez, M., O'Neal, L., McLeod, L.,

    Delacqua, G., Delacqua, F., Duda, S. N., & REDCap Consortium (2019). The REDCap

    consortium: Building an international community of software platform partners. *Journal of*

    Biomedical Informatics, *95*, 103208. https://doi.org/10.1016/j.jbi.2019.103208

Harris, P. A., Taylor, R., Thielke, R., Payne, J., Gonzalez, N., & Conde, J. G. (2009). A

    metadata-driven methodology and workflow process for providing translational research

    informatics support. *Journal of Biomedical Informatics*, *42*(2), 377-381.

    https://doi.org/10.1016/j.jbi.2008.08.010

[11]Hartnett, D., Carr, A., & Sexton, T. (2016). The Effectiveness of Functional Family Therapy in

    Reducing Adolescent Mental Health Risk and Family Adjustment Difficulties in an Irish

    Context. *Family Process, 55*(2), 287-304. https://doi.org/10.1111/famp.12195

Haug, T., Nordgreen, T., Öst, L. G., Tangen, T., Kvale, G., Hovland, O. J., Heiervang, E. R., &

    Havik, O. E. (2016). Working alliance and competence as predictors of outcome in

    cognitive behavioral therapy for social anxiety and panic disorder in adults. *Behaviour*

    Research and Therapy, *77*, 40-51. https://doi.org/10.1016/j.brat.2015.12.004

Hedges, L. V. (1994). Fixed effect models. In H. E. Cooper, & L. V. Hedges (Ed.), *Handbook of*

    research synthesis (pp. 285-299). Russell Sage Foundation.

Hedges, L. V., Tipton, E., & Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods*, *1*(1), 39-65. https://doi.org/10.1002/jrsm.5

Henggeler, S. W. (1999). Multisystemic therapy: An overview of clinical procedures, outcomes, and policy implications. *Child Psychology and Psychiatry Review*, *4*(1), 2-10. https://doi.org/10.1097/00004583-199911000-00006

[12]Henggeler, S. W., Pickrel, S. G., & Brondino, M. J. (1999). Multisystemic treatment of substance-abusing and-dependent delinquents: Outcomes, treatment fidelity, and transportability. *Mental Health Services Research, 1*(3), 171-184. https://doi.org/ 10.1023/a:1022373813261

Herschell, A. D., Quetsch, L. B., & Kolko, D. J. (2019). Measuring Adherence to Key Teaching Techniques in an Evidence-Based Treatment: A Comparison of Caregiver, Therapist, and Behavior Observation Ratings. *Journal of Emotional and Behavioral Disorders*, 1-12. https://doi.org/10.1177/1063426618821901

Higgins, J. P., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*, *21*(11), 1539-1558. https://doi.org/10.1002/sim.1186

Hilsenroth, M. J., Ackerman, S. J., Blagys, M. D., Baity, M. R., & Mooney, M. A. (2003). Short-term psychodynamic psychotherapy for depression: An examination of statistical, clinically significant, and technique-specific change. *The Journal of Nervous and Mental Disease*, *191*(6), 349-357. https://doi.org/10.1097/01.NMD.0000071582.11781.67

Hoagwood, K., Jensen, P. S., Petti, T., & Burns, B. J. (1996). Outcomes of mental health care for children and adolescents: I. A comprehensive conceptual model. *Journal of the American Academy of Child & Adolescent Psychiatry*, *35*(8), 1055-1063. https://doi.org/10.1097/00004583-199608000-00017

Hoffart, A., Sexton, H., Nordahl, H. M., & Stiles, T. C. (2005). Connection between patient and therapist and therapist's competence in schema-focused therapy of personality problems. *Psychotherapy Research*, *15*(4), 409-419. https://doi.org/10.1080/10503300500091702

Hofmann, S. G., Asnaani, A., Vonk, I. J., Sawyer, A. T., & Fang, A. (2012). The efficacy of cognitive behavioral therapy: A review of meta-analyses. *Cognitive Therapy and Research*, *36*(5), 427-440. https://doi.org/10.1007/s10608-012-9476-1

Hogue, A., Dauber, S., Lichvar, E., Bobek, M., & Henderson, C. E. (2015). Validity of therapist self-report ratings of fidelity to evidence-based practices for adolescent behavior problems: Correspondence between therapists and observers. *Administration and Policy in Mental Health and Mental Health Services Research*, *42*(2), 229-243. https://doi.org/10.1007/s10488-014-0548-2

[13]Hogue, A., Henderson, C. E., Dauber, S., Barajas, P. C., Fried, A., & Liddle, H. A. (2008). Treatment adherence, competence, and outcome in individual and family therapy for adolescent behavior problems. *Journal of Consulting and Clinical Psychology*, *76*(4), 544-555. https://doi.org/10.1037/0022-006X.76.4.544

[28]Hogue, A., Liddle, H. A., Dauber, S., & Samuolis, J. (2004). Linking Session Focus to

Treatment Outcome in Evidence-Based Treatments for Adolescent Substance

Abuse. *Psychotherapy: Theory, Research, Practice, Training*, *41*(2), 83-96.

https://doi.org/10.1037/0033-3204.41.2.83

Hogue, A., Liddle, H. A., & Rowe, C. (1996). Treatment adherence process research in family

therapy: A rationale and some practical guidelines. *Psychotherapy: Theory, Research,*

Practice, Training, 33(2), 332–345. https://doi.org/10.1037/0033-3204.33.2.332

Holder, N., Holliday, R., Williams, R., Mullen, K., & Surís, A. (2018). A preliminary

examination of the role of psychotherapist fidelity on outcomes of cognitive processing

therapy during an RCT for military sexual trauma-related PTSD. *Cognitive Behaviour*

Therapy, 47(1), 76-89. https://doi.org/10.1080/16506073.2017.1357750

[14]Holth, P., Torsheim, T., Sheidow, A. J., Ogden, T., & Henggeler, S. W. (2011). Intensive

quality assurance of therapist adherence to behavioral interventions for adolescent substance

use problems. *Journal of Child & Adolescent Substance Abuse, 20*(4), 289-313.

https://doi.org/10.1080/1067828X.2011.581974

Howitt, D. & Cramer, D. (2014) *Introduction to Statistics*. (6th ed.). Pearson.

Huedo-Medina, T. B., Sánchez-Meca, J., Marín-Martínez, F., & Botella, J. (2006). Assessing

heterogeneity in meta-analysis: Q statistic or I² index? *Psychological Methods, 11*(2), 193–

206. https://doi.org/10.1037/1082-989X.11.2.193

[15]Huey Jr, S. J., Henggeler, S. W., Brondino, M. J., & Pickrel, S. G. (2000). Mechanisms of change in multisystemic therapy: reducing delinquent behavior through therapist adherence and improved family and peer functioning. *Journal of Consulting and Clinical Psychology*, *68*(3), 451-467. https://doi.org/10.1037/0022-006X.68.3.451

Hunsley, J., & Mash, E. J. (2005). Introduction to the special section on developing guidelines for the evidence-based assessment (EBA) of adult disorders. *Psychological Assessment*, *17*(3), 251-255. https://doi.org/10.1037/1040-3590.17.3.251

Hunsley, J., & Mash, E. J. (2007). Evidence-based assessment. *Annual Review of Clinical Psychology*, *3*, 29-51. https://doi.org/10.1146/annurev.clinpsy.3.022806.091419

IBM Corp. Released 2019. IBM SPSS Statistics for Windows, Version 26.0 [Computer software]. Armonk, NY: IBM Corp.

Insel, T. R., & Gogtay, N. (2014). National Institute of Mental Health clinical trials: new opportunities, new expectations. *JAMA Psychiatry*, *71*(7), 745-746. 10.1001/jamapsychiatry.2014.426

IntHout, J., Ioannidis, J. P., Rovers, M. M., & Goeman, J. J. (2016). Plea for routinely presenting prediction intervals in meta-analysis. *BMJ*, *6*(7), e010247. http://dx.doi.org/10.1136/bmjopen-2015-010247

Jacob, N., Neuner, F., Maedl, A., Schaal, S., & Elbert, T. (2014). Dissemination of psychotherapy for trauma spectrum disorders in postconflict settings: A randomized

controlled trial in Rwanda. *Psychotherapy and Psychosomatics*, *83*(6), 354-363. https://doi.org/10.1159/000365114

Judd, C. M., & Kenny, D. A. (1981). Process analysis: Estimating mediation in treatment evaluations. *Evaluation Review*, *5*(5), 602-619. https://doi.org/10.1177/0193841X8100500502

Kazdin, A. E. (2003). *Research Design in Clinical Psychology.* Allyn & Bacon.

Kazdin, A. E. (2007). Mediators and mechanisms of change in psychotherapy research. *Annual Review of Clinical Psychology*, *3*, 1-27. https://doi.org/10.1146/annurev.clinpsy.3.022806.091432

Kendall, P. C., & Hedtke, K. (2006). *Cognitive-Behavioral Therapy for anxious children: Therapist Manual.* 3rd ed. Workbook Publishing

Khoury, B., Lecomte, T., Fortin, G., Masse, M., Therien, P., Bouchard, V., Paquin, K., & Hofmann, S. G. (2013). Mindfulness-based therapy: a comprehensive meta-analysis. *Clinical Psychology Review*, *33*(6), 763-771. https://doi.org/10.1016/j.cpr.2013.05.005

Kossmeier, M., Tran, U. S., Voracek, M. (2020). *metaviz*: Forest Plots, Funnel Plots, and Visual Funnel Plot Inference for Meta-Analysis. R package version 0.3.1. https://CRAN.R-project.org/package=metaviz

Kushner, M. G., Maurer, E. W., Thuras, P., Donahue, C., Frye, B., Menary, K. R., ... & Van Demark, J. (2013). Hybrid cognitive behavioral therapy versus relaxation training for co-

occurring anxiety and alcohol disorder: A randomized clinical trial. *Journal of Consulting and Clinical Psychology*, *81*(3), 429-442. https://doi.org/10.1037/a0031301

Kuyken, W., & Tsivrikos, D. (2009). Therapist competence, comorbidity and cognitive-behavioral therapy for depression. *Psychotherapy and Psychosomatics*, *78*(1), 42-48. https://doi.org/10.1159/000172619

Lagattuta, K. H., Sayfan, L., & Bamford, C. (2012). Do you know how I feel? Parents underestimate worry and overestimate optimism compared to child self-report. *Journal of Experimental Child Psychology*, *113*(2), 211-232. https://doi.org/10.1016/j.jecp.2012.04.001

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1) 159-174. https://doi.org/10.2307/2529310

[16]Lange, A. M., van der Rijken, R. E., Busschbach, J. J., Delsing, M. J., & Scholte, R. H. (2017). It's not just the therapist: Therapist and country-wide experience predict therapist adherence and adolescent outcome. *Child & Youth Care Forum*, *46*(4), 455-471. https://doi.org/10.1007/s10566-016-9388-4

[17]Lange, A. M., van der Rijken, R. E., Delsing, M. J., Busschbach, J. J., & Scholte, R. H. (2019). Development of therapist adherence in relation to treatment outcomes of adolescents with behavioral problems. *Journal of Clinical Child & Adolescent Psychology*, *48*(sup1), S337-S346. https://doi.org/10.1080/15374416.2018.1477049

Laws, H. B., Constantino, M. J., Sayer, A. G., Klein, D. N., Kocsis, J. H., Manber, R., Markowitz, J. C., Rothbaum, B. O., Steidtman, D., Thase, M. E. & Arnow, B. A. (2017).

Convergence in patient–therapist therapeutic alliance ratings and its relation to outcome
in chronic depression treatment. *Psychotherapy Research*, *27*(4), 410-424.

Lenhard, W. & Lenhard, A. (2016). Calculation of Effect Sizes. Retrieved from:
https://www.psychometrica.de/effect_size.html. Dettelbach (Germany): Psychometrica.
https://doi.org/10.13140/RG.2.2.17823.92329

[18]Liber, J. M., McLeod, B. D., Van Widenfelt, B. M., Goedhart, A. W., van der Leeden, A. J.,
Utens, E. M., & Treffers, P. D. (2010). Examining the relation between the therapeutic
alliance, treatment adherence, and outcome of cognitive behavioral therapy for children
with anxiety disorders. *Behavior Therapy*, *41*(2), 172-186.
https://doi.org/10.1016/j.beth.2009.02.003

Lipsey, M. W. (2003). Those confounded moderators in meta-analysis: Good, bad, and ugly. *The
Annals of the American Academy of Political and Social Science*, *587*, 69-81.
https://doi.org/10.1177/0002716202250791

Lipsey, M. W., & Wilson, D. B. (2001). *Applied social research methods series; Vol. 49.
Practical meta-analysis.* Sage Publications, Inc.

[19]Löfholm, C. A., Eichas, K., & Sundell, K. (2014). The Swedish implementation of
multisystemic therapy for adolescents: Does treatment experience predict treatment
adherence? *Journal of Clinical Child & Adolescent Psychology*, *43*(4), 643-655.
https://doi.org/10.1080/15374416.2014.883926

Lopez, M. A., & Basco, M. A. (2015). Effectiveness of cognitive behavioral therapy in public
mental health: Comparison to treatment as usual for treatment-resistant

depression. *Administration and Policy in Mental Health and Mental Health Services Research*, *42*(1), 87-98. https://doi.org/10.1007/s10488-014-0546-4

Luborsky, L., McLellan, A. T., Woody, G. E., O'Brien, C. P., & Auerbach, A. (1985). Therapist success and its determinants. *Archives of General Psychiatry*, *42*(6), 602-611. https://doi.org/10.1001/archpsyc.1985.01790290084010

Margison, F. R., Barkham, M., Evans, C., McGrath, G., Clark, J. M., Audin, K., & Connell, J. (2000). Measurement and psychotherapy: Evidence-based practice and practice-based evidence. *The British Journal of Psychiatry*, *177*(2), 123-130. https://doi.org/10.1192/bjp.177.2.123

Martin, D. J., Graske, J. P., & Davis, M. K. (2000). Relation of the therapeutic alliance with outcome and other variables: A meta-analytic review. *Journal of Consulting and Clinical Psychology*, *68*(3), 438–450. https://doi.org/10.1037/0022-006X.68.3.438

Martinez, R. G., Lewis, C. C., & Weiner, B. J. (2014). Instrumentation issues in implementation science. *Implementation Science*, *9*(1), 118. https://doi.org/10.1186/s13012-014-0118-8

Marziali, E. A. (1984). Prediction of outcome of brief psychotherapy from therapist interpretive interventions. *Archives of General Psychiatry*, *41*(3), 301-304. https://doi.org/10.1001/archpsyc.1984.01790140091011

Mash, E. J., & Hunsley, J. (2005). Evidence-based assessment of child and adolescent disorders: Issues and challenges. *Journal of Clinical Child and Adolescent Psychology*, *34*(3), 362-379. https://doi.org/10.1207/s15374424jccp3403_1

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemica Medica*, *22*(3), 276-282. https://doi.org/10.11613/BM.2012.031

McHugh, R. K., Murray, H. W., & Barlow, D. H. (2009). Balancing fidelity and adaptation in the dissemination of empirically-supported treatments: The promise of transdiagnostic interventions. *Behaviour Research and Therapy*, *47*(11), 946-953. https://doi.org/10.1016/j.brat.2009.07.005

McLeod, B. D. (2011). Relation of the alliance with outcomes in youth psychotherapy: A meta-analysis. *Clinical Psychology Review*, *31*(4), 603-616. https://doi.org/10.1016/j.cpr.2011.02.003

McLeod, B. D., Smith, M. M., Southam-Gerow, M. A., Weisz, J. R., & Kendall, P. C. (2015). Measuring treatment differentiation for implementation research: The Therapy Process Observational Coding System for Child Psychotherapy Revised Strategies Scale. *Psychological Assessment, 27*(1), 314–325. https://doi.org/10.1037/pas0000037

McLeod, B. D., Southam-Gerow, M. A., Jensen-Doss, A., Hogue, A., Kendall, P. C., & Weisz, J. R. (2019). Benchmarking treatment adherence and therapist competence in individual cognitive-behavioral treatment for youth anxiety disorders. *Journal of Clinical Child & Adolescent Psychology*, *48*(sup1), S234-S246. https://doi.org/10.1080/15374416.2017

McLeod, B. D., Southam-Gerow, M. A., Tully, C. B., Rodríguez, A., & Smith, M. M. (2013). Making a case for treatment integrity as a psychosocial treatment quality indicator for youth mental health care. *Clinical Psychology: Science and Practice, 20*(1), 14-32. https://doi.org/10.1111/cpsp.12020

McLeod, B. D., Southam-Gerow, M. A., & Weisz, J. R. (2009). Conceptual and methodological issues in treatment integrity measurement. *School Psychology Review*, *38*(4), 541-546.

Meier, A., McGovern, M. P., Lambert-Harris, C., McLeman, B., Franklin, A., Saunders, E. C., & Xie, H. (2015). Adherence and competence in two manual-guided therapies for co-occurring substance use and posttraumatic stress disorders: clinician factors and patient outcomes. *The American Journal of Drug and Alcohol Abuse*, *41*(6), 527-534. https://doi.org/10.3109/00952990.2015.1062894

Miller, S. J., & Binder, J. L. (2002). The effects of manual-based training on treatment fidelity and outcome: A review of the literature on adult individual psychotherapy. *Psychotherapy: Theory, Research, Practice, Training*, *39*(2), 184-198. https://doi.org/10.1037/0033-3204.39.2.184

Moncher, F. J., & Prinz, R. J. (1991). Treatment fidelity in outcome studies. *Clinical Psychology Review*, *11*(3), 247-266. https://doi.org/10.1016/0272-7358(91)90103-2

Mowbray, C. T., Holter, M. C., Teague, G. B., & Bybee, D. (2003). Fidelity criteria: Development, measurement, and validation. *American Journal of Evaluation, 24*(3), 315–340. https://doi.org/10.1177/109821400302400303

Muse, K., & McManus, F. (2013). A systematic review of methods for assessing competence in cognitive–behavioural therapy. *Clinical Psychology Review*, *33*(3), 484-499. https://doi.org/10.1016/j.cpr.2013.01.010

O'Malley, S. S., Foley, S. H., Rounsaville, B. J., Watkins, J. T., Sotsky, S. M., Imber, S. D., & Elkin, I. (1988). Therapist competence and patient outcome in interpersonal

psychotherapy of depression. *Journal of Consulting and Clinical Psychology*, *56*(4), 496-501. https://doi.org/10.1037/0022-006X.56.4.496

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716. https://doi.org/10.1126/science.aac4716

Owen, J., & Hilsenroth, M. J. (2014). Treatment adherence: The importance of therapist flexibility in relation to therapy outcomes. *Journal of Counseling Psychology*, *61*(2), 280-288. https://doi.org/10.1037/a0035753

Page-Gould, E. (2015). *What is the formula to calculate the critical value of correlation?* ResearchGate. https://www.researchgate.net/post/What_is_the_formula_to_calculate_the_critical_value_of__correlation

Peng, P., Lin, X., Ünal, Z. E., Lee, K., Namkung, J., Chow, J., & Sales, A. (2020). Examining the mutual relations between language and mathematics: A meta-analysis. *Psychological Bulletin*, 2. https://doi.org/10.1037/bul0000231

Perepletchikova, F. (2006a). *Implementation of Treatment Integrity Procedures Scale.* http://treatmentintegrity.com/HTMLDocs/itips.htm

Perepletchikova, F. (2006b). *Implementation of Treatment Integrity Procedures Scale: Rater manual.* http://treatmentintegrity.com/HTMLDocs/itipsratermanual.html

Perepletchikova, F. (2006c). *Literature search procedures: List of terms.* http://treatmentintegrity.com/HTMLDocs/listofterms.htm

Perepletchikova, F. (2011). On the topic of treatment integrity. *Clinical Psychology: Science and Practice*, *18*(2), 148-153. https://doi.org/10.1111/j.1468-2850.2011.01246.x

Perepletchikova, F., & Kazdin, A. E. (2005). Treatment integrity and therapeutic change: Issues and research recommendations. *Clinical Psychology: Science and Practice*, *12*(4), 365-383. https://doi.org/10.1093/clipsy.bpi045

Perepletchikova, F., Treat, T. A., & Kazdin, A. E. (2007). Treatment integrity in psychotherapy research: analysis of the studies and examination of the associated factors. *Journal of Consulting and Clinical Psychology, 75*(6), 829-841. https://doi.org/10.1037/0022-006X.75.6.829

Peterson, L., Homer, A. L., & Wonderlich, S. A. (1982). The integrity of independent variables in behavior analysis. *Journal of Applied Behavior Analysis*, *15*(4), 477-492. https://doi.org/10.1901/jaba.1982.15-477

Pigott, T. D. (2001). Missing predictors in models of effect size. Evaluation & the Health Professions, 24(3), 277-307. https://doi.org/10.1177/01632780122034920

Pigott, T. D., & Polanin, J. R. (2020). Methodological Guidance Paper: High-Quality Meta-Analysis in a Systematic Review. *Review of Educational Research*, *90*(1), 24-46. https://doi.org/10.3102/0034654319877153

Piper, W. E., Debbane, E. G., Bienvenu, J. P., de Carufel, F., & Garant, J. (1986). Relationships between the object focus of therapist interpretations and outcome in short-term individual psychotherapy. *British Journal of Medical Psychology*, *59*(1), 1-11. https://doi.org/10.1111/j.2044-8341.1986.tb02659.x

[20]Podell, J. L., Kendall, P. C., Gosch, E. A., Compton, S. N., March, J. S., Albano, A.-M., Rynn, M. A., Walkup, J. T., Sherrill, J. T., Ginsburg, G. S., Keeton, C. P., Birmaher, B., & Piacentini, J. C. (2013). Therapist factors and outcomes in CBT for anxiety in youth. *Professional Psychology: Research and Practice, 44*(2), 89-98. https://doi.org/10.1037/a0031700

Polanin, J. R., Tanner-Smith, E. E., & Hennessy, E. A. (2016). Estimating the difference between published and unpublished effect sizes: A meta-review. *Review of Educational Research*, *86*(1), 207-236. https://doi.org/10.3102/0034654315582067

Proctor, E., Silmere, H., Raghavan, R., Hovmand, P., Aarons, G., Bunger, A., ... & Hensley, M. (2011). Outcomes for implementation research: conceptual distinctions, measurement challenges, and research agenda. Administration and Policy in Mental Health and Mental Health Services Research, 38(2), 65-76. https://doi.org/10.1007/s10488-010-0319-7

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/.

Rapley, H. A., & Loades, M. E. (2019). A systematic review exploring therapist competence, adherence, and therapy outcomes in individual CBT for children and young people. *Psychotherapy Research, 29*(8), 1010-1019. https://doi.org/10.1080/10503307.2018

Revelle, W. (2019) *psych*: Procedures for Personality and Psychological Research, Northwestern University, Evanston, Illinois, USA, https://CRAN.R-project.org/package=psych. Version = 1.9.12.

Riley, R. D., Higgins, J. P., & Deeks, J. J. (2011). Interpretation of random effects meta-analyses. *BMJ*, *342*, d549. https://doi.org/10.1136/bmj.d549

[30]Robbins, M. S., Feaster, D. J., Horigian, V. E., Puccinelli, M. J., Henderson, C., & Szapocznik, J. (2011). Therapist adherence in brief strategic family therapy for adolescent drug abusers. *Journal of Consulting and Clinical Psychology*, *79*(1), 43-53. https://doi.org/10.1037/a0022146

Rosenthal D. A., Hoyt W. T., Ferrin J. M., Miller S., Cohen N. D. (2006). Advanced methods in meta-analytic research: Applications and implications for rehabilitation counseling research. *Rehabilitation Counseling Bulletin, 49*(4), 234–246. https://doi.org/10.1177/00343552060490040501

[21]Rowe, C., Rigter, H., Henderson, C., Gantner, A., Mos, K., Nielsen, P., & Phan, O. (2013). Implementation fidelity of Multidimensional Family Therapy in an international trial. *Journal of Substance Abuse Treatment*, *44*(4), 391-399. https://doi.org/10.1016/j.jsat.2012.08.225

Ryum, T., Stiles, T. C., Svartberg, M., & McCullough, L. (2010). The effects of therapist competence in assigning homework in cognitive therapy with cluster C personality disorders: Results from a randomized controlled trial. *Cognitive and Behavioral Practice*, *17*(3), 283-289. https://doi.org/10.1016/j.cbpra.2009.10.005

RStudio Team (2020). *RStudio: Integrated Development for R. RStudio, Inc., Boston, MA* URL http://www.rstudio.com/.

Sachs, J. S. (1983). Negative factors in brief psychotherapy: An empirical assessment. *Journal of Consulting and Clinical Psychology*, *51*(4), 557-564. https://doi.org/10.1037/0022-006X.51.4.557

Sanetti, L. M. H., & Kratochwill, T. R. (2011). An evaluation of the treatment integrity planning protocol and two schedules of treatment integrity self-report: Impact on implementation and report accuracy. *Journal of Educational and Psychological Consultation*, *21*(4), 284-308. https://doi.org/10.1080/10474412.2011.620927

Sasso, K. E., Strunk, D. R., Braun, J. D., DeRubeis, R. J., & Brotman, M. A. (2016). A re-examination of process–outcome relations in cognitive therapy for depression: Disaggregating within-patient and between-patient effects. *Psychotherapy Research*, *26*(4), 387-398. https://doi.org/10.1080/10503307.2015.1026423

Schoenwald, S. K., Chapman, J. E., Sheidow, A. J., & Carter, R. E. (2009). Long-term youth criminal outcomes in MST transport: The impact of therapist adherence and organizational climate and structure. *Journal of Clinical Child & Adolescent Psychology*, *38*(1), 91-105. https://doi.org/10.1080/15374410802575388

Schoenwald, S. K., Garland, A. F., Chapman, J. E., Frazier, S. L., Sheidow, A. J., & Southam-Gerow, M. A. (2011). Toward the effective and efficient measurement of implementation fidelity. *Administration and Policy in Mental Health and Mental Health Services Research*, *38*(1), 32-43. https://doi.org/10.1007/s10488-010-0321-0

Schoenwald, S. K., Halliday-Boykins, C. A., & Henggeler, S. W. (2003). Client-level Predictors of Adherence to MST in Community Service Settings. *Family Process*, *42*(3), 345-359. https://doi.org/10.1111/j.1545-5300.2003.00345.x

[23]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. https://doi.org/10.1037/a0013788

[24]Schoenwald, S. K., Sheidow, A. J., & Letourneau, E. J. (2004). Toward effective quality assurance in evidence-based practice: Links between expert consultation, therapist fidelity, and child outcomes. *Journal of Clinical Child and Adolescent Psychology*, *33*(1), 94-104. https://doi.org/10.1207/S15374424JCCP3301_10

[22]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. https://doi.org/10.1023/A:1026229102151

Schulte, A. C., Easton, J. E., & Parker, J. (2009). Advances in treatment integrity research: Multidisciplinary perspectives on the conceptualization, measurement, and enhancement of treatment integrity. *School Psychology Review*, *38*(4), 460-475.

Sedgwick, P. (2012). Pearson's correlation coefficient. *BMJ, 345*, e4483. https://doi.org/10.1136/bmj.e4483

Serralta, F. B., Pole, N., Tiellet Nunes, M. L., Eizirik, C. L., & Olsen, C. (2010). The process of change in brief psychotherapy: Effects of psychodynamic and cognitive–behavioral

prototypes. *Psychotherapy Research*, *20*(5), 564-575.

https://doi.org/10.1080/10503307.2010.493537

Sexton, T. L., Ridley, C. R., & Kleiner, A. J. (2004). Beyond common factors: multilevel-
process models of therapeutic change in marriage and family therapy. *Journal of Marital
and Family Therapy, 30*(2), 131-149. https://doi.org/10.1111/j.1752-0606.2004.tb01229.x

Shadish, W. R., & Haddock, C. K. (1994). Combining estimates of effect size. In H. Cooper & L.
V. Hedges (Eds.), *The Handbook of Research Synthesis* (pp. 261-281). Russell Sage
Foundation.

Shelton, R. C., Cooper, B. R., & Stirman, S. W. (2018). The sustainability of evidence-based
interventions and practices in public health and health care. *Annual Review of Public
Health*, *39*, 55-76. https://doi.org/10.1186/1748-5908-7-17

Shiffman, S., Stone, A. A., & Hufford, M. R. (2008). Ecological momentary assessment. *Annual
Review of Clinical Psychology*, *4*, 1-32.
https://doi.org/10.1146/annurev.clinpsy.3.022806.091415

Shrout, P. E., Spitzer, R. L., & Fleiss, J. L. (1987). Quantification of agreement in psychiatric
diagnosis revisited. *Archives of General Psychiatry*, *44*(2), 172-177.
https://doi.org/10.1001/archpsyc.1987.01800140084013

Sikkema, K. J., Ranby, K. W., Meade, C. S., Hansen, N. B., Wilson, P. A., & Kochman, A.
(2013). Reductions in traumatic stress following a coping intervention were mediated by
decreases in avoidant coping for people living with HIV/AIDS and childhood sexual

abuse. *Journal of Consulting and Clinical Psychology*, *81*(2), 274-283.

https://doi.org/10.1037/a0030144

Silverman, W. K., & Hinshaw, S. P. (2008). The second special issue on evidence-based

psychosocial treatments for children and adolescents: A 10-year update. *Journal of Clinical*

Child & Adolescent Psychology, *37*(1), 1-7. https://doi.org/10.1080/15374410701817725

[25]Smith, D. C., Hall, J. A., Jang, M., & Arndt, S. (2009). Therapist adherence to a motivational-

interviewing intervention improves treatment entry for substance-misusing adolescents with

low problem perception. *Journal of Studies on Alcohol and Drugs*, *70*(1), 101-105.

https://doi.org/10.15288/jsad.2009.70.101

Smith, M. M., McLeod, B. D., Southam-Gerow, M. A., Jensen-Doss, A., Kendall, P. C., &

Weisz, J. R. (2017). Does the delivery of CBT for youth anxiety differ across research and

practice settings? *Behavior Therapy*, *48*(4), 501-516.

https://doi.org/10.1016/j.beth.2016.07.004

Society for Implementation Research and Collaboration. (2018). Proceedings of the 4th Biennial

Conference of the Society for Implementation Research Collaboration (SIRC) 2017:

implementation mechanisms: what makes implementation work and why? part

2. *Implementation Science, 13*(39). https://doi.org/10.1186/s13012-018-0715-z

Southam-Gerow, M. A., & McLeod, B. D. (2013). Advances in applying treatment integrity

research for dissemination and implementation science: Introduction to special

issue. *Clinical Psychology: Science and Practice*, *20*(1), 1-13.

https://doi.org/10.1111/cpsp.12019

Stanick, C. F., Halko, H. M., Dorsey, C. N., Weiner, B. J., Powell, B. J., Palinkas, L. A., &

    Lewis, C. C. (2018). Operationalizing the 'pragmatic'measures construct using a

    stakeholder feedback and a multi-method approach. *BMC Health Services Research*, *18*.

    https://doi.org/10.1186/s12913-018-3709-2

StataCorp. (2020). Released 2019. Stata Statistical Software: Release 16. [Computer software].

    College Station, TX: StataCorp LLC

Stiles, W. B. (1988). Psychotherapy process-outcome correlations may be

    misleading. *Psychotherapy: Theory, Research, Practice, Training*, *25*(1), 27-35.

Stiles, W. B., & Shapiro, D. A. (1989). Abuse of the drug metaphor in psychotherapy process-

    outcome research. *Clinical Psychology Review*, *9*(4), 521-543. https://doi.org/10.1016/0272-

    7358(89)90007-X

Stiles, W. B., & Shapiro, D. A. (1994). Disabuse of the drug metaphor: psychotherapy process-

    outcome correlations. *Journal of Consulting and Clinical Psychology*, *62*(5), 942-948.

    https://doi.org/10.1037/0022-006X.62.5.942

Stirman, S. W., Kimberly, J., Cook, N., Calloway, A., Castro, F., & Charns, M. (2012). The

    sustainability of new programs and innovations: a review of the empirical literature and

    recommendations for future research. *Implementation Science*, *7*(1), 17.

    https://doi.org/10.1186/1748-5908-7-17

Stirman, S. W., Marques, L., Creed, T. A., Gutner, C. A., DeRubeis, R., Barnett, P. G., Kuhn, E.,

    Suvak, M., Owen, J., Vogt, D., Jo, B., Schoenwald, S., Johnson, C., Mallard, K.,

Beristianos, M., & La Bash, H. (2018). Leveraging routine clinical materials and mobile technology to assess CBT fidelity: the Innovative Methods to Assess Psychotherapy Practices (imAPP) study. *Implementation Science*, *13*, 69. https://doi.org/10.1186/s13012-018-0756-3

[26]Sundell, K., Hansson, K., Löfholm, C. A., Olsson, T., Gustle, L. H., & Kadesjö, C. (2008). The transportability of multisystemic therapy to Sweden: short-term results from a randomized trial of conduct-disordered youths. *Journal of Family Psychology*, *22*(4), 550-560. https://doi.org/10.1037/a0012790

Tanana, M., Hallgren, K. A., Imel, Z. E., Atkins, D. C., & Srikumar, V. (2016). A comparison of natural language processing methods for automated coding of motivational interviewing. *Journal of Substance Abuse Treatment*, *65*, 43-50. https://doi.org/10.1016/j.jsat.2016.01.006

Tanner-Smith, E. E., & Tipton, E. (2013). Robust variance estimation with dependent effect sizes: practical considerations including a software tutorial in Stata and SPSS. *Research Synthesis Methods*, *5*(1), 13-30. https://doi.org/10.1002/jrsm.1091

Tate, R. F. (1954). Correlation between a discrete and a continuous variable. Point-biserial correlation. *The Annals of Mathematical Statistics*, *25*(3), 603-607. https://doi.org/10.1214/aoms/1177728730

[27]The Multisite Violence Prevention Project. (2014). Implementation and process effects on prevention outcomes for middle school students. *Journal of Clinical Child & Adolescent Psychology*, *43*(3), 473-485. https://doi.org/10.1080/15374416.2013.814540

Tipton, E. (2015). Small sample adjustments for robust variance estimation with meta-regression. *Psychological Methods, 20*(3), 375-393. https://doi.org/10.1037/met0000011

Tipton, E., Pustejovsky J. E., & Ahmadi H. (2019a). A history of meta-regression: Technical, conceptual, and practical developments between 1974 and 2018. *Research Synthesis Methods*, *10*(2), 161-179. https://doi.org/10.1002/jrsm.1338.

Tipton, E., Pustejovsky, J. E., & Ahmadi, H. (2019b). Current practices in meta-regression in psychology, education, and medicine. *Research Synthesis Methods*, *10*(2), 180-194. https://doi.org/10.1002/jrsm.1339

United States Census Bureau. (2020). *Try out our new way to explore data.* Census.gov. https://www.census.gov/data.html

Valentine, J. C., Pigott, T. D., & Rothstein, H. R. (2010). How many studies do you need? A primer on statistical power for meta-analysis. *Journal of Educational and Behavioral Statistics*, *35*(2), 215-247. https://doi.org/10.3102/1076998609346961

Viechtbauer, W. (2010). Conducting meta-analyses in R with the *metafor* package. *Journal of Statistical Software*, *36*(3), 1-48. https://doi.org/10.18637/jss.v036.i03

Wallace, B. C., Small, K., Brodley, C. E., Lau, J., & Trikalinos, T. A. (2012). Deploying an interactive machine learning system in an evidence-based practice center: abstrackr. In *Proc. of the ACM International Health Informatics Symposium (IHI)*, p.819-824.

Waltz, J., Addis, M. E., Koerner, K., & Jacobson, N. S. (1993). Testing the integrity of a

    psychotherapy protocol: assessment of adherence and competence. *Journal of Consulting*

    and Clinical Psychology, *61*(4), 620-630. https://doi.org/10.1037//0022-006x.61.4.620

Webb, C. A., DeRubeis, R. J., & Barber, J. P. (2010). Therapist adherence/competence and

    treatment outcome: A meta-analytic review. *Journal of Consulting and Clinical*

    Psychology,78(2), 200-211. https://doi.org/10.1037/a0018912

Webb, C. A., DeRubeis, R. J., Dimidjian, S., Hollon, S. D., Amsterdam, J. D., & Shelton, R. C.

    (2012). Predictors of patient cognitive therapy skills and symptom change in two

    randomized clinical trials: the role of therapist adherence and the therapeutic

    alliance. *Journal of Consulting and Clinical Psychology*, *80*(3), 373-381.

    https://doi.org/10.1037/a0027663

Weck, F., Richtberg, S., Jakob, M., Neng, J. M., & Höfling, V. (2015). Therapist competence

    and therapeutic alliance are important in the treatment of health anxiety

    (hypochondriasis). *Psychiatry Research*, *228*(1), 53-58.

    https://doi.org/10.1016/j.psychres.2015.03.042

Weisz, J. R., Kuppens, S., Ng, M. Y., Eckshtain, D., Ugueto, A. M., Vaughn-Coaxum, R.,

    Jensen-Doss, A., Hawley, K. M., Krumholz Marchette, L. S., Chu, B. C., Weersing, V. R.,

    & Fordwood, S. R. (2017). What five decades of research tells us about the effects of youth

    psychological therapy: A multilevel meta-analysis and implications for science and

    practice. *American Psychologist, 72*(2), 79–117. https://doi.org/10.1037/a0040360

Westra, H. A., Constantino, M. J., Arkowitz, H., & Dozois, D. J. A. (2011). Therapist differences in cognitive–behavioral psychotherapy for generalized anxiety disorder: A pilot study. *Psychotherapy*, *48*(3), 283-292. https://doi.org/10.1037/a0022011

Wickstrom, K. (1995). *A study of the relationship among teacher, process, and outcome variables with school-based consultation.* [Doctoral dissertation, Louisiana State University]. LSU Historical Dissertations and Theses.

Wickstrom, K. F., Jones, K. M., LaFleur, L. H., & Witt, J. C. (1998). An analysis of treatment integrity in school-based behavioral consultation. *School Psychology Quarterly, 13*(2), 141 - 154. https://doi.org/10.1037/h0088978

Wilson, D. B. (n.d.). *Practical Meta-Analysis Effect Size Calculator*. Campbell Collaboration. https://campbellcollaboration.org/research-resources/effect-size-calculator.html

Xiao, B., Imel, Z. E., Georgiou, P. G., Atkins, D. C., & Narayanan, S. S. (2015). "Rate My Therapist": Automated Detection of Empathy in Drug and Alcohol Counseling via Speech and Language Processing. *PloS ONE*, *10*(12). https://doi.org/10.1371/journal.pone.0143055

Figure 1. PRISMA Flow Diagram
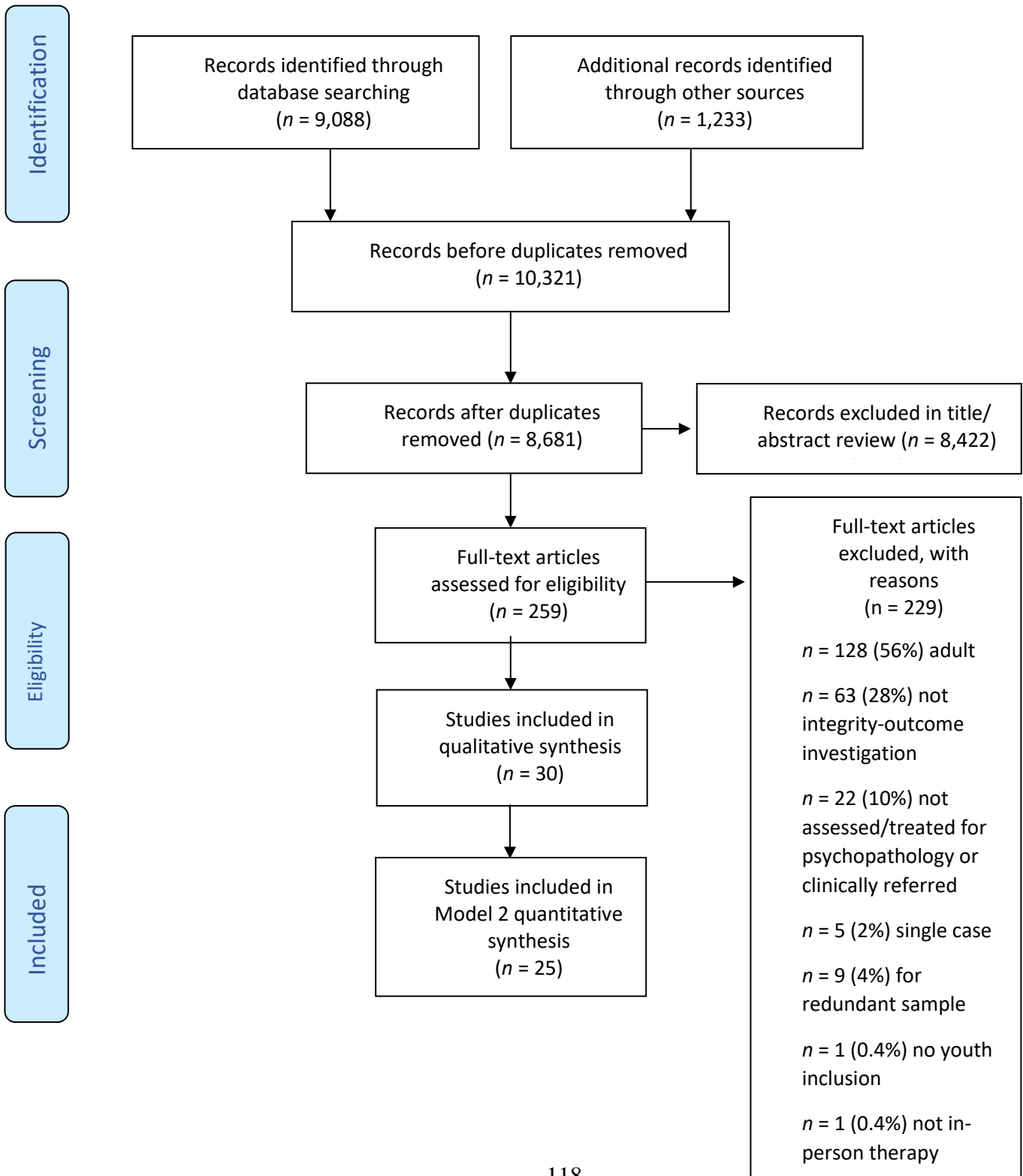
Table 1

A Summary of Included Studies (N = 30)

Study number	Year	Journal	Problem area	Intervention(s)	Sample sizes	ITIPS total(s)
1	2018	BMC Health Service Research	Trauma	Trauma-focused CBT	$N = 281$	59 (AA)
2	2018	Clinical Psychology & Psychotherapy	Anxiety	Individual CBT; Group CBT	$n = 91$; $n = 88$	79;79 (A)
3	2013	Evaluation and Program Planning	Conduct/Behavior problems	Early Risers Conduct Problems Prevention Program	$N = 262$	64(AA)
4	2018	Behavior Therapy	Attention problems	Planning My Life; Solution-focused Treatment	$n = 30$; $n = 31$	71;71 (A)
5	2018	Behavior Therapy	Conduct/Behavior problems	Group CBT	$N = 119$	73 (A)
6	2009	Substance Abuse Treatment	Substance use	Adolescent Community Reinforcement Approach	$N = 399$	49 (AA)
7	2017	Journal of Substance Abuse Treatment	Substance use/Conduct/Behavior problems	Multisystemic Therapy	$N = 40$	61 (AA)
8	2006	Journal of Abnormal Child Psychology	Depression	Penn Resiliency Program	$N = 271$	54 (AA)
9	2012	Child Care Youth Forum	Anxiety	CBT	$N = 32$	68 (A)
10	2014	Journal of Family Therapy	Conduct/Behavior problems	Functional Family Therapy	$N = 118$	49 (AA)

11	2016	Family Process	Conduct/Behavior problems	Functional Family Therapy	$N = 42$	48 (AA)
12	1999	Mental Health Services Research	Conduct/Behavior problems	Multisystemic Therapy	$N = 118$	55 (AA)
13	2008	JCCAP	Substance use	Multidimensional Family Therapy; Individual CBT	$n = 74$; $n = 62$	76;76 (A)
14	2011	JCCAP	Substance use	Multisystemic Therapy	$N = 41$	58 (AA)
15	2000	JCCP	Conduct/Behavior problems	Multisystemic Therapy	$N = 118$	55 (AA)
16	2017	Child & Youth Care Forum	Conduct/Behavior problems	Multisystemic Therapy	$N = 4,290$	51 (AA)
17	2019	JCCAP	Conduct/Behavior problems	Multisystemic Therapy	$N = 848$	53 (AA)
18	2010	Behavior Therapy	Anxiety	Group CBT, Individual CBT	$n = 23$; $n = 33$	58 (AA)
19	2014	JCCAP	Conduct/Behavior problems	Multisystemic Therapy	$N = 973$	55 (AA)
20	2013	Professional Psychology: Research and Practice	Anxiety	CBT	$N = 279$	71 (A)
21	2013	Journal of Substance Abuse Treatment	Substance Use	Multidimensional Family Therapy (2 samples)	$N = 212$ $N = 171$	61;61 (AA)
22	2003	Mental Health Services Research	Conduct/Behavior problems	Multisystemic Therapy	$N = 666$	51 (AA)
23	2009	JCCP	Conduct/Behavior problems	Multisystemic Therapy	$N = 1,979$	57 (AA)
24	2004	JCCAP	Conduct/Behavior problems	Multisystemic Therapy	$N = 452$	53 (AA)

25	2009	Journal of Studies on Alcohol and Drugs	Substance Use	Strengths-Oriented Family Therapy	$N = 54$	56 (AA)
26	2008	Journal of Family Psychology	Conduct/Behavior Problems	Multisystemic Therapy	$N = 156$	55 (AA)
27	2014	JCCAP	Conduct/Behavior Problems	Multisite Violence Prevention Program selective intervention	$N = 334$	54 (AA)
28	2004	Psychotherapy	Substance use	Multidimensional Family Therapy; Individual CBT	$n = 25$; $n = 26$	59 (AA)
29	2008	Dissertation	Substance use	Multidimensional Family Therapy; Individual CBT	$n = 65$; $n = 54$	80 (A)
30	2011	JCCAP	Substance use	Brief Strategic Family Therapy	$N = 480$	69 (A)

Note. Study number refers to and corresponds with the identical superscript in the reference section.

AA = Approaching Adequacy. A = Adequate. JCCAP = Journal of Clinical Child & Adolescent Psychology. JCCP = Journal of Consulting and Clinical Psychology.

Table 2

Methodological Characteristics of the Included Studies (N = 30)

Variable								κ	ICC
	North America only		Outside of North America		Inside/Outside North America				
Study region	19 (63.3%)		10 (33.3%)		1 (3.3%)			1	-
	Mean % male		Range % male		Range % female				
% male	66.9%		22.9 – 86.0 %		14 – 53			-	1
	M age (SD)		Range of M age (years)		Range of SD age				
M age (SD)	14.06 (2.43)		6.92 – 16.30		.93 – 3.85			-	.13
	Mean % White (SD)								
Race	49.48% (29.7)							-	NV
	Recruited for study		Clinically-referred/ treatment seeking		Involuntary		Mixture		
Recruitment	8 (26.7%)		13 (43.3%)		5 (16.7%)		4 (13.3%)	1	-
	Criminal justice	University outpatient	Non-university outpatient	School	Home	Other or multiple	Unknown		
Treatment setting	1 (3.3%)	2 (6.7%)	12 (40%)	3(10%)	7(23.3%)	3 (10%)	2 (6.9%)	.59	-
	Randomly assigned		Not randomly assigned		Unknown				

Client assignment	20 (66.7%)		7 (23.3%)		3 (10%)		1	-

Assignment method	Simple randomization	Block	Stratified	Covariate adaptive	Unclear but randomized	Unknown/not randomized		
	5 (16.7%)	3 (10%)	1 (3.3%)	9 (30%)	1 (3.3%)	11 (37.9%)	1	-

Payment	Subjects received incentives	Subjects did not receive incentives	Unknown	
	8 (26.7%)	12 (40%)	10 (33.3%)	.11

Target problem	Oppositional/ conduct problems	Depression (mood) problems	Anxiety problems	Trauma or stressor-related	Substance/ alcohol use	Multiple problems		
	14 (46.7%)	1 (3.3%)	4 (13.3%)	1 (3.3%)	8 (26.7%)	2 (6.7%)	1	-

Target problem diagnosis?	No diagnosis	Partial diagnostic criteria	Full diagnostic criteria	Unknown		
	14 (46.7%)	1 (3.3%)	13 (43.3%)	2 (6.7%)	1	-

Target problem diagnosis reliable?	Unreliable	Reliable	Unknown		
	4 (13.3%)	11 (36.7%)	15 (50%)	1	-

Study focus measure development?	Yes	No	
	1 (3.3%)	29 (96.7%)	1

Note. κ = Cohen's Kappa. ICC = Intraclass correlation. NV = No Value.

Table 3

Treatment-level Descriptors of Included Studies (N = 37)

Variable					κ	ICC
	Inadequate	Approaching adequate		Adequate		
ITIPS sum score	0 (0%)	26 (68%)		12 (32%)	-	-
	Cognitive-behavioral	Multi-system		Client-centered		
Treatment type	14 (37.8%)	22 (59.5%)		1 (2.7%)	.77	-
	Individual	Group	Family	Multi-system		
Treatment format	10 (27%)	4 (10.8%)	3 (8.1%)	20 (54.1%)	1	-
	Yes	No				
Significant individual contact (>25%)?	37 (100%)	0 (0%)			1	-
	Yes	No				
Significant parent component (>25%)	23 (62%)	14 (38%)			1	-
	Yes	No or unknown				
Significant family component (>25%)?	20 (67%)	17 (33%)			NV	-
	M age (years)	SD age				
Therapist age	41.19	5.08			-	NV
	% without majority trainee providers					
Majority trainee therapists?	100%				-	1

Note. κ = Cohen's Kappa. ICC = Intraclass correlation

Table 4

Independent Variable-level Characteristics of Included Studies (N = 53)

Variable					κ	ICC
	Adherence only	Competence only	Combined components - no definition	Combined components - definition		
Treatment integrity conceptualization	38 (71.1%)	12 (22.6%)	1 (1.9%)	1 (1.9%)	.8	-
	Unknown/not reported	Adherence to prescribed interventions	Adherence to principles/goals of therapy	Other		
Adherence conceptualization	1 (1.9%)	18 (34%)	18 (34%)	3 (5.7%)	.81	-
	Unknown/not reported	Frequency	Presence/ Absence	Extensiveness	Other	
Adherence scoring	8 (15.1%)	1 (1.9%)	2 (3.8%)	9 (17%)	18 (34%)	.5	-
	Not reported/unknown		Technical or "domain-limited" competence			
Competence conceptualization	2 (3.8%)		10 (18.9%)		1	-
	Client report	Other report	Direct observation	Indirect observation		
Collection method	6 (11.3%)	13 (24.7%)	26 (49.1%)	7 (13.2%)	.80	-
	Therapist	Supervisor/ trainer	Client	Caregiver- or other-	Independent observer	
Reporter	6 (11.3%)	6 (11.3%)	3 (5.7%)	10 (18.9%)	25 (47.2%)	1	-
	Yes		No			
Naïve reporter?	3 (5.7%)		26 (49.1%)		1	-

Homegrown treatment integrity measure?	Yes		No		
	19 (35.8%)		32 (60.4%)		1	-
Treatment integrity assessment timing	Once	Multiple times	Session-by-session		
	2 (3.8%)	34 (64.2%)	13 (24.5%)		.80	-

Note. κ = Cohen's Kappa. ICC = Intraclass correlation.

Table 5

Dependent Variable-level Characteristics of Included Studies (N = 95)

Variable							κ	ICC
	Symptoms/diagnosis	Functioning	Consumer perspectives	Environments	Systems	Other		
Outcome conceptualization	54 (56.8%)	25 (26.3%)	2 (2.1%)	11 (11.6%)	1 (1.1%)	1 (1.1%)	.90	-
	Not one clear problem area		Not matched		Matched			
Outcome matched to problem area?	1 (1.1%)		17 (17.9%)		77 (81.1%)		1	-
	Therapist	Supervisor	Client	Caregiver	Independent observer	Combined		
Outcome reporter	4 (4.2%)	1 (1.1%)	34 (35.8%)	35 (36.8%)	15 (15.8%)	4 (4.2%)	1	-
	Unclear		Not naïve		Naïve			
Reporter naïve to condition?	48 (50.5%)		43 (45.3%)		4 (4.2%)		.83	-
	Questionnaire	Observation/ independent evaluation	Objective data counts	Clinical interview	Unknown/ unclear			
Collection method	75 (78.9%)	2 (2.1%)	10 (10.5%)	4 (4.2%)	4 (4.2%)		1	-

Note. κ = Cohen's Kappa. ICC = Intraclass correlation.

Table 6

Correlational Confound Descriptions in Included Studies (N = 30)

Variable				κ	ICC
	Not measured *n* (%)	Measured/not applied *n* (%)	Measured and applied *n* (%)		
Alliance measured?	21 (72.4%)	2 (6.9%)	6 (20.7%)	1	-
	Not established *n* (%)	Partially established *n* (%)	Completely established *n* (%)		
Temporal precedence established?	1 (3.4%)	21 (72.4%)	7 (24.1%)	1	-
Other confounding variables identified	Baseline symptom severity, MST procedures, Child age (SD), Clinical experience, formal CBT training, minority status, session length (min), treatment format, therapist experience, therapist sex, language, years of team activity, implementation wave, treatment experience, treatment history, number of sessions, therapist education, month in treatment				

Note. κ = Cohen's Kappa. ICC = Intraclass correlation.

Vita

Ruben G. Martinez was born on April 21, 1989 in Muncie, Indiana. He graduated from Greenwood Community High School, Greenwood, Indiana in 2007. He attended Indiana University, Bloomington, where he graduated with a Bachelor of Arts in Psychology in May 2011. Subsequently, he worked for two years as a research assistant at the Cognitive Development Lab and a lab/project coordinator in the Training Research and Implementation in Psychology labs at Indiana University before entering the Clinical Psychology doctoral program at Virginia Commonwealth University in Richmond, Virginia, where he earned his Master's of Science in 2017. In 2019, Ruben matched as a clinical psychology intern at the Jane & Terry Semel Institute for Neuroscience and Human Behavior at the University of California, Los Angeles. He will complete his internship at the end of June 2020 and begin a postdoctoral fellowship (National Research Service Award T32) at the Semel Instite in summer 2020.