Assessment of the Therapeutic Alliance Scales: A Reliability and Validity Meta-Analytic Evaluation

Danielle L. Bouchard

Dissertation submitted to the

Faculty of Graduate and Postdoctoral Studies

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Clinical Psychology

School of Psychology

Faculty of Social Sciences

University of Ottawa

© Danielle Bouchard, Ottawa, Canada, 2018

Abstract

Extensive research has been conducted out on the construct of therapeutic alliance. With the growing emphasis on evidence-based practice in psychology it is vital that measures used in both clinical and research settings are empirically well-suited for the population under investigation. However, many measurement issues related to the reliability and validity of the alliance construct remain unaddressed or unresolved. Two studies were designed to add to the scientific evidence on the therapeutic alliance by establishing empirical evidence of the psychometric properties of this construct's most commonly used measures, with the intention of identifying the most psychometrically sound alliance measures. This was first done by systematically reviewing the literature to identify studies that used the most commonly used alliance measures. Next, key psychometric properties of each measure (internal reliability and predictive validity) were reviewed to determine if the alliance was assessed in the context of individual adult psychotherapy. In the first study, I conducted a reliability generalization analysis to (a) estimate the average reliability coefficient of each alliance measure identified in the systematic review and (b) examine the potential influence that study characteristics may have had on the reliability estimates. Six different alliance measures were included (Agnew Relationship Measure, California Psychotherapy Alliance Scales, Counselor Rating Form, Penn Alliance Scales, Therapeutic Bond Scale, Working Alliance Inventory), in various formats and rater versions, resulting in a total of 17 alliance measure variants for this first analysis. In the second study, I conducted a validity generalization analysis using only those studies from the first study that were identified as containing outcome data. The purpose of this study was to synthesize the alliance-outcome effect sizes that have been reported for

the most commonly used therapeutic alliance measures and to assess the potential impact study characteristics may have on those effect sizes. Five different alliance measures (California Psychotherapy Alliance Scales, Counselor Rating Form, Penn Alliance Scales, Therapeutic Bond Scale, Working Alliance Inventory, Vanderbilt Therapeutic Alliance Scale) in various formats and rater versions, resulting in a total of 15 alliance measure variants were included in this analysis. This second study was different from previous alliance-outcome meta-analyses as I only included studies that (a) could be identified as providing psychotherapy, as opposed to other mental health services, (b) assessed the alliance from individual adult psychotherapy, (c) were identified as using the most commonly used alliance measures, and (d) measured the alliance at the midpoint of treatment, or earlier. This second study also differed from previous meta-analyses as I conducted separate analyses for correlational data and partial correlational data. The reliability generalization study found that majority of the alliance measures were good choices for assessing the alliance based on their mean reliability coefficients. The validity generalization study found relatively no difference in the early alliance's ability to predict treatment outcomes in individual adult psychotherapy between full correlation data (r = .24) and partial correlation data (r = .23). There was also no difference found among the different alliance measures, or their variants, in their ability to predict treatment outcomes, suggesting that no one alliance measure is statistically better at predicting outcomes. The results from both studies suggest that, based on their overall level of reliability as well as their ability to predict treatment outcomes, both researcher and clinicians should consider these measures, with few exceptions, as comparably good choices for assessing the alliance in adult individual psychotherapy.

Statement of Co-Authorship

The two manuscripts included in this dissertation were prepared in collaboration with my dissertation supervisor, Dr. John Hunsley. I was the primary author and Dr. Hunsley was the secondary author for the first manuscript, entitled "Assessing Therapeutic Alliance Scales: A Reliability Generalization Meta-Analytic Evaluation," and the second manuscript entitled "Assessing Therapeutic Alliance Scales: Meta-Analytic Evaluation of Adult Individual Psychotherapy." As the primary author on both manuscripts, I was responsible for the conceptualization of the research questions and methods, planning and execution of the statistical analyses, interpretation of the results, and preparation of the manuscripts. Dr. Hunsley provided guidance and assistance on all aspects of this dissertation, especially in the research methods and editing of the manuscripts.

Acknowledgements

First and foremost, I would like to offer my sincere thanks to my supervisor, Dr. John Hunsley, for all the support, guidance, patience, and insight he has given me over the years. I cannot recall the countless times, John, you helped me see the forest as I got lost amongst the trees. I would especially like to express my gratitude for always making yourself available to provide direction and feedback regardless of your many other obligations. Lastly, thank you for all the support and encouragement you have given me as I struggled at times to find work-life balance between graduate school and raising a young family. In addition to my supervisor, I would like to thank my thesis committee members, Drs. Patrick Gaudreau, Sophie Lebel, and George Tasca for their valuable feedback and insight.

As for my friends, a very special thanks to Diane Laroche, who took time out of her tremendously busy schedule to help me double code my data. Your help on this project means more to me than you will ever know. Thank you to all those friends I have made along the way in the program, your endless support through all the accomplishments and hardships of graduate school have made finishing this degree possible. For all my friends external to the program, I want to thank you for constantly reminding me that life does exist outside of school and that I need to continue to live it.

To my family, I would first like to thank my parents for their constant support, encouragement, and believing in me. To my wonderful husband, Tony, who has sacrificed more than I could ever repay, who has seen me at my best and worst, who has picked me up time and time again as I stumbled, your unwavering love and support continues to amaze me. I would not be the person I am today without you. Thank you. Last but certainly

not least in my heart, to my two boys, Isaac and Lucas, who have only known “Mommy” as a student, thank you for keeping me grounded and teaching me what is really important in life, I cannot wait to start this new chapter in our lives.

Table of Contents

Abstract ..... **ii**

Statement of Co-Authorship ..... **iv**

Acknowledgements ..... **v**

List of Tables ..... **ix**

List of Figures ..... **x**

CHAPTER 1: Assessment of the Theraputic Alliance Scales: A Reliability and Validity Meta-Analytic Evaluation **1**

An Overview of the Dissertation **5**

Historical Background and Early Conceptualization of the Therapeutic Alliance 6

The Therapeutic Alliance within Different Treatment Approaches 9

Psychodynamic therapies 9

Cognitive-behavioural therapies (CBT) 10

Humanistic therapies 11

Summary 13

Measurement of Therapeutic Alliance **14**

Psychometric Properties of Clinical Research Measures 21

Reliability and Reliability Generalization **23**

Reliability 23

Reliability Generalization 25

Validity and Validity Generalization **27**

Validity 27

Validity Generalization 30

Summary 30

CHAPTER 2: Assessing Therapeutic Alliance Scales: A Reliability Generalization Meta-Analytic Evaluation **32**

Method **38**

Literature Search 38

Inclusion/Exclusion Criteria 39

Selection of Studies 42

Coding Sample Characteristics and Potential Moderators 44

Results **47**

Q statistic and I^2 index 49

Analysis of Moderator Variables 50

Publication Bias 54

Orwin's fail-safe N 54
Funnel Plots 54

Discussion **55**
The Effects of Sample Characteristics on Reliability Estimates 57
Limited Reporting of Reliability 60
Limitations 60
Conclusions and Recommendations 62

References **64**

Appendix A **101**

Appendix B **104**

CHAPTER 3: Assessing Therapeutic Alliance Scales: Meta-Analytic Evaluation of Adult Individual Psychotherapy **111**
Measurement Considerations 113
Study Inclusion Considerations 116
The Present Meta-Analysis 119

Method **120**
Literature Search 120
Inclusion/Exclusion Criteria 122
Selection of Studies 124
Coding Sample Characteristics and Potential Moderators 126

Results **129**
Q statistic and I^2 index 130
Analysis of Moderator Variables 130
Publication Bias 132
Rosenthal's fail-safe N 132
Funnel Plots 133

Discussion **134**
Limitations 136
Conclusions 138

References **140**

Appendix A **161**

Appendix B **164**

Appendix C **169**

Conclusion **173**

References **182**

List of Tables

Table 1 Six Core Alliance Measures 15

Table 2 Criteria for assessing psychometric adequacy of a measure: Select Categories 22

Table 3 List of Measure-related Search Terms 91

Table 4 Descriptive Statistics for the Mean Reliability Coefficient 93

Table 5 Simple Meta-Regression for Mean Reliability Coefficients and Sample Characteristics for Significant Continuous Moderators 94

Table 6 Analysis of Variance between Mean Reliability Coefficients and Sample Characteristics 96

Table 7 List of Measure-related Search Terms 156

Table 8 Descriptive Statistics for the Mean Alliance-Outcome Coefficients 158

List of Figures

Figure 1 Flow Chart of Study Selection 97
Figure 2 Funnel Plot for CALPAS-24-C 98
Figure 3 Funnel Plot for WAI-T 98
Figure 4 Funnel Plot for WAI-S-T 99
Figure 5 Funnel Plot for WAI-C 99
Figure 6 Funnel Plot for WAI-S-C 100
Figure 7 Funnel Plot for WAI-SR-C 100
Figure 8 Flow Chart of Studies Selection 159
Figure 9 Funnel Plot Publication Bias (full correlation) 160
Figure 10 Funnel Plot Publication Bias (partial correlation) 160

CHAPTER 1: Assessment of the Therapeutic Alliance Scales: A Reliability and Validity Meta-Analytic Evaluation

Danielle L. Bouchard

University of Ottawa

Assessment of the Therapeutic Alliance Scales: A Reliability and Validity Meta-Analytic Evaluation

The view that the therapeutic alliance is a vital component of the therapeutic process that is linked to treatment outcomes is a longstanding belief among psychotherapists. This important psychological construct, which is also known as working alliance, helping alliance, alliance, or therapeutic bond, is one of the more extensively studied variables in the psychotherapy literature and spans multiple client populations (e.g., child, adult, couples). According to Horvath, Del Re, Fluckiger, and Symonds (2011), a 2009 search using the above terms resulted in the identification of over 7,000 articles, chapters, and books. The growing number of empirical reviews and meta-analyses conducted on the therapeutic alliance also demonstrate the considerable amount of research that has been carried out on this psychological construct, (for examples see, Elvins & Green, 2010; Fluckiger, Del Re, Wampole, Symonds, & Horvath, 2011; Horvath & Symonds, 1991; Horvath et al., 2011; Martin, Garske, & Davis, 2000; Sharf, Primavera, & Diener, 2010).

Although the concept of the therapeutic alliance has existed since the early 20th century, it was not a key topic of interest within the research literature until the 1970s and early 1980s. Its rise in popularity is due, in part, to a number of high profile studies that found comparable effects across different treatment orientations (Luborsky, Singer, & Luborsky, 1975; Stiles, Shapiro, & Elliot, 1986). These findings led many researchers to search for common underlying mechanisms that could account for the observed changes within their clients (Horvath et al., 2011). The therapeutic alliance was one of those underlying mechanisms that seemed to cut across different therapeutic approaches. Since

these early days, there has been considerable research conducted on this psychological construct.

One of the more extensively studied areas within the therapeutic alliance is the relation between alliance and treatment outcome. To date, there have been hundreds of studies and numerous meta-analyses conducted on the alliance-outcome relation. Of these many meta-analyses, five have focused on therapy results with either adults or adults and older adolescents, with four focused on the alliance and general treatment outcomes and one focused on the relation between alliance and treatment dropout. These five meta-analyses all found comparable summary effect sizes (Outcomes in general: Horvath & Symonds, 1991, r = .26; Martin et al., 2000, r = .22; Horvath & Bedi, 2002, r = .21; Horvath et al., 2011, r = .275. Dropout: Sharf et al., 2010, d = .55 [approximately r = .27]). Moreover, the data indicated that the therapeutic alliance is a consistent predictor of outcomes across therapeutic approaches.

In recent years the scope of issues examined in alliance research has expanded. For example, a number of studies have looked at how the alliance develops over time. By analyzing the fluctuation in the alliance over the course of treatment, researchers have been able to identify a number of developmental patterns: stable alliance, linear alliance, and quadratic or "u-shape," and "v-shaped" alliance (Ardito & Rabellino, 2011). The last two patterns reflect the ruptures and repairs that can occur in the alliance. Here, the ruptures refer to quick declines in the alliance, whereas, the repairs are a resolution of the problem(s) that was responsible for the rupture (Safran & Muran, 2000). In addition to developmental trajectories, researchers have also considered the possibility of a "halo effect," that is, the strength of association between alliance and outcome may be inflated

when both the alliance and outcome are rated by the client. However, Horvath and colleagues (2011) found that there the difference between independent and same source effect sizes was statistically nonsignificant. Lastly, efforts have been made to specifically train therapists to enhance the alliance, with an alliance-fostering training model and treatment manual having been developed and empirically tested (Crits-Christoph, Crits-Christoph, & Gibbons, 2010). Results from this study were encouraging, with alliance scores increasing over the course of treatment. The results of these lines of alliance research are important as they help to inform training and practice in psychotherapy.

With the growing emphasis on evidence-based practice (EBP) within the field of professional psychology, clinicians are being strongly encouraged to use scientific data to inform and direct their treatment practices (American Psychological Association Presidential Task Force on Evidence Based Practice, 2006). This push for EBP has also been observed within the area of the therapeutic alliance literature. In 2009, the American Psychological Association Divisions of Psychotherapy and Clinical Psychology commissioned a task force to update their review of evidence-based therapy relationships, with the purpose of identifying, operationalizing, and disseminating empirically supported data on the therapeutic relationship (Norcross, 2011). From this empirical evidence, in their resolution on psychotherapy effectiveness, APA has come to acknowledge the therapeutic alliance being a key component of the effectiveness of psychological treatment as well as endorsed the alliance as being a bond between therapist and client that involves agreement on tasks and goals (APA, 2013).

One way to ensure the quality of EBP is through the use of evidence-based assessment (EBA), for the precision of the EBP is only as good as the assessment

foundation that it is built upon (Hunsley & Mash, 2007). An essential ingredient to EBA is the use of psychometrically strong measures within one's research and treatment practices (Hunsley & Mash, 2007). This requires ensuring that the psychometric properties of the measures that are being used within a study, or in clinical practice, are appropriate for the population and purpose under investigation. In other words, it is critical that the measures in question have been shown to yield scores that are reliable and valid for purpose for which they are being used. In the case of alliance research that would mean ensuring that therapeutic alliance measures are psychometrically suitable for the sample and purpose being assessed.

An Overview of the Dissertation

Given that the therapeutic alliance plays an important role in therapeutic process and the general trend towards using evidence-based practices in psychology, this dissertation will add to the literature by providing updated empirical evidence regarding the psychometric properties of the most commonly used measures of the therapeutic alliance. The dissertation studies are based on the results of a systematic literature search of the therapeutic alliance measures identified in Elvins and Green's empirical review (2008), and will include articles and dissertations published up until January 2017. Research identified in this review form the data set for the first study, in which a meta-analytic procedure known as reliability generalization (RG) was used in order to calculate the mean reliability estimates of the measures across studies. Using relevant studies from this literature search, the second study examined the predictive validity of these measures using another meta-analytic method known as validity generalization (VG). In both studies,

moderator analyses were conducted to determine whether selected study characteristics exerted a statistically significant influence on these mean reliability and validity estimates.

In the following sections, I provide an overview of the history and initial conceptualization of the therapeutic alliance construct. Next, I briefly review the therapeutic alliance within the context of the different treatment approaches. After this, I address how the therapeutic alliance is measured and discuss a number of measurement issues concerning this construct. From there, I focus on the psychometric properties of clinical research measures in general and then, more specifically, discuss the reliability and reliability generalization of these measures. Next, I discuss the validity and validity generalization of clinical research measures.

Historical Background and Early Conceptualization of the Therapeutic Alliance

The concept of a therapeutic relationship and its importance to the therapeutic process can be traced back the writings of Freud. In his writings on transference he made note of working with his clients collaboratively in the effort to help them face and work through their pain (Freud, 1971). This concept of collaboration, also referred to as "unobjectionable or positive transference" (Freud, 1971, p. 105), arose from Freud questioning the irony of having some patients continue with treatment once their defenses were activated when, logically, he felt they should leave treatment. Transference was a process in which a bond developed between therapist and client, allowing them to work together towards eliminating the client's symptoms and was seen by Freud to be a key aspect of change (Elvins & Green, 2008; Freud, 1971). Other psychoanalysts, such as Sterba (1934), Zetzel (1956), and Greenson (1965, 1967), have elaborated further on this concept.

Many researchers have recognized Zetzel (1956) as being the theorist who formally introduced the term *therapeutic alliance* (Horvath, et al., 2011; Hatcher, 2010; Horvath & Bedi, 2002). However, credit is also given to Bibring for using the term alliance in her 1956 article, where she put forth the idea that a strong therapeutic alliance was essential in order for the analysis to be effective for the client (Safran & Muran, 2000; Zetzel, 1956). In Greenson's (1965, 1967) discussion on transference, he used the term "working alliance," and made the distinction between working alliance and therapeutic alliance. He wrote that the working alliance referred to the client's ability to do the work necessary for therapy, whereas the therapeutic alliance referred strictly to the relational bond that develops between the therapist and client.

Around the same time, formulations of the therapeutic alliance began to emerge from outside the psychoanalytic tradition. Certain core aspects of alliance, such as unconditional positive self-regard, the emphasis on empathic identification, and the communication of empathic understanding by the therapist appeared in the works of Rogers (Elvins & Green, 2008; Rogers, 1965). In Elvins and Green's (2008) review, they noted that Anderson (1962) was the first to operationalize the concepts of empathy and therapeutic rapport. Orlinsky and Howard (1975) elaborated further on these concepts and begin to subject them to empirical testing. Incorporating their results into previous theory, they put forth a three dimensional understanding of the therapeutic alliance: working alliance (a collaboration between therapist and client working together), empathic resonance, and mutual affirmation (which is very similar to the concept of unconditional positive self-regard) (Howard & Orlinsky, 1972; Orlinsky & Howard, 1975).

Bordin (1979) was another psychotherapy researcher who reconceptualized the alliance in a manner distinct from its psychodynamic roots. Using Sterba's (1934), Zetzel's (1956), and Greenson's (1965) work as a starting point, Bordin departed from starting with the concept of transference and, instead, emphasized how therapists and clients actually work together (Bordin, 1979; Hatcher, 2010; Krause et al., 2011; Orlinsky & Rønnestad, 2000). Focusing more on this collaborative relationship, he reworked the Orlinsky and Howard's three dimensional understanding of alliance into his own three-component working alliance theory: 1) agreement on *goals*, 2) collaboration on *tasks* that make up the therapy, and 3) establishment/development of a *bond* between client and therapist (Bordin, 1979). Bordin (1979) argued that the working alliance between the client and therapist applied not only to psychoanalysis, but to all forms of therapeutic approaches, and was the key to change for the client. He also suggested that different treatment approaches would focus, or place more emphasis on, different aspects of the alliance. He referred to this as embedded alliances.

In parallel to Bordin, Luborsky (1976) was also moving the concept of alliance away from its psychoanalytic roots. Like Bordin, he believed that alliance was present in all treatment approaches (Krause et al., 2011). However, his conceptualization of the alliance concept was somewhat different from Bordin's approach. Elaborating and extending on Zetzel's (1956) and Stone's (1961) work, Luborsky proposed that the alliance occurring between the client and therapist happened in two distinct phases: Type I and Type II (Horvath et al., 2011). Type I alliance develops in the early stages of treatment and represents a helping relationship. This helping relationship depends on the client experiencing the therapist as being supportive and helpful, and as valuing and respecting

the client. Type II alliance represents a working collaboration between the therapist and client. Here, the client and therapist work together in a joint effort towards overcoming the client's problems (Luborsky, 1994).

The Therapeutic Alliance within Different Treatment Approaches

Although the therapeutic alliance is present in all treatment approaches, there are a number of discernable differences with respect to how the alliance is conceptualized. Considering three broad treatment approaches (i.e., psychodynamic, cognitive-behavioural, and humanistic therapies), Muran and Barber (2010) noted that each one considered the alliance to be an important treatment variable and that the difference among approaches lies within which aspects of alliance are emphasized and utilized in the therapeutic process.

Psychodynamic therapies. As previously mentioned, the concept of alliance originated from within this particular treatment orientation. Since the initial conceptualization rooted in the transferential aspects of therapy, the concept of alliance has grown and expanded. Early conceptualizations described the alliance as occurring between the analyst's interpretations and the client's ability and willingness to take in these interpretations and reflect on them (Messer & Wolizky, 2010). There was an expectation for both participants to remain flexible and to be able to switch back and forth between reflection of experience and actual experiencing. This was seen as necessary in order for change to occur (Messer & Wolizky, 2010). The biggest conceptual change with respect to the function of the alliance was the heightened emphasis on its curative aspects. These therapeutic properties, in and of themselves, are now seen as mechanism of change, rather than only a prerequisite for insight to occur.

As Messer and Wolizky (2010) explained, most current psychodynamic writers do not dismiss the value of insight, but instead now see the benefits made by insight as being on par with the benefits gained by having a "good enough" relationship with a therapist. A "good enough" therapist has been described as one who is more caring and understanding than the client's parents were. It is from working with a "good enough" therapist that the client will be receptive to gaining insight from the transference, as well as experience a new good enough relationship, which is therapeutic in its own right. According to some researchers, there is evidence that supports the view that a more positive relationship has therapeutic benefits in its own right (Wallerstein, 1986; Zuroff & Blatt, 2006) and that, therefore, insight may not be the only therapeutic action involved in psychodynamic therapy (Messer & Wolitzky, 2010).

Cognitive-behavioural therapies (CBT). CBT therapists have a long history of recognizing that the therapeutic relationship is a significant contributor to the process of change. However, it is only relatively recent that this construct has received sizeable attention as an operationalized research variable (Castonguay, Constantino, McAleavey, & Goldfried, 2010). Like other treatment orientations, the alliance is seen as an important treatment variable, however, there are some clear differences in the theoretical conception of the alliance in the psychodynamic and CBT orientations. It has been argued that the primary distinction between CBT and other approaches is that CBT places more emphasis on collaboration and teamwork (Raue & Goldfried, 1994; Castonguay et al., 2010). This client-therapist collaboration is known as *collaborative empiricism* (Dattilio & Hanna, 2012). The concept of collaborative empiricism entails that the therapist and client work together, as a team, to identify the client's problems as well as to identify solutions to those

problems. This is accomplished through collaboratively devising a treatment plan and exploring together, through the process of discovery and experimentation, the aspects of the clients' functioning that contribute to their difficulties.

CBT-oriented writers have traditionally placed more emphasis on specific techniques than on the interpersonal sensitivity needed to help engage and move the client through treatment (Wright & Davis, 1994). This emphasis on techniques can also be seen within the concept of collaborative empiricism. It is through collaborative empiricism that the client will develop a number of behavioural and cognitive strategies that will help them challenge the thoughts and situations that are causing them distress (Dattilio & Hanna, 2012). Another reason empirical collaboration is important is that it guides the therapist's case conceptualization. This helps to minimize the assumptions therapists have to make concerning clients, to lead to a better understanding of the clients' beliefs, and to strengthen the therapeutic relationship (Dattilio & Hanna, 2012).

A further distinction between CBT and other approaches is the way the alliance is conceptualized in its role in the process of change. CBT theorists recognize the alliance as a factor that enhances the adherence to specific techniques, but suggest that it is not the mechanism by which change occurs. In other words, the relationship fosters an environment that encourages change and supports clients' motivation to change, so that clients will engage in the treatment techniques that are needed to produce change (Beck, Rush, Shaw & Emery, 1979; Castonguay et al., 2010; Dattilio & Hanna, 2012).

Humanistic therapies. For the humanistic approaches (e.g., Gestalt, emotion-focused therapy) the therapeutic relationship is seen as collaborative and central to treatment (Elliot, Watson, Goldman & Greenberg, 2004; Rogers, 1951; Watson &

Kalogerakos, 2010). It has been noted that, without collaboration, there is no alliance (Berdondini, Elliot & Shearer, 2012). In these approaches the alliance has been defined as involving an agreement between client and therapist on goals, as well as the bond that develops between the dyad (Watson & Kalogerokos 2010). However, for humanistic approaches, an effective therapeutic relationship is conceptualized primarily as an emotional connection that occurs between the client and therapist; this connection is considered to be the most important aspect of the alliance and is viewed as necessary for the client to change (Berdondini et al., 2012; Watson & Kalogerakos, 2010).

This view of the alliance originated from Rogers' theory that pathology arises from an inconsistency between a client's internal experience of their real self and their internalized feelings of self-worth learned from early caretakers. It is this conflict that Rogers believed led to anxiety and a tendency for clients to deny both the conflicting feelings and their own experience. He suggested that creating a warm, accepting, empathic environment and encouraging a positive self-regard for the client was needed to help clients improve their feelings of worth and self-acceptance. More specifically, he hypothesized that it is the relationship that provides clients a safe and non-threatening environment in which they can explore their inner world and make the interpersonal and intrapersonal changes needed to function adaptively: it through this relational process that the client is believed to develop greater awareness, responsibility, and self-actualization (Watson & Kalogerakos, 2010).

In these treatment approaches, the therapist balances guidance and acceptance while making process observations and suggestions to the client that will help them make shifts in their experiencing in the session. The therapist's role is to help clients become

aware of all of their experiences and help them work out ways of expressing and balancing needs and feelings in manner that they fully own and accept. Here, the therapist distinguishes between the working conditions and relationship conditions in therapy (Watson & Greenberg, 1994). The former refers to the collaborative aspect of the alliance, which reflects working together on agreed upon goals and therapeutic tasks. The latter refers to the bond of the alliance and reflects the connection the client and therapist have with each other. Emphasis is placed on being empathic, authentic, and attuned to the client's experience in order to provide a safe environment to facilitate client changes (Berdondini et al., 2012).

Summary. Even though specific aspects of alliance are emphasized and weighted differently among the treatment orientations just described, there is evidence that supports Bordin's assumption that the therapeutic alliance is an important variable within the process of change across all forms of psychotherapy. In the most recent alliance-outcome meta-analysis, contrasting the average effect sizes of four different treatment approaches (i.e., CBT, r= .35, Interpersonal Therapy (IPT*),* r = .39, Psychodynamic*,* r = 29, and Substance Abuse treatments*,* r = .23), Horvath et al. (2011) found the strength of the association between alliance and outcome of each treatment to be highly statistically significant. However, the contrast analysis between the treatment categories was not statistically significant, indicating that there was no statistically significant difference in terms of the impact of the alliance on treatment outcome across the various treatment approaches.

Measurement of Therapeutic Alliance

Considering the importance and the wealth of research conducted on the therapeutic alliance, two important questions arise: How is the construct defined and how is it measured? More specifically, how has the concept of the therapeutic alliance been operationalized so that it can be studied? In the more recent literature, the therapeutic alliance has been broadly defined as a collective and affective bond that develops between a client and therapist (Martin et al., 2000). More precisely, it has been described as consisting of interpersonal processes, or interactions, that occur within the relationship of the client and therapist, that are independent of the type of treatment provided or theoretical orientation of the therapist (Elvins & Green, 2010). Although it has received incredible amounts of attention, this psychological construct still remains vaguely defined within much of the current literature. There has been a recent attempt to demystify this complex construct by clarifying its underlying dimensions and how these dimensions are represented in various measures (Krause et al., 2011). However, the vast majority of the existing alliance research is based on a loosely defined variable, which may be contributing to measurement error among the different alliance measures (Ardito & Rabellino, 2011; Elvins & Green, 2008; Krause et al., 2011).

Even though these modern conceptualizations of the therapeutic alliance share a number of common features, such as being pan-theoretical in nature and having an emphasis on collaboration and consensus between the therapist and client, they often fail to clearly define the alliance construct (Horvath et al., 2011). This has afforded researchers and clinicians the liberty to integrate this loosely defined concept into theoretically different models of the therapeutic process (Horvath et al., 2011). Alliance measures have

therefore been developed and tested based on a range of different versions of the alliance concept, which has led to a number of measurement problems within the alliance literature (Elvins & Green, 2008; Horvath et al., 2011).

One of the key measurement problems is the proliferation of alliance measures that occurred mostly between 1978 and 1986. Based on Horvath and colleagues' (2011) review of the therapeutic alliance literature, there are over 38 adult alliance measures that have been used to assess this construct. Out of this considerable number of measures, there are six that have been most commonly used and were recommended as core measures of the construct by Martin and Elvins (Martin et al., 2000; Elvins & Green, 2008) (see Table 1). These six measures are: 1) Working Alliance Inventory (WAI, Horvath & Greenberg, 1989), 2) the California Psychotherapy Alliance Scales (CALPAS, Gaston & Ring, 1992; Marmar, Weiss & Gaston, 1989), 3) the Pennsylvania (Penn) scales, (Helping Alliance Questionnaire - HAq-II, Luborsky, 1976), 4) the Vanderbilt scales (Vanderbilt Psychotherapy Process Scale and Vanderbilt Therapeutic Alliance Scale, (VPPS/VTAS), Hartley & Strupp, 1983), 5) the Toronto Scales (TARS, Marziali, 1981) and 6) the Therapeutic Bond Scales (TBS, Saunders, Howard & Orlinsky, 1989).

Table 1

Six Core Alliance Measures

Measure	Date developed	Conceptual background	Rating system	Sources of alliance rater and different versions of measure	Psychometric properties as reported in Elvins & Green (2008) and Ardito & Rabellino (2011)
California scales (CALPAS) (Gaston & Marmar, 1994; Marmar et al.,	1986	Primarily from a psychodynamic framework. The CALPAS is comprised of the	The CALPAS consists of 24-items rated on a seven-point scale. Clinical observer	Therapists, clients, clinical observers	Results from factor analysis have shown support for the four aspects of the

1989a; Marmar, Weiss & Gaston, 1989b)		California Therapeutic Alliance Rating Scale (CALTARS) and the California Psychotherapy Alliance Scales (CALPAS). The CALPAS is a revised version of the CALTARS and was developed to rate four aspects of alliance outlined by Gaston: commitment, therapist understanding and involvement, patient-therapist agreement on goals, and strategies.	version consists of 30 and has 4 subscales	A short version of the client measure exists	alliance. Internal consistencies reported for older scales that are highly similar to the CALPAS are acceptable (0.94-0.76) and 0.84 for the whole client scale.
Pennsylvania (Penn) alliance scales (Helping alliance questionnaire – HAq-II) (Alexander & Luborsky, 1986; Luborsky, 1976; Luborsky et al., 1985; Luborsky et al., 1983)	1976	From a psychodynamic perspective, these scales empirically test Luborsky's Type 1 and Type 2 alliance. These scales consist of Helping Alliance Rating Counting Signs Method (HAcs), Helping Alliance Rating Method (HAr), and Helping Alliance Questionnaire Method (HAq).	A score on a 5 (HAcs), a 10 (HAr), and a 6 (HAq) point rating scale (e.g., 1 = very low, 5= very high) is assigned to a series of items grouped into sub-scales based on the type of alliance being measured.	Therapist (HAr), Clients (HAq), and clinical observers (HAcs). The HAq is the widest used version of this measure. The HAq also has a shorten version (HAq-II) and has been translated into French	High levels of internal consistency for the therapist and observer scale have been reported (0.93).
Therapeutic bond scales (TBS) (Saunders et al., 1989)	1989	Based on Orlinsky and Howard's psychotherapeutic model and is based on the Therapy Session report (TSR). The TBS contains a Global Bond Scale that consists of 3 subscales measuring 3 dimensions of alliance	This 50 items measure is comprised of the 3 dimensions: working alliance scale, 15 items, empathic resonance scale, 17 items, and mutual affirmation scale, 18 items. Items are rated on a 21-point rating system.	Client and clinical observer	Internal reliabilities of each scale as well as the global scale have been reported as ranging from 0.62-0.82.
Toronto scales (TARS) (Marziali, 1984; Marziali, 1981)	1981	These scales are based on a psychodynamic conceptualization of the alliance as well as Bordin's integrative model. It mainly focuses on the affective aspects of alliance. Marziali and colleagues combined items from the VPPS, VTAS and the HAcs to develop the Therapeutic Alliance Rating Scale (TARS).	All three versions consist of 42 items rated on a 6-point scale.	Client, therapist, and clinical observer	Has shown to have adequate internal consistencies
Vanderbilt therapeutic scales (Hartley & Strupp, 1983; O'Malley et al,	1978	From adult psychodynamic therapy, influenced by Orlinsky and Howard conceptualization	The VPPS measures a segment of therapy and is rated on a 80 item 5-point scale.	Clinical observer	Factor analysis demonstrated that the VTAS and the VPPS has similar

1983; Suh et al., 1986)		of alliance. The Vanderbilt Psychotherapy Process Scale (VPPS) measures the relationship between client and therapist as well as the therapeutic process. The Vanderbilt Therapeutic Alliance Scale (VTAS) was modified from the VPPS to more specifically measure the alliance.	The VTAS is a 44-item 6-point likert observer scale designed to rate tapes of treatment sessions.		factor structures. The VPPS has been shown to have 0.96-0.82 levels of internal consistency and the VTAS has demonstrated solid internal consistency.
Working alliance inventory (WAI) Horvath & Greenberg (1986, 1989)	1986	The WAI was designed to measure the three dimensions of Bordin's working alliance in adults across all forms of psychotherapy.	The WAI is a 36-item 7-point Likert type scale. The shorten version is a 12-item 7-point scale. Many studies have adopted the shorten version of the WAI.	Client, therapist, clinical observer There is a shorter version of the WAI. It has been translated in a number of different languages. A couples version have also been recently developed (Symonds & Horvath, 2004)	Internal consistency for the WAI-p been reported to be 0.93 for the full scale and ranges from 0.85 to 0.88 for the subscales. The WAI-o's internal consistency has been reported as 0.98 and the WAI-s' range from (0.85-0.93).

Note. Adapted from "Therapeutic Alliance and Outcome of Psychotherapy: Historical excurses, Measurements, and Prospects for research" by R. B. Ardito and D. Rabellino, 2011, *Frontiers in Psychology for Clinical Settings, 2,* p. 3-6, Copyright 2011 Ardito and Rabellino and "The Conceptualization and Measurement of Therapeutic Alliance: An Empirical Review, by R. Elvins and J. Green, 2008, *Clinical Psychology Review, 28,* p. 1170-1178, Copyright Elsevier B.V. 2008.

Although research has demonstrated that there are some common underlying themes within these measures, the measures are not based on a common definition of the alliance construct (Hatcher & Barends, 1996; Hatcher, 2010). In addition, no single measure contains representative items for the different components of the alliance concept that have emerged over time. The measures vary on the number of items they contain and which dimensions of the alliance construct they address. Furthermore, the underlying factor structure, or how the items group together to measure a specific dimension, does not

necessarily represent the conceptual subscales included as part of the development of the instruments.

Examples of these measurement issues have been reported in a number of studies (e.g., Elvins & Green, 2008; Krause et al., 2011; Hatcher & Barends, 1996; Horvath & Bedi, 2002). For instance, Elvins and Green (2008) highlighted the point that the WAI bond scale, which has its conceptual roots embedded in Bordin's working alliance theory, does a good job of capturing the therapist's contribution to the personal relationship. However, the WAI places less emphasis on the client's capacity and motivation to engage in therapeutic work. These conclusions were based on Hatcher and Barends' (1996) factor analytic research, where the factors were not found to capture Bordin's central theme of active purposive mutual work. It has also been suggested that, in completing the WAI, clients may not be differentiating between tasks and goals, whereas therapist are likely to consider the distinction between these two subscales (Horvath & Bedi, 2002). Elvins and Green (2008) also noted that the WAI bond scale highly correlates with the Barrett-Lennard Relationship Inventory (BLRI: Barrett-Lennard, 1986) a measure of empathy. Another example provided by Elvins and Green (2008) in their review is the Vanderbilt scales, which were influenced by Orlinsky and Howard's conceptualization of the alliance. With this measure, they noted that the scales place more emphasis on the client and the dyadic components of alliance, compared to aspects of the therapist's involvement. These scales also fail to include any items that capture the purposive, mutual work aspect of the alliance.

Notable differences across the measures, as well as between the different informant versions (e.g., client versus therapist) of the same measure, have been reported. For example, although the WAI does not emphasize the client's capacities, these capacities are

directly measured in the CALPAS (Horvath & Bedi, 2002). Negative transference, or contributions, is measured in the TARS, CALPAS, and VTAS, but not in the other measures (Gaston & Marmar, 1994; Hovarth & Bedi, 2002). The CALPAS and the HAq-II focus primarily on the therapists' clinical impressions of the client, whereas the WAI places more weight on questions about how the dyad functions. In other words, compared to these two measures, the WAI puts more emphasis on capturing the agreement or consensus between the therapist and client (Krause et al., 2011; Horvath & Bedi, 2002). Item sampling biases were also found for three of the core measures (WAI, CALPAS and HAq-II) based on cognitive, affective, and behavioural dimensions used by Krause et al., (2011) in their review of alliance measures. That is, these instruments were found to differ on the number of items that pertained to these three domains. They found a slight sampling bias in the WAI for items referring to the affective dimension of the alliance in the therapist's version and a slight sampling bias for the cognitive dimension in the client's version. Biases were observed in the CALPAS as well, with a slight bias towards the behavioural dimension in the therapist's version and a slight bias for the cognitive in the client's version.

Results from Horvath et al. (2011) meta-analysis provided additional empirical support for the existence of differences between the various alliance measures. In their moderator analysis of the alliance-outcome relation, comparing four core alliance measures (i.e., CALPAS, HAq, VPPS, and WAI) and a group of "other" alliance measures, they found homogeneity of effects only in the data for the CALPAS and VPPS. In other words, for the HAq, the WAI, and the "other" alliance measures, the different versions of the same measure (i.e., different rater versions, short forms) did not yield the same results in predicting treatment outcome. The authors concluded that these findings might be the

result of CALPAS and VPPS having fewer versions than the WAI or HAq. However, a more detailed analysis is clearly needed to determine if any substantive measurement differences exist within the various forms of the WAI and HAq, and if there are, what these differences might be.

Horvath et al. (2011) found another potential measurement issue when they examined the interactions between the type of measure used in alliance-outcome studies and the alliance rater. Here, they observed that the interaction between these two variables accounted for 23% of the variance within the combined outcome variable. Although these results could not be generalized to the entire alliance-outcome literature, as the studies reporting data from multiple raters accounted for only 25% of the entire sample of studies, they concluded that these results suggest that these findings are likely partially due to the variety of ways that alliance is measured. Where the actual differences lie within this interaction has yet to be determined. Further analysis is needed to examine the effects between a specific alliance measure and a specific treatment outcome (i.e., treatment outcomes in general versus a focus on specific outcomes such as premature termination).

As intriguing as these findings are, it is important to recognize that there are aspects of Horvath et al.'s (2011) work that may have influenced these moderation analyses by increasing the variability of effect sizes included in the meta-analysis. For example, although the focus of the meta-analysis was on adult psychotherapy, a number of the studies included in the analysis included adolescent clients. In addition, a number of studies included in the meta-analysis evaluated the alliance-outcome relation for mental health services that were not psychotherapy. For example, when the alliance-outcome relation was assessed within the context of a mental health team, the alliance was assessed

with other treating clinicians, such as a physiotherapist, rather than the treating psychologist. This underlines the need to attend closely to study inclusion criteria when examining the literature on therapeutic alliance in adult psychotherapy, as the alliance-outcome relation may differ by developmental stage of client or by type of treatment provided.

Psychometric Properties of Clinical Research Measures

Having an evidence-based assessment tool is essential to practicing the evidence-base standards set out by APA. Without the use of a scientifically supported measure, the research, and ultimately treatment practices, resulting from the data will be fallible. All too frequently many of the assessment measures being used by psychologists have not been supported by sufficient scientific evidence (Hunsley & Mash, 2008). There seems to be a common misconception among researchers that psychometric properties derived from test development studies, or other studies, are immutable and, therefore are sufficient to establish psychometric applicability of the measure for any population or purpose (Vacha-Haase, Henson, & Caruso, 2002).

In order for an assessment measure to be considered as evidence-based it needs to demonstrate solid scientific evidence of its psychometric suitability. This is established through ascertaining empirical evidence that illustrates that the measure yields scores that are reliable and valid for the specific sample and/or conditions for which it is being used. However, as the assessment literature is so vast in scope, there is no predetermined number of studies that can establish the psychometric properties for an instrument (Hunsley & Mash, 2008). This problem leaves us with the question of what constitutes "good enough" psychometric criteria (Hunsley & Mash, 2007).

In an effort to address this concern, Hunsley and Mash (2007, 2008) proposed a rating system consisting of nine psychometric categories to help determine an instrument's psychometric adequacy. These nine categories are: norms, internal consistency, inter-rater reliability, test-retest reliability, content validity, construct validity, validity generalization, sensitivity to treatment change, and clinical utility. A rating of less than adequate, adequate, good, excellent, unavailable, and not applicable is assigned to each of these categories. What constitutes an adequate, good, or excellent can vary from category to category. Generally, though, a rating of (a) adequate indicates that a measure meets a minimal level of scientific rigor, (b) good indicates that a measure possesses solid scientific support, and (c) excellent indicates there is extensive, high quality supporting evidence for the measure. In contrast, a rating of less than adequate indicates the measure does not meet the minimum level set out in the criteria. Within a clinical context, Hunsley and Mash (2008) recommended using only those measures that have meet, at a minimum, the criteria of good. As the focus of this dissertation is on internal consistency and predictive validity, only the categories of internal consistency and validity generalization will be used to assess the results of the two meta-analyses. The decision to solely focus on internal consistency and predictive validity was made as, first, alpha is the most commonly reported reliability estimate, and, second, the predictive validity of the alliance is the most clinically relevant with respect to treatment outcomes. The criteria for these two categories are summarized in the table below: (See Table 2).

Table 2

Criteria for assessing psychometric adequacy of a measure: Select categories

Internal consistency:
Adequate: Preponderance of evidence indicates alpha values of .70-.79.

Good: Preponderance of evidence indicates alpha values of .80-.89.
Excellent: Preponderance of evidence indicates alpha values of ≥ .90.
Validity generalization:
Adequate: Some evidence supports the use of this instrument with either (a) more than one specific group (based on sociodemographic characteristics such as age, gender and ethnicity) or (b) in multiple settings (e.g., home, school, primary care setting, inpatient setting).
Good: Preponderance of evidence supports the use of this instrument with either (a) more than one specific group (based on sociodemographic characteristics such as age, gender and ethnicity) or (b) in multiple settings (e.g., home, school, primary care setting, inpatient setting).
Excellent: Preponderance of evidence supports the use of this instrument with more than one specific group (based on sociodemographic characteristics such as age, gender and ethnicity) and across multiple settings (e.g., home, school, primary care setting, inpatient setting).

Note. From "Developing Criteria for Evidence-Based Assessment: An Introduction to Assessment That Work," by J. Hunsley and E. J. Mash, (Eds.), 2008, *A Guide to Assessment That Works,* p. 8-9. Copyright 2008 by Oxford University Press

Reliability and Reliability Generalization

Reliability

Reliability is an essential psychometric property in psychological measurement that needs to be taken into consideration when evaluating the utility of a measure (Geisinger, 2013; Hunsley & Mash, 2008). Reliability refers to the consistency, or stability, of the observed scores on a specific measure, and is inversely related to the amount of measurement error associated with a score of an instrument (Coaley, 2010; DeVon et al., 2007; Henchy, 2013; Miller, 2010). More simply, high reliability means low measurement error, and low reliability means high measurement error. This measurement error stems from two sources, systematic error and random error. Systematic errors are errors that consistently affect all of the individuals' or group of individuals' scores, such as poorly

worded test items; in contrast, random errors are errors that impact different individuals' score in different ways (Tyron & Bernstein, 2003).

In classical test theory (CTT), any observed test score is the sum of two independent additive components: true score and error score (Miller, 2010; Nimon et al., 2012). The error score within CTT refers to random error only, as it is assumed that systematic errors can be controlled for (Miller, 2010; Tyron & Bernstein, 2003, p. 30). As measures are not without their flaws, and therefore, to some degree, inaccurate, the true score and the reliability of that score can never be determined. As a consequence, it is the reliability of the observed score that is calculated and is more accurately referred to as a reliability estimate (Dimitrov, 2002; Henchy, 2013). Reliability estimates range from 0.00, indicating an absence of reliability, to 1.00, indicating perfect reliability. In order for a measure to obtain perfect reliability, every inter-item correlation of that measure would need to have a value of 1.00.

There are a of number of ways to categorize and estimate reliability, the selection of which one to use depends on the type of measure used (Cook, 2006; Downing, 2004; Miller, 2010). Internal consistency is used to assess the consistency of the items within the measure, that is, the homogeneity of the items. Test-retest reliability is used to assess the stability of the measure over time, and inter-rater reliability is used to assess the consistency between different raters.

These reliability estimates help to locate the source of the error within the scores obtained on a measure, as they represent the proportion of the variance in the test scores that can be attributed to the true score (Miller, 2010). According to Hunsley and Mash (2008), estimate values for internal consistency that range from 0.70 to 0.79 should be

considered to indicate adequate reliability, 0.80 to 0.89 to indicate good reliability, and any value over 0.89 to indicate excellent reliability. These suggestions are for coefficient alpha values, which is the most commonly used statistic to estimate reliability.

It is important to take into account a measure's reliability coefficient when determining effects within a study. The reliability of scores on a measure can either attenuate or accentuate the magnitude of the effects that have been obtained from that measure (Henson, 2001; Kieffer, 2011; Reinhardt, 1996). Given that the reliability of the scores on a measure can influence the statistical results of a study, it is imperative that these coefficients be reported so that the reader may independently interpret the reported results. It is for these reasons that American Psychological Association (APA) publication standards indicate that reliability values should be reported for data generated with all the measures used with each sample in a study regardless of whether the focus of the research is on psychometrics (APA, 2008; Wilkinson, 1999).

Reliability Generalization

Unfortunately, there are far too many published studies that do not provide reliability estimates for their samples, as most researchers fail to report the reliability estimates for their own data (Vacha-Haase, 1998; Vacha-Haase & Thompson, 2011; Yin & Fin, 2000). Most often, researchers appear to be under the misperception that once a measure's reliability has been established, the reliability estimate becomes a property of the instrument itself. As a result, they use the measure's initial reliability estimate, or estimates from similar samples, as evidence that their measure is reliable. Vacha-Haase, Kogan, and Thompson (2000), refer to this erroneous process as *reliability induction.* To illustrate the extent of this problem, in a recent review of the literature Vacha-Haase and

Thompson (2011) found that in 12, 994 primary studies, 54.6% of the researchers did not even mention reliability, let alone report reliability estimates. Moreover, 15% of the 12,994 primary studies that did mention reliability estimates did so by simply inducting previously reported reliability values only when it applied to their own data. Out of these induction cases, 48% referred to only the test manual and not to other studies where the same measure was used. Although the average of the mean coefficient alpha values reported by Vacha-Haase and Thompson was .80 (*SD* = .09), the average range for these estimates was .45 to .95, with the smallest mean alpha reported being .17. Bearing in mind that many of these estimates are based on induction from the test manual, these lower values seem to suggest that relying on reliability estimates from the original study of a measure is likely to overestimate the reliability value obtained in subsequent studies. Given that reliability is not a property of measure itself but, rather, a function of the scores obtained on a specific measure by a particular group under particular circumstances, reliability should be examined with every study (Coaley, 2010; Henchy, 2013; Vacha-Haase, 1998).

It is possible, however, to develop an estimate of the range of likely reliability estimates associated with an instrument, across samples and contexts. With a sufficient number and variety of studies reporting reliability values for scores on a measure, this range should provide a reasonable approximation of what reliability estimates are likely to be obtained in future studies if they are conducted with samples comparable to those used in prior research. Vacha-Haase (1998) introduced a meta-analytic method called *reliability generalization* (RG) to examine this variability of reliability of scores across studies. This method is modelled after Schmidt and Hunter's (1977) validity generalization method that

uses the means, standard deviations, and other descriptive statistics to compute validity coefficients across studies. RG analyses permit the examination of the effects of specific study characteristics on reliability estimates across studies. In other words, these analyses can provide a mean score reliability for sample-specific variables for a particular measure, thus giving researchers a baseline to consider when deciding to use a measure (Graham et al., 2011; Henson, 2001).

RG can also identify (a) the upper and lower bounds (i.e., a confidence interval) of score reliability estimates produced by a measure and (b) the study and sample characteristics that contribute to the variability of the reliability estimates across studies (Kieffer, 2011). The information that a RG provides gives researchers a more robust starting point when choosing a measure compared to relying on a single reliability estimate (Graham et al., 2011). As Graham et al. (2011) stated, "RG studies will answer the question, 'how reliable are the scores produced by the measure across different sample and study characteristics?'" (p. 42).

Validity and Validity Generalization

Validity

Validity is another important psychometric property that needs to be taken into consideration when evaluating the utility of a measure. Validity generally refers to the degree to which scores on an instrument accurately reflect the construct it was designed to measure (DeVon et al., 2007; Hunsley & Mash, 2008). More specifically, an instrument is said to be valid when the construct it is measuring actually exists and changes or variations on different attributes of that construct within the individual produce changes or variations

on the corresponding attributes within the measure (Borsboom, Mellenbergh, & van Heerden, 2004; Haynes, Smith & Hunsley, 2011).

There are three main types of measurement validity that have been identified: (1) construct validity, (2) content validity, and (3) criterion-related validity (Nunnally, 1978; Tyron & Bernstein, 2003). Construct validity refers to the broader conceptualization of validity, that is, the degree to which an instrument reflects the phenomenon it purports to measure. Content validity refers to the extent the items on a measure fully and accurately represents all aspects of the construct it was designed to assess. Criterion-related validity refers to the extent the measure relates to some criterion variable that is relevant to the construct (Hunsley & Lee, 2010).

Predictive validity is a specific type of criterion-related validity that involves the degree to which the scores on a measure predict the criterion-related variable at some future point in time (Tyron & Bernstein, 2003). For example, how well the scores of the WAI obtained during treatment predict treatment outcomes or unilateral termination (i.e., dropout) provides predictive validity data for the WAI. Treatment outcome and unilateral termination are important criterion variables because of how they are hypothesized to relate to the alliance's role within the therapeutic process. The alliance relates to outcomes through the contribution it makes to the changes in functioning within the client. It is these changes that ultimately lead to the outcome of treatment. The association between alliance and unilateral termination functions in a similar manner as was the case for outcomes, with studies showing that a weaker alliance was associated with higher dropout rates (Sharf, Primavera, & Diener, 2010; Swift & Greenberg, 2012).

For predictive validity, the degree to which validity has been established is usually determined by a correlation or regression coefficient (i.e., *r* statistic) obtained between the measure and the relevant criterion (Tyron & Bernstein, 2003). These *r*-values range from +1 to – 1, with 1 indicating a total positive correlation, 0 indicating no correlation, and -1 indicating a total negative correlation. Although Cohen's (1992) benchmarks have been a standard for interpreting *r*-values (i.e., .10 small effect, .30 moderate effect, and .50 large), it has been argued that slightly different values may be more suitable for interpreting effects in psychology (i.e., less than .20 is a small effect, between .20 and .30 is a moderate effect, and greater than .30 is a large effect; Hemphill, 2003). In addition, although having statistical benchmarks for interpreting the magnitude of a correlation is helpful, these empirical guidelines are still somewhat artificial and do not tell us the whole story. That is, on their own, they are missing valuable contextual information, such as the purpose of the study or the study characteristics from which the values were derived, which helps determine their clinical importance (Hunsley & Westmacott, 2007). Therefore, caution should always be taken when interpreting the meaning and importance of the *r*-value.

Unlike the reliability of scores on a measure, which can exist in the absence of the scores on the measure being valid, the validity of the scores on the measure is dependent on the reliability of scores on the measure (Graziano & Raulin, 2004). That is, reliability is a prerequisite for validity. Like reliability, validity is not an innate characteristic of the measurement itself. One cannot state that a measure is simply valid or invalid (Suen, 1990). Rather, this psychometric property is context-sensitive and varies across different sample characteristics or populations (Graziano & Raulin, 2004; Hunsley & Mash, 2008). In other words, establishing validity for a measure for a specific purpose or population does not

inherently mean that this measure will be valid for a different population or purpose. The issue of validity then becomes one of interpretation and the ways in which a measure's scores are used (Suen, 1990; Furr & Bacharack, 2008). When selecting a measure, it is important that researchers base these interpretations and uses on empirical evidence that supports the measure being utilized as an appropriate instrument for their specific population and purpose.

Validity Generalization

Validity generalization (VG), a meta-analytic method developed by Schmidt and Hunter (1977), is used to help assess the extent to which validity coefficients from scores on a measure can be generalized across different sample characteristics. This meta-analytic method uses the means, standard deviations, correlation coefficients, and other descriptive statistics to compute a VG coefficient across different studies that used the same measure. As in RG, VG analyzes other study characteristics (e.g., sample size, sample characteristics) to help explain the varying validity coefficients that were observed across studies (Vacha-Haase, 1998). By calculating a mean *r*-value and the corresponding 95% confidence interval, it is possible to obtain a best estimate of what the validity estimate will likely be in subsequent studies. In other words, we can determine how generalizable the validity coefficients are across the different samples and studies.

Summary

Having evidence-based assessment tools are essential to practicing in an evidence-based manner. Therefore, it is important to ensure that the measures used to assess the construct under study are psychometrically sound and appropriate for their intended use. Although the therapeutic alliance has been extensively researched, many measurement

issues with the reliability and validity of this construct's multitude of measures remain unaddressed. For these reasons, establishing strong empirical evidence for the use of specific alliance measures with a specific population or specific purpose is very important.

This dissertation aims to add to the field of evidence-based practice by first systematically identifying studies that used the most commonly used alliance measures. Next, key psychometric properties of each measure (internal reliability and predictive validity) will be reviewed in order to evaluate if the alliance was assessed in the context of individual adult psychotherapy. It is imperative that only studies that deal with an adult population and that are of psychotherapy (not other mental health services) are included in this dissertation so that conclusions of the alliance can be drawn for this specific population for this specific purpose. The first study, a reliability generalization analysis, examines the internal reliability of the most commonly used alliance measures identified in the systematic review. The purpose of this study is to obtain an average reliability estimate for each of the alliance measures as well as to examine the potential influence study characteristics, such as age or type of mental disorder treated, may have on the reliability estimates. The second study, a validity generalization analysis, uses only those studies from the first study that were identified as containing outcome data. The purpose of this second study is to synthesize the alliance-outcome effect sizes that have been reported for the most commonly used therapeutic alliance measures and to assess the potential impact study characteristics may have on those effect sizes. This second study will differ from the previous alliance-outcome meta-analyses, as it will address the number of selection criteria issues outlined in previous sections.

CHAPTER 2: Assessing Therapeutic Alliance Scales: A Reliability Generalization Meta-Analytic Evaluation

Assessing Therapeutic Alliance Scales: A Reliability Generalization Meta-Analytic Evaluation

The therapeutic alliance has long been seen by both clinicians and researchers as an essential component of the therapeutic process. This important psychological construct is one of the more extensively studied variables in the psychotherapy literature and spans multiple forms of intervention (Elvins & Green, 2008; Fluckiger, Del Re, Wampole, Symonds, & Horvath, 2012; Horvath & Symonds, 1991; Horvath, Del Re, Fluckiger, & Symonds, 2011; Martin, Garske, & Davis, 2000; Sharf, Primavera, & Diener, 2010). Although, initially rooted in the psychoanalytic tradition, other formulations of the therapeutic alliance construct have emerged over time, with different theorists emphasizing different core aspects of the alliance (Horvath et, al., 2011; Safran & Muran; 2000). Certain alliance concepts, such as therapist empathy and therapeutic rapport, became operationalized and subjected to empirical testing (Elvins & Green, 2008). An emphasis on the how the therapist and client work together towards the client's therapeutic goals also emerged and has been the focus of considerable empirical evaluation (Bordin, 1979; Hatcher, 2010; Horvath et al., 2011; Krause, Altimir, & Horvath, 2011; Orlinsky & Rønnestad, 2000 Luborsky, 1976).

Although it has received extensive attention, the therapeutic alliance construct still remains vaguely defined within much of both the professional and research literature. Even though contemporary conceptualizations of the therapeutic alliance may share a number of common features, such as such as being pan-theoretical in nature and having an emphasis on collaboration and consensus between the therapist and client, many researchers often fail to clearly define the alliance construct they are studying (Horvath et al., 2011). This has

resulted in the development of alliance measures based on a range of different conceptualizations of the alliance construct—currently there are at least 30 different alliance measures that have been used in psychotherapy research studies. Despite substantial differences how alliance is conceptualized in these measures, they are frequently treated as being conceptually and functionally equivalent. This confusion about the precise nature of these alliance measures, along with considerable variability in the thoroughness of efforts taken to develop and validate an alliance measure, a number of measurement problems have arisen within the alliance literature (Ardito & Rabellino, 2011; Elvins & Green, 2008; Horvath et al., 2011; Krause et al., 2011). For example, the measures differ on the dimensions of the alliance construct that are included, the number of items used to operationalize an alliance dimension, and the extent to which factor analytic investigations provide empirical support for the subscales intended to represent the various alliance dimensions included in a measure (Hatcher & Barends, 1996; Hatcher, 2010). Additionally, notable inconsistencies between the different informant versions (e.g., client versus therapist) of the same measure have been reported (e.g., Gaston & Marmar, 1994; Horvath & Bedi, 2002; Krause et al., 2011).

Because of issues such as these it can be challenging for researchers to know which alliance measure to use for their research. To begin to provide a sense of the psychometric characteristics of these measures, in the present study I conduct meta-analyses in order to synthesize the reliability estimates that have been reported for the most commonly used therapeutic alliance measures in the adult psychotherapy literature. I will use meta-analytic procedures to determine (a) summary reliability values (both mean values and

confidence intervals) for the measures and (b) the potential moderating influence of sample and study characteristics on these reliability values.

Reliability refers to the consistency, or stability, of the observed scores of a specific measure and can be seen as the degree to which a test score is free from random measurement error (American Educational Research Association, American Psychological Association, National Council on Measurement in Education (AERA, APA, NCME), 2014 Geisinger, 2013; Miller, 2010). In classical test theory (CTT), the observed test score on a measure is comprised of two components: the true score and the error score (Miller, 2010). As measures are not without their flaws, and therefore, to some degree, inaccurate, the true score and reliability of that score can never be determined. As a consequence, it is the reliability of the observed score that is calculated and is more accurately referred to as a reliability estimate (Dimitrov, 2002; Henchy, 2013). (Dimitrov, 2002; Henchy, 2013). These reliability estimates can help to locate the source of the error within the scores obtained on a measure, as they represent the proportion of the variance in the test scores that can be attributed to the true score (Miller, 2010). In other words, reliability estimates demonstrate how consistent the obtained test scores are. Depending on the type of measure used, there are three ways to categorize and estimate reliability, internal consistency, test-retest, and inter-rater reliability. As the reliability of a measure's scores can impact the magnitude of the effects obtained with that measure, it is imperative that reliability estimates' are reported in a study (Henson, 2001; Kieffer & MacDonald, 2011; Reinhardt, 1996).

Unfortunately it is all too common for authors not to report reliability estimates for the specific samples used in their studies (Vacha-Haase, 1998; Vacha-Haase & Thompson, 2011; Yin & Fin, 2000). Most often, researchers appear to be under the misperception that once a measure's reliability has been reported for one sample, it will be the same for other samples. In other words, a reliability estimate is seen as a property of the instrument itself, rather than as an estimate derived from a specific sample. As a result, researchers use the initial reliability estimate reported in the development of a measure, or estimates from similar samples, as evidence that the measure itself is reliable. The erroneous process has been referred to as *reliability induction* (Vacha-Haase, Kogan& Thompson, 2000). However, reliability is a property of the scores obtained from the measure (Vacha-Haase, 1998; Henchy, 2013; Hunsley & Mash, 2008). Because reliability is not a property of the instrument itself, reliability can be influenced by a study's participants or setting, and therefore, can fluctuate from study to study (Thompson, 1999).

It is possible, however, to develop an estimate of the range of likely reliability estimates associated with an instrument, across samples and contexts. With a sufficient number and variety of studies reporting reliability values for scores on a measure, a mean reliability estimate, along with an associated 95% confidence interval, should provide a reasonable approximation of what reliability estimates are likely to be obtained in future studies if they are conducted with samples comparable to those used in prior research. Vacha-Haase (1998) introduced a meta-analytic method called reliability generalization (RG) to examine this variability of reliability of scores across studies. This method is modelled after Schmidt and Hunter's (1977) validity generalization method that uses the means, standard deviations, and other descriptive statistics to compute validity coefficients

across studies. RG analyses permit the examination of the effects of specific study characteristics on reliability estimates across studies. In other words, these analyses can provide a mean score reliability for sample-specific variables for a particular measure, thus giving researchers a baseline to consider when deciding to use a measure (Graham, Diebels, & Barnow, 2011; Henson, 2001). RG can also identify (a) the upper and lower bounds (i.e., a confidence interval) of score reliability estimates produced by a measure and (b) the study and sample characteristics that contribute to the variability of the reliability estimates across studies (Kieffer & MacDonald, 2011). The information that a RG provides gives researchers a more robust starting point when choosing a measure compared to relying on a single reliability estimate (Graham et al., 2011).

To date, only one RG study focusing on alliance measures has been published. Hanson, Curry, and Bandalos (2002) conducted RG analyses on the Working Alliance Inventory, a commonly used alliance measure. In this study, data were included for several measure variants, including the client and therapist versions of the Working Alliance Inventory (WAI) and the Working Alliance Inventory-Short (WAI-S), as well as the Working Alliance Inventory-Observer (WAI-O). The authors reported high reliability estimates for both the client and therapist's versions of the WAI (.93 and .91, respectively), for the both the client and therapist's versions of the WAI-S (.95 and .93 respectively), as well as for the WAI-O (.97). Although results from this RG suggested the WAI (client & therapist), the WAI-S (client & therapist), and the WAI-O scores are likely to be highly reliable, there are important limitations to the study that warrant the inclusion of the WAI in the current study. These limitations include: (a) conducting analyses with small sample sizes (most analyses included only 3-5 studies); (b) only including data from published studies (i.e., no

attempt was made to include grey literature such as dissertations); (c) failing to contact the authors of published studies to obtain reliability coefficients not reported in the publications; and (d) reporting only mean reliability values (i.e., no confidence interval values). Additionally, the most recent articles included in the Hanson et al. analyses were published in 2002, and many studies using variants of the WAI have been published since then.

Method

Literature Search

The search for studies using therapeutic alliance measures was conducted through searches of three electronic databases: PsycINFO (1806 – January 09, 2017; OVID platform), PubMed (1800 – January 09, 2017; Medline platform), and Dissertations and Theses (1861 – January 09, 2017; ProQuest platform) (see figure 1 for flowchart regarding study selection). As PsycINFO and PubMed are key databases for the field of psychology, the vast majority of the therapeutic alliance research would be accessible through these two databases. Therefore, SCOPUS, Web of Science, CINHAL, or Social Work abstracts were not included in the search. In addition, as Dissertations and Theses contain unpublished dissertations, no other grey literature was searched. In general, if results from a study were reported in both a dissertation and a published article, results from the published article were retained for the meta-analysis to ensure that the data are only used once. However, in the cases where the dissertation provided more usable data (e.g., sample characteristics data) than the published study, the dissertation was retained instead.

The terms used in the searches were based on Elvins and Green's (2008) review of the conceptualization and measurement of the therapeutic alliance. As their study was an

empirical review of the literature, no additional alliance measure was searched for, as these 33 measures were found to be the most used measures. The decision to include only the 33 measures identified as the most common, as opposed to less frequently used alliance measures, including any measure developed after Elvins and Green's (2008) review was made due to concerns about having adequate statistical power to be able to detect potential differences between measures. In addition, as the intent for the present study was to examine measures of alliance in the adult psychotherapy literature (i.e., clients/participants aged 18 years and older), 10 alliance measures designed to assess child, adolescence, or family populations were not included in the database searches. Therefore, the literature was searched using the 23 of the 33 measures identified. Search terms used the measure's full name, any or all the acronyms for the measure (if applicable), and the acronym(s) of the measure in conjunction with the term alliance (see Table 1 for a list of these search terms). Finally, the results of this search were compared to the studies containing one (or more) of the 23 alliance measures that were included in the Horvath et al. (2011) alliance meta-analysis. Studies included in that meta-analysis, but not identified with the current search strategy, were then reviewed for possible inclusion in the current study (see Appendix A for differences between studies used in the Horvath et al. (2011) meta-analysis and the current RG study). No further authors were contacted, or study citations reviewed, in the effort to identify additional studies.

Inclusion/Exclusion Criteria

The following were used as inclusion criteria for this RG study: (a) The study was clinical as opposed to analogue (i.e., the study had to have taken place within treatment conditions, not simply a simulation of psychotherapy). This exclusion criterion has been

labelled *Analogue*; (b) Alliance was measured in a psychotherapy context, as opposed to the alliance being measured in other service provision situations such as career counselling. This criterion has been labelled *Not Psychotherapy*; (c) The study was an empirical study and not a review or critique of the literature, and was not a book chapter or a book review. This criterion was labelled *Review/Book*; (d) The study had a minimum of 5 adult participants (this requirement was based on the inclusion criteria used in Horvath et al.'s, (2011) meta-analysis of the adult therapeutic alliance literature). As most of these studies consisted of case studies presented in the context of an in-depth examination of another topic, this criterion was included with the criterion labelled *Review/Book*; (e) The psychotherapy provided in the study had been conducted by a licensed service provider. Service providers who were in a graduate level professional training program were also included, as they were under the supervision of a licensed service provider. This criterion was included in order to ensure a certain level of competency on the part of those providing the treatment. Studies were excluded if it was not clear who provided the psychotherapy services or if it was a mixed sample of providers providing the psychotherapy (e.g., multidisciplinary team – not all providers licensed) and the reliability estimate was reported for the group of providers as whole. Studies in this category were included with those in the criterion of *Not Psychotherapy*; (f) The client must have direct contact with a therapist. Studies where psychotherapy was provided via a computer program were removed. This criterion was included as part of the criterion *Not Psychotherapy*; (g) The study contained an adult sample only (i.e., 18 years of age and older). However, exceptions were made if the study contained a small number of 17-year olds in the sample. This criterion was labelled *Inexact Treatment Context*; (h) The alliance

was measured within the context of individual therapy, rather than group, couple, or family therapy. In the event a study contained a mixed sample (e.g., couples and individuals, licensed versus non-licensed provider), the data were included if the authors reported separate reliability scores for those groups that met inclusion criteria (e.g., a reliability estimate for a licensed provider subgroup was provided). This criterion was included as part of the criterion *Inexact Treatment Context*; (i) The study was written in English or French. This criterion was labelled *Study Not Accessible* (j) A print or electronic version of the study report had to be accessible. Studies that were retrievable via the university library or through interlibrary loan were included. However, studies that could not be accessed by these means were excluded. This criterion was also included under the heading *Study Not Accessible;* (k) Descriptive information needed to evaluate the study's eligibility for the RG study was reported in the article (e.g., age of participants, type of treatment used). If this information was not clearly presented in the article, the study was excluded from further review. In the case where the authors reported that the descriptive information was recorded elsewhere (i.e., made reference to another article), this source was searched for the missing information. This criterion was labelled *Insufficient Information;* (l) The therapeutic alliance had to have been measured during therapy. Studies where the alliance was assessed only at intake, feedback, only at termination, or for a consultation were eliminated. This criterion was labelled *Incorrect Time of Assessment*; (m) The study had to have used the specific alliance measure that was entered into the literature search. The specific alliance measure also had to be in its complete form. Studies that reported using a modified version of the measure (e.g., items were removed or reworded) were eliminated. This criterion was labeled *Wrong Measure*; (n) As described

previously, in the case of a study that was described in both a dissertation and a published article, the dissertation was eliminated from the pool of studies unless it contained more complete information than the article, in which the article was removed. This criterion was labelled *Duplicates*; (o) Any duplicate publications were eliminated from the pool of studies. Included here were studies in which the data reported in a particular study was also reported in another article. This was done to help control for dependence of the data. The study that had the least amount of data pertinent to the RG analyses was removed. This criterion was labelled *Duplicates;* and (p) After the literature search was completed, and all other inclusion criteria had been met, the alliance measures that did not provide a minimum of two studies per version (e.g., short form) and per informant (e.g., client) were eliminated. For alliance measures where there were at least two studies but fewer than five reliability estimates per version and informant, moderator analyses were not conducted due to power considerations. However, mean reliability estimates and confidence intervals were calculated and reported for these measure variants.

Selection of Studies

Phase I of the literature review consisted of excluding the alliance measures that were intended to assess child, adolescent, or family samples (see Figure 1). This resulted in 10 out of 33 alliance measures being removed from the search. There were 3,256 reports identified in this initial database search of the remaining 23 therapeutic alliance measures. Phase II of the literature review involved the removal of duplicate studies. This resulted in eliminating 1,224 studies from the total. Phase III of the search consisted of reading the abstracts in order to remove any study that did not fit the inclusion criteria. This evaluation resulted in eliminating a total of 1,101 studies. Phase IV included a full article review to

determine fit with study inclusion criteria; this step resulted in the elimination of 633 studies. Phase V involved both abstract and full article review of the articles used by Horvath et al. (2011) but not identified in the search. As a result of this review, an additional 28 studies were added. At the end of these five phases, a total of 326 studies are available to be included in this meta-analysis.

With regards to the alliance measures, 11 out of the 23 (i.e., ARM, BLRI, CALPAS, CEI, CRF, HAq, TARS, TBS, TSRS, Vanderbilt scales, WAI) were used in the available studies. Analyses conducted for this study were based on reliability coefficients for these alliance measures within the 326 studies.

Prior to the analyses being conducted the data were screened for missing reliability coefficients. For studies in which reliability coefficients were not reported, attempts were made via email to contact the corresponding authors to obtain the missing information. The authors were asked to send information regarding the reliability coefficients obtained in the studies in question, or to send their datasets so that the reliability coefficients could be calculated. Out of a total of 129 requests that were sent to authors, 17 usable reliability coefficients were provided and 42 authors indicated that they no longer had the data requested. This information, when added to the studies in which a sample reliability coefficient was reported, resulted in data from 123 studies being available for use in this RG study. Once these data were obtained, the alliance measure variants that did not meet the requirement of a minimum of two reliability coefficient requirement (per version, per informant) were removed from further consideration. This resulted in the removal of an additional 5 alliance measures as well as 2 variants of other measures (the WAI-SR-T and

the HAq-I-T). Thus, there were 6 different alliance measures, in varying formats, included in the planned analyses. Of these 6 different alliance measures in their varying formats, only the full scale internal consistency (i.e., the correlation between all the items in the specific measure) was analyzed. In other words, because the focus of the study was on the broad construct of therapeutic alliance, the reliability of the specific subscales for any of the alliance measures was not assessed.

To control for potential non-independence of data in the planned analyses, each study was allowed to contribute only one reliability coefficient per measure and informant. For studies that reported multiple reliability coefficients for an alliance measure variant, the coefficient from data obtained nearest to the third session was selected. This time point was selected as the therapeutic alliance has been shown to be established by the third session (Horvath & Symonds, 1991).

Coding Sample Characteristics and Potential Moderators

Consistent with psychometric theory, a large number of RG studies have found that several different moderators have a significant effect on reliability estimates (López-López, Botella, Sánchez-Meca, & Marin-Martinez, 2012). The standard deviation of test scores, sample composition test length, and whether reliability estimates are obtained from original or adapted versions of a measure are frequently examined for their potential moderating effects within RG analyses (Botella & Ponte, 2011; Kieffer & MacDonald, 2011; Therrien & Hunsley, 2013). Indeed, it has been suggested that measure characteristics such as these should be routinely included when examining predictive models of heterogeneity of variance within an RG study (Botella & Ponte, 2011; López-López et al., 2012).

Accordingly, these characteristics, along with a number of other variables, were coded for each study and then tested as potential moderators if heterogeneity of variance was present (see Appendix B for complete coding manual).

The coded variables were: (a) *Year of publication* (coded as continuous); (b) *Language of article* (coded as 0 for English, and 1 for other); (c) *Type of report* (coded as 0 for journal, 1 for dissertation); (d) *Language of measure* (coded 0 for English, 1 for other); (e) *Alliance informant therapist* (coded 0 if alliance was not assessed by therapist, 1 if assessed by therapist); (f) *Alliance informant client* (coded 0 if alliance was not assessed by client, 1 if assessed by client); (g) *Alliance informant observer* (coded 0 if alliance was not assessed by observer, 1 if assessed by observer); (h) *Type of population* (coded as 0 for clinical sample, 1 for community sample, 2 for university student sample); (i) *Type of mental disorder treated* (coded as 0 for depression, 1 for anxiety, 2 for psychosis, 3 for personality disorder, 4 health-related problems, 5 for samples with multiple disorders, 6 for other); (j) *Type of treatment provided* (coded as 0 for forms of cognitive-behavioural therapy [CBT], 1 for forms of experiential therapy, 2 for forms of psychodynamic therapy, 3 for forms of interpersonal psychotherapy [IPT], 4 for a mixed sample of different psychotherapies, and 5 for other forms of therapy); (k) *Number of therapist and clients* (coded separately as continuous); (l) *Mean age and standard deviation of age of therapist and client* (both variables coded separately as continuous); (m) *Percentage of female therapists and percentage of female clients* (coded separately as continuous); (n) *Mean and standard deviation of years of experience of therapists* (coded separately as continuous); (o) *Mode of treatment* (coded 0 for face-to-face, 1 for distance communication, 2 for mixed); (p) *Sample size* (coded as a continuous variable); and (p) *Mean and standard deviation of*

test scores for therapists and clients (both variables coded as a continuous variable).

In order to examine the reliability of the coding process, a second rater coded 20% (*n*=24) of the articles coded by the primary investigator (*n*=117). Inter-rater reliability was assessed for continuous variable codes using intraclass correlation coefficients (ICC), (Shrout & Fleiss, 1979). Benchmarks recommended by Hunsley and Mash (2008) were used (i.e., ICC values of .70-.79 as adequate, ICC values of .80-.89 as good, and ≥ .90 as excellent). Reliability for the continuous moderator variables ranged from .79 to 1.00. Only one variable scored in the adequate range (i.e., age of client, ICC = .79), the remainder of the ICC scores were .80 or higher. Discrepancies among the raters for the age of client variable was determined to be the result of a simple error, as one rater coded the mean age value as missing, whereas the other rater coded had recorded a mean age value. Inter-rater reliability for categorical codes was assessed using the kappa benchmarks recommended by Hunsley and Mash (2008) (i.e., *k* values of .60-.74 as adequate, *k* values of .75 - .84 as good, and *k* values of ≥ .85 excellent). Reliability for the categorical moderator variables ranged from .65 to 1.00. As with the continuous variable, only one categorical variable had a value below the good range (i.e., type of population, Kappa = .65), the remainder of the variables scored .80 or higher. Difficulties in coding this variable originated from differences in interpreting the type of setting as either clinical or community based on the descriptions of the setting reported in the article. A number of settings could either be interpreted as either a hospital or community mental health center. To help clarify the setting, raters jointly searched the Internet to obtain more details specific to the setting in question. Discrepancies in coding were resolved through joint discussion and review of the articles until a unanimous agreement was reached.

Results

All data were analyzed using Comprehensive Meta-Analysis, Version 3 (CMA; Borenstein, 2016). The analytic procedures outlined by Borenstein, Hedges, Higgins, and Rothstein (2009) were followed, including having all analyses weighted by the inverse variance of the reliability coefficient (Graham et al., 2011). Most meta-analyses do not use *r* values in their computations, as these values tend not to be normally distributed. Therefore Fisher-*z* transformations were performed to normalize the sampling distribution of the *r* values (Borenstein et al., 2009). After the Fisher's *z* score and the standard deviation of the score were used in the analyses, the transformation process was reversed and the meta-analytic results were presented in their original reliability metrics (i.e., Cronbach's alpha). This transformation process is done automatically by the CMA program. Table 4 presents the descriptive statistics for the 6 alliance measures and their variants included in these analyses. CMA was used to calculate the mean reliability coefficient for each alliance measure variant, along with the 95% confidence intervals. Table 4 also presents the total number of the reliability estimates (*k*) contributed to the analysis by each measure and the range of reliability estimates for each of the alliance measures.

As previously noted, Hunsley and Mash (2008) described measures as having excellent internal consistency values when the preponderance of evidence demonstrates an alpha value of ≥.9, good internal consistency with an alpha value between .80-.89, and adequate internal consistency with an alpha value between .70-.79. Based on these guidelines, 10 alliance measure variants had mean values that demonstrated excellent internal consistency (ARM-28-C, CALPAS-24-C, CRF-S-C, HAq-O, WAI-C, WAI-S-C, WAI-SR-C, WAI-S-T, WAI-T). The three alliance measures in observer format obtained similar

estimates with a reliability average of *r* = .97, *k* =2 and a 95% CI = [.85, .99] for the WAI-O, *r* = .96, *k* = 3 with a 95% CI = [.93, .97] for the CALPAS-O, and *r* = .95, *k* = 2 with 95% CI = [.90, .98] for the HAq-O. The CRF-S-C also had a comparable mean alpha at *r* =.94, *k* = 6 and a 95% CI = [.93, .96].

The WAI-C and the WAI-SR-C had an average reliability estimates of *r* =.93, *k* = 37, 17, respectively, whereas the WAI-T had an average reliability estimate of *r* =.92, *k* = 21. For the WAI-C and WAI-T, these results are comparable to those reported by Hanson et al., (2002). With a large number of studies contributing to this WAI-C estimate, as well as the 95% CI = [.91, .94], the WAI-C's reliability estimate is fairly robust. Both the WAI-SR-C and WAI-T had similar impressive 95% CI = [.91, .94] and CI = [.88, .95], respectively. Although the WAI-SR-C and WAI-T were used in fewer studies than was the WAI-C, the results suggest that these three forms of the WAI are likely to provide highly reliable scores.

The ARM-28-C, WAI-S-C, and WAI-S-T each had an average reliability coefficient of *r* = .90, *k* = 2, 49, 21, respectively, with similar 95% CI = [.83, .94], [.89, .91], and [.88-.92]. As the mean for the ARM-28-C was calculated on only 2 studies, caution should be taken when considering this result.

There were four alliance measure variants that showed good internal consistency estimates, with mean alpha values greater than *r*= .80 (ARM-5-T, *k* = 2, CALPAS-24-C, *k* = 8, HAq-I-C, *k* = 8 and HAq-II-T *k* = 2). All four also had 95% confidence intervals with a lower bound greater than or equal to .80. The HAq-II-C had a similar mean alpha value of *r* =.84. However, the HAq-II-C 95% CI= [.73, .91] were less impressive. Lastly, the ARM-5-C and the TBS-C both were shown to have adequate reliability estimates. The ARM-5-C had an average reliability estimate of *r* =.78, *k* =4 with a 95 % CI = [.67, .86], whereas the TBS-C

had an average reliability estimate of r =.74, k =2 with a 95 % CI = [.48, .88]. As with other average estimates being calculated on small k, care should be taken when drawing conclusion from the results. However, in addition to the small k, particular attention should be drawn to the confidence intervals of these two measures. As confidence intervals help to establish precision of the effect size, with larger ranges indicating less precision, the accuracy of these two measures' mean reliability estimates is limited compared to other alliance measures included in this RG, with the exception of the HAq-II-C. Furthermore, when considering the lower end of these confidence intervals, these measures would be considered to have less than adequate reliability values according to the criteria set out by Hunsley and Mash (2008).

Q statistic and I^2 index

Heterogeneity of effects was calculated using the Q statistic and the I^2 index. The Q statistic and its p value were used to determine significance (Huedo-Medina, Sanchez-Meca, Marin-Martinez, & Botella, 2006), as a significant Q statistic provides evidence for true effect size variance among the different mean reliability coefficients for each measure. Next, the I^2 index was used to determine how much of the observed variance is due to true differences in effect sizes across studies (Borenstein et al., 2009). This statistic is expressed as a percentage (i.e., from 0% to 100%), with values moving closer to 100% indicating a likelihood the observed variance is non-random and may be explained by covariates. As presented in Table 4, approximately half of the alliance measures (ARM-5-C, CALPAS-24-C, HAq-II-C, TBC-C, WAI-C, WAI-S-C, WAI-SR-C, WAI-S-T, WAI-T) had significant Q statistics and high I^2 values. These results suggest that some of the observed variance is not be due to random sampling error alone and that moderator analyses are warranted. As

it has been noted that a small sample size affects the precision of these statistics (Borenstein et al., 2009; von Hippel, 2015), regardless of significant Q and high I^2, values, only those measures for which a minimum of 5 reliability coefficients were available were included in the moderator analyses. The HAq-II-C, TBS-C, and ARM-5-C were not included in the moderator analyses for this reason. Although the CRF-S-C did meet the required 5 reliability coefficient minimum, results did not indicate significant heterogeneity, and therefore, moderator analyses were not conducted on this measure. Finally, it should be noted that, although many measures (HAq-II-T, HAq-I-C, HAq-O, ARM-28-C, ARM-5-T, CALPAS-O, WAI-O) failed to obtain a significant Q statistic, this the may be due to the low number of reliability coefficients ($k \leq 3$) included in these analyses rather than the effect sizes of these measures truly being homogeneous in nature.

Analysis of Moderator Variables

As multiple comparisons were planned for the moderator analyses, there was a need to find a balance between possible Type I and Type II errors. Striking a balance between these two types of error can be accomplished by comparing the risks and benefits associated with each error rate with the purpose of the analysis (Borenstein et al., 2009; Minium, Clarke, & Coladarci, 1999). The purpose of this RG meta-analysis is to detect small but discernable differences between the alliance measures among different populations. As a result, the consequences associated with committing Type I error were considered to be relatively benign and unlikely to be detrimental to the domain of therapeutic alliance research. On the other hand, having lower power, and thus less ability to detect the small but discernable differences, may result in committing a Type II error. Committing this type

of error could have an impact on results not being further studied by alliance researchers. Therefore, the decision was made to use, for all analyses, a significance level of $p < .05$.

For those measures that met the inclusion criteria of $k = 5$ or greater and that had a significant Q value (i.e., CALPAS-24-C, WAI-C, WAI-S-C, WAI-SR-C, WAI-S-T, and WAI-T) moderator analyses were performed to test the relations between sample characteristics and each measure's mean reliability estimate. To assess the relation between the obtained effect sizes and the potential moderators that are continuous in nature, a series of simple meta-regressions were conducted. Sample characteristics included in these analyses were sample size (both number of clients and therapists), publication year, mean and standard deviation of test scores (both client and therapist versions), mean and standard deviation of age (both clients and therapists), mean and standard deviation of years of experience of therapists, and percentage of women in the sample (both clients and therapists). To be included in an analysis for a given measure, values for potential moderators must have been available in at least three studies because the CMA program cannot compute meta-regressions with data from fewer studies.

Moderator analyses with statistically significant results are presented in Table 5. For the WAI-S-C, the mean age of therapists, the therapists' years of experience and the standard deviation of the clients' scores were found to explain some of the variance in the measure's mean reliability coefficients. The WAI-S-T had the mean age of therapists and the percentage of female therapists as significant moderators. The clients' as well as the therapists' sample size also explained some of the variance for the measure's reliability estimate. However, upon a closer look at the clients' samples sizes, it was observed that one

study's sample size was dramatically larger than the other 20 studies included in this analysis (i.e., 35% greater than next largest sample size and 180% greater than third largest sample size). Once removed from the analysis, the clients' sample size was no longer a significant moderator for the WAI-S-T.

The WAI-T had three significant moderators: The mean and standard deviation of the therapists' age as well as the therapists' sample size. Whereas, the WAI-SR-C had two significant moderators: The mean and the standard deviation of the clients' scores. As for the WAI-C and the CALPAS-24-C, both these alliance measures had one significant moderator, with the publication year explaining some of the variance for the WAI-C and the standard deviation of the clients' score explaining some variance for the CALPAS-24-C.

Subgroup analyses were conducted in order to assess the potential moderating effects on the reliability coefficients of sample characteristics that were categorical in nature. Following methods described by Borenstein et al. (2009), a variant of analysis of variance (ANOVA) was used to compare the subgroup means for the following sample characteristics: (a) population, (b) type of treatment provided, (c) primary mental disorder in the study sample, (d) language of measure, and (e) mode of treatment. The CALPAS-24-C, WAI-C, WAI-S-C, WAI-SR-C, WAI-S-T, and WAI-T were included in the subgroup analysis.

When there are five or few studies within a specific subgroup, it is likely that estimates of tau-squared for the analyses will be imprecise (Borenstein, Hedges, Higgins, & Rothstein, 2010). Therefore, it has been suggested to conduct these analyses using the tau-squared pooled estimate as the increase in accuracy gained by pooling more studies is likely to be greater than true differences between the subgroups (Borenstein et al., 2010).

This approach was taken in analyses where (a) there were uneven numbers between the subgroups and, (b) most subgroups had five or fewer studies contributing data, or (c) there were only two subgroups and one subgroup had data from five or fewer studies.

As noted in Table 5, only three of moderators in three different measure variants were found to be significant. For WAI-S-C, the type of treatment provided helped explain the variance within this measure. For type of treatment, experiential therapies had the highest reliability estimates (.95), treatments based on a mix of therapeutic orientations was next (.90), followed by cognitive therapies (.88), and then psychodynamic therapies (.87). As the *k* was not equal for each of these moderators, with some types of treatment contributing only one estimate and others as many as 15 estimates, these results should be seen as tentative.

For the WAI-T the moderator of population was shown to be significant. Again, with the caveat of unequal contributions to average effect sizes, the alliance measured within samples drawn from community settings had the highest reliability estimate (.94). The alliance being measured in samples drawn from mixed settings (i.e., community and clinical) also had an excellent reliability average at .92, whereas the alliance measured in samples drawn from clinical settings had a good reliability average at .82.

As for the WAI-C, the type of mental disorder/problem treated was significant. Both samples with depression and samples with personality disorders were found to have an excellent reliability estimates (both estimates at .94). Both the “mixed” and “other” categories showed good reliability estimates (mixed = .89, other = .88).

Publication Bias

Orwin's fail-safe *N*

As this RG study included data only from published articles and dissertation studies, it is important to assess whether publication bias may have influenced study results. To address this concern, Orwin's fail-safe *N* (1983) was calculated to obtain an estimate of how many studies with an average reliability of .50 would be needed to drop the average reliability of a measure below .70 (Graham & Christiansen, 2009; Therrien & Hunsley, 2013). As seen in Table 4, there are a number of measure variants with low fail-safe values. For example, the CALPAS-24-C would need 18 studies with a reliability below .50 to bring down its mean reliability (α=.88) to .70, and the HAq-II-C would need only 5 studies with a reliability below .50 to lower its mean reliability (α=.88) to .70. Low values such as these suggest that the results from these measures should be viewed as tentative, although they do provide the best reliability estimates currently available. In the case of the WAI, all versions of the measure had relatively high Orwin's *N* values, ranging from 35 to 86 studies. This indicates that mean reliability estimates for these measures are fairly robust and unlikely to be negatively affected by unpublished research findings.

Funnel Plots

Funnel plot diagrams help to assess the distribution of the effect sizes of the studies included in a meta-analysis and, thus, provide another way to examine possible publication bias effects (Borenstein et al., 2010). A funnel plot that has its point distributed symmetrically around the mean indicates the likely absence of publication bias as the sampling error is considered random. This study's funnel plots have the standard error on the vertical axis (*y*-axis) and the effect sizes (i.e., Fisher's *z*; higher scores representing

higher reliability coefficients) on the horizontal (*x*-axis). The use of the standard error on the *y*-axis, as opposed to the sample size or variance, makes it easier to identify asymmetry. By plotting the data in this fashion, it allows for the points to be spread out on the bottom half of the diagram. This is where the studies with smaller sample sizes are plotted, typically due to having more sampling error variation (Borenstein et al., 2010). To ensure there is adequate power to distinguish chance from real asymmetry, having a 10-study minimum when testing for publication bias is recommended (Higgins & Green, 2011). Therefore, only those measures that had a *k* of 10 or more were included in this analysis (see Figures 2 to 7). The asymmetrical distribution of effect sizes for the WAI-C, WAI-SR-C, WAI-T, and the WAI-S-T suggest a potential publication bias for these measures. Egger's tests were then conducted to provide further evidence of asymmetry (Borenstein, 2005). Results indicated that there was significant asymmetry only with the WAI-S-T (Egger's test = 3.62, $p = 0.007$). Next a Trim and Fill test was performed to help account for the possible missing studies. Results of this test suggest that only 4 studies with reliability coefficients smaller than the mean (i.e., < .9) would be needed to account for the missing data (Borenstein, 2005). However, once the missing studies were accounted for, the mean reliability value for the WAI-S-T was only marginally impacted ($r = .88$).

Discussion

The purpose of the present study was to (a) explore the typical score reliability of the therapeutic alliance measures most commonly used in the adult psychotherapy research literature and (b) identify specific study characteristics that may influence these reliability estimates. After reviewing the available literature, six different alliance measures (ARM, CALPAS, CRF, HAq, TBS, WAI), in various formats and rater versions, were included

in the analyses, resulting in a total of 17 alliance measure variants being included in this RG study. Of the measures examined in this study, 10 (ARM-28-C, CALPAS-O, CRF-S-C, HAq-O, WAI-C, WAI-O, WAI-S-C, WAI-SR-C, WAI-S-T, WAI-T) showed evidence of having high reliability coefficients.

By far, the WAI is the most frequently used alliance measure within the literature, so it was not completely unexpected that all six of the WAI variants included in this RG demonstrated excellent reliability estimates. These high reliability values are consistent with the results of previous research (i.e., Hanson et al., 2002) and add further support to statements made within the literature concerning the psychometric strengths of the various versions of the WAI.

It should be noted that, although the WAI is the most frequently used alliance measure, the observer version (WAI-O), along with four other measure variants (ARM-28-C, CALPAS-O, CRF-S-C, HAq-O), had mean reliability estimates that were derived from relatively few studies. As such, it is important to view the results from these measures with appropriate caution. Also, of some significance, is that three out of these five alliance measures (WAI-O, CALPAS-O, HAq-O) are observer-rated alliance measure variants. The limited number of available reliability coefficients for the three observer alliance measure variants may be due, in part, to researchers focusing more on the interrater reliability rather than internal consistency reliability when using observer data. However, these two types of reliability are assessing two different aspects of a measure (interrater – consistency across raters, internal consistency – consistency between the items), and therefore, should both be reported for a particular sample in each study.

As for the ARM-5-T, CALPAS-24-C, HAq-I-C, HAq-II-C, and HAq-II-T, based on their overall reliability estimates, they appear to be fairly good choices for assessing the alliance, although the reliability estimates are lower than those obtained for the versions of the WAI. The ARM-5-C and the TBS-C only showed adequate reliability estimates based on the results of this study and, therefore, may not be the best option for measuring the therapeutic alliance.

Results of this RG found no substantial difference between the therapist and client versions of the same measure, with the exception of the ARM. The therapist and client versions for all of the formats of the WAI, as well as the HAq-II, differed only by a small amount (e.g., mean alpha values of .90 versus .92). These results suggest that, although there appear to be notable differences in the item content between the different versions completed by clients and therapists (e.g., Hatcher & Barends, 1996; Horvath & Bedi, 2002), these differences appear not to have an impact on the measure's ability to reliably assess the alliance.

The Effects of Sample Characteristics on Reliability Estimates

Moderator analyses were conducted on 6 of the 17 alliance measure variants, with 5 of these measures being a different format or version of the WAI. Although the CRF-S-C and the HAq-I-C had at least five reliability coefficients in our sample of studies, these two measures did not demonstrate heterogeneity of variance in their estimates and, therefore, moderator analysis was not warranted. Given the limited number of reported reliability estimates available for the other measures, the CALPAS-C-24 was the only other measure beyond the WAI variants I was able to examine the potential moderating effects of study characteristics and effect size. In general, results indicated that the WAI-S-C produced

higher reliability estimates in samples with younger and less experienced therapists as well as in samples where there was higher variability among the clients' scores on the alliance measure. Findings also suggested that the type of treatment being provided impacted the reliability estimates for the WAI-S-C. These results indicated that samples that provided experiential therapy or samples that had a combination of different orientations providing therapy produced higher reliabilities than the other treatment modalities for this measure. Notwithstanding that additional data are needed to substantiate these results, the findings here suggest that the WAI-S-C provides reliable alliance scores. However, these scores appear to be influenced by several different study characteristics. Research, therefore, should bear this in mind when determining to use the short version of the WAI.

Like the WAI-S-C, the WAI-S-T also produced higher reliability estimates in samples with younger therapists. Higher reliability estimates were also observed in samples where there were a larger percentage of female therapists. In addition, the clients' and therapists' sample size had an effect on the WAI-S-T's reliability estimates, with larger samples yielding lower reliability estimates for both variables. The moderator effect for the clients' sample size may be anomalous, however, as it was only one large study that influenced the results—once this study was removed from the analysis, sample size was no longer a significant moderator.

Similar to the two previous measure variants, the WAI-T produced higher reliability estimates in samples with younger therapists. Reliability estimates for the WAI-T were found to be higher when there was more variability among the therapists' age, when samples were drawn from a community setting, and with smaller therapists' sample sizes. For the WAI-SR-C, higher reliability estimates were found in samples where clients scored

higher on the alliance measure and when those scores were more dispersed. This pattern of larger dispersion among the clients' scores producing higher reliability estimates was also found for the CALPAS-24-C. As for the WAI-C, results indicate that studies that had older publication dates and samples that were being treated for depression or a personality disorder had higher reliability estimates. All of these potential influences on score reliability should be considered when researchers are determining which alliance measure to use. With this being said, the influence from one of these moderating variables will have less of a meaningful impact on a study's reliability estimate when the alliance is assessed from a measure with high reliability and narrow confidence interval levels.

Although these moderator analyses shed some light on the impact a number of study characteristics have on an alliance measure's reliability estimate, the moderator analyses were limited by the infrequency with which study authors provided the necessary details in their reports. Consistent with other research (e.g., Hanson et al., 2002), many articles did not contain details such as the mean and standard deviation of alliance scores for both clients and therapists, years of experience of therapists, type of treatment provided, and the main client presenting problem. For this RG, most of the above moderators could only contribute a small *k* to the analysis and the impact of many other potential moderators could not be examined because of insufficient studies containing the required information. Nevertheless, based on these results, researchers should be aware of a number of sample specific characteristics (e.g., distribution of scores, experience of therapists) that may have an impact on the likely score reliability of their chosen alliance measure.

Limited Reporting of Reliability

Out of 326 studies eligible to be included in this RG, 56% did not report a reliability estimate based on scores from study participants (33% used reliability induction and 23% failed to mention reliability). Despite the repeated emphasis being placed on the importance of researchers reporting their sample's reliability estimates, (e.g., APA, 2008; Therrien & Hunsley, 2013; Vacha-Haase & Thompson, 2011; Wilkinson et al., 1999), failure to do so is still commonplace. Although these numbers are rather discouraging, results from this RG may be pointing towards a shift in reporting study-specific reliability, as the reporting rate is above the average observed in Vacha-Haase and Thompson's (2011) review of the RG literature. In this review the authors found, on average, RG studies were based on 18.7% of the primary studies reporting their sample's reliability, whereas, the primary studies' reporting rate for the present RG was 34.4%. What is more, in the studies used in the present RG there was an increase in reporting sample-specific reliability estimates in recent years, with an average reporting rate of 14.5 studies per publication year from 2010 onward, compared to an average reporting rate of 5.7 studies per publication year for the years between 1982 to 2009. Even though this increase may be relatively small, it is somewhat encouraging as it suggests that the reporting of reliability estimates by researchers is moving in the right direction.

Limitations

This study has a number of limitations. First, as with all meta-analyses, this study's main limitation relates to the identification and retrieval of studies that used the targeted therapeutic alliance measures. To address this challenge, I searched the two most common databases related to the field of psychology, PsycInfo and PubMed. It is likely, however, that

a number of unpublished studies were not located. In addition, a number of studies identified in the initial search were not retrievable. Of those studies identified as "not accessible" approximately 60% were written in a language other than English or French (most frequently, German, Spanish, and Chinese), 20% were unpublished dissertations, and 20% were in the unavailable journals. Attempts were made to retrieve the unpublished dissertations identified in Horvath et al. (2011) meta-analysis, but many of these studies were not accessible.

In addition, as there are databases that were not searched (e.g., CINHAL), it is conceivable that a number of counselling, nursing, or social work articles may have been missed. However, articles identified from the PsycInfo and PubMed databases did include a number of studies that were specific to these three disciplines and/or were in journals with a focus on these disciplines.

Second, as in the case with other RG studies, the majority of retrieved studies did not provide reliability estimates for the sample. Attempts were made to obtain additional reliability estimates by contacting the articles' corresponding authors. However, these efforts were met with little success which, in turn, led to the exclusion of a number of alliance measures from this study. This lack of data also resulted in the majority of the measure variants having only a small number of studies contributing to the mean reliability estimates and the moderator analyses. This, in turn, affected the statistical power of the meta-analyses, as well as the generalizability of a number of the findings. That being said, given that the present analyses and results are based on studies containing data from many clients and therapists, the findings of this RG currently provide the best reliability estimates available for these alliance measures.

As the Orwin's fail-safe *N* takes into consideration possible missing data (i.e., unpublished studies, published studies that failed to report reliability estimates), analyses to estimate fail-safe values were conducted to address these two limitations. The high Orwin's *N* values obtained for the 5 WAI measure variants (WAI-C, WAI-S-C, WAI-SR-C, WAI-S-T, WAI-T) suggest that these 5 mean reliability estimates are likely to provide an accurate perspective on the measures' reliability estimates and are unlikely to be negatively impacted by either data in unpublished studies or missing reliability values in published studies. Unfortunately, the same cannot be said with regard to the other alliance measures, as low Orwin's *N* values were obtained for them.

Funnel plots were also utilized to address these limitations. The lack of evidence of an asymmetrical distribution for the WAI-C, WAI-S-T, WAI-SR-C, and WAI-T provide further support that unpublished studies or missing reliabilities values would be unlikely to negatively impact the measure variants' mean reliability coefficients. The asymmetrical distribution of the WAI-S-T reliability values suggested a publication bias; however, when adjusted to account for possible missing studies, there was only minimal effect on the measure's mean reliability coefficient. Unfortunately, the other 12 alliance measure variants could not be tested for publication bias using funnel plots.

Conclusions and Recommendations

Numerous measures have been developed to assess the therapeutic alliance between client and therapist. With more emphasis being placed on researchers and clinicians to use evidence-based measures, it becomes increasingly important to determine which of these alliance measures are the most psychometrically sound. Based on our analyses, 15 out of 17 measure variants (ARM-5-T, ARM-28-C, CALPAS-24-C, CALPAS-O,

CRF-S-C, HAq-I-C, HAq-II-C, HAq-II-T, HAq-O, WAI-O, WAI-C, WAI-S-C, WAI-SR-C, WAI-S-T, and WAI-T) demonstrated sufficiently strong reliability evidence for their use in measuring the alliance. Of these 15 measure variants, the WAI-C, WAI-S-C, WAI-SR-C, WAI-S-T, and WAI-T were the most robust with respect to their high mean reliability estimates and narrow confidence interval levels, although results of moderator analyses and funnel plots do indicate that there are factors that might affect the reliability of scores obtained in studies using these measures.

Although this study provides some evidence that reporting practices may be improving, failure to report study-specific reliability continues to be all too common. It is for this reason that I strongly encourage researchers to calculate and report reliability coefficients for their own sample rather than simply relying on reliability induction and/or assuming that scores on a measure will be reliable just because other researchers frequently use the measure. Additionally, I also strongly encourage all researchers to report full details on their samples, as required by the APA publication manual (APA, 2010). This includes providing information on both those who receive treatments and those who provide the treatments. For a large number of studies in this RG authors failed to provide any details on the therapists, particularly when alliance measure was completed only by clients. This is unfortunate, as this lack of information limited my ability to analyze of these sample-specific characteristics and the influence they may exert on the reliability estimates of scores on therapeutic alliance measures.

References

References with an asterisk (*) indicate the studies that were included in the RG analysis.

American Psychological Association (APA). (2010). *Publication Manual of the American Psychological Association* (6th ed.). Washington, DC: American Psychological Association

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

American Psychological Association Publications and Communications Board Working Group on Journal Article Reporting Standards (2008). Reporting standards for research in psychology: Why do we need them? What might they be? *American Psychologist, 63,* 839-851. doi: 10.1037/0003-066X.63.9.839

*Andersson, G., Paxling, B., Wiwe, M., Vernmark, K., Felix, C. B., Lundborg, L., … Carlbring, P. (2012). Therapeutic alliance in guided internet-delivered cognitive behavioural treatment of depression, generalized anxiety disorder and social anxiety disorder. *Behaviour Research and Therapy, 50*, 544-550. http://dx.doi.org/10.1016/j.brat.2012.05.003

*Andrade-Gonzalez, N., & A. Fernandez-Liria, A. (2015). Spanish adaptation of the Revised Helping Alliance Questionnaire (HAq-II). *Journal of Mental Health, 24,* 155-161. doi: 10 3109/09638237 2015 1036975

Ardito, R. B., & Rabellino, D. (2011). Therapeutic alliance and outcome of psychotherapy: Historical excursus, measurements, and prospects for research. *Frontiers in Psychology*, *2*, 270. http://doi.org/10.3389/fpsyg.2011.00270

*Arnow, B. A., Blasey, C., Manber, R., Constantino, M. J., Markowitz, J. C., Klein, D. N., … Rush, A. (2007). Dropouts versus completers among chronically depressed outpatients. *Journal of Affective Disorders, 97,* 197-202. http://dx.doi.org/10.1016/j.jad.2006.06.017

*Arnow, B. A., Steidtmann, D., Blasey, C., Manber, R., Constantino, M. J., Klein, D. N., … Kocsis, J. H. (2013). The relationship between the therapeutic alliance and treatment outcome in two distinct psychotherapies for chronic depression. *Journal of Consulting and Clinical Psychology, 81*, 627-638. doi: 10 1037/a0031530

*Auszra, L., Greenberg, L. S., & Herrmann, I. (2013). Client emotional productivity-optimal client in-session emotional processing in experiential therapy. *Psychotherapy Research, 23*, 732-746. http://dx.doi.org/10.1080/10503307.2013.816882

*Bachelor, A., Meunier, G., Laverdiere, O., Gamache, D. (2010). Client attachment to therapist: Relation to client personality and symptomatology, and their contributions to the therapeutic alliance. *Psychotherapy: Theory, Research, Practice, Training, 47*, 454-468. http://dx.doi.org/10.1037/a0022079

*Baldwin, S. A., Wampold, B. E., & Imel, Z. E. (2007). Untangling the alliance-outcome correlation: Exploring the relative importance of therapist and patient variability in the alliance. *Journal of Consulting and Clinical Psychology, 75*, 842-852. http://dx.doi.org/10.1037/0022-006X.75.6.842

*Barber, J. P., Connolly, M. B., Crits-Christoph, P., Gladis, L., & Siqueland, L. (2000). Alliance predicts patients' outcome beyond in-treatment change in symptoms. *Journal of Consulting and Clinical Psychology, 68*, 1027-1032. http://dx.doi.org/10.1037/0022-006X.68.6.1027

*Barber, J. P., Gallop, R., Crits-Christoph, P., Barrett, M. S., Klostermann, S., McCarthy, K. S., & Sharpless, B. A. (2008). The role of the alliance and techniques in predicting outcome of supportive-expressive dynamic therapy for cocaine dependence. *Psychoanalytic Psychology, 25*, 461-482. http://dx.doi.org/10.1037/0736-9735.25.3.461

*Barrowclough, C., Meier, P., Beardmore, R., & Emsley, R. (2010). Predicting therapeutic alliance in clients with psychosis and substance misuse. *The Journal of Nervous and Mental Disease, 198*, 373-377. doi: 10 1097/NMD 0b013e3181da4d4e

*Birringer, J. A. (1999). *The in-session thoughts and feelings of counselor trainees in relationship to measures of the working alliance, counselor self-estimate and adult attachment style* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 1999-95015-012)

Bordin, E. S. (1979). The generalizability of the psychoanalytic concept of the working alliance. *Psychotherapy: Theory, Research and Practice, 16*, 252–260. http://dx.doi.org/10.1037/h0085885

Borenstein, M., (2005). Software for Publication Bias. In H. R. Rothstein, H. R., Sutton, A. J., & Borenstein, M. (Eds), *Publication Bias in Meta-Analysis: Prevention, Assessment and Adjustments* (pp. 193-220). John Wiley & Sons, Ltd, Chichester, UK. doi: 10.1002/0470870168

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

Borenstein, M., Hedges, L. V., Higgins, & J. P. T., Rothstein, H. R. (2010). A basic

introduction to fixed-effect and random-effects models for meta-analysis. *Research Synthesis Methods, 1,* 97-111. doi: 10.1002/jrsm.12

Borenstein, M., Hedges, L., Higgins, J. P. T., & Rothstein, H. R. (2016). Comprehensive Meta-Analysis Version 3. Englewood, NJ: Biostat.

Botella, J., & Point, G. (2011). Effects of the heterogeneity of the variances on reliability generalization: An example with the Beck Depression Inventory. *Psicothema, 23*, 516-522.

*Brenes, G. A., Miller, M. E., Williamson, J. D., McCall, W., Knudson, M., & Stanley, M. A. (2012). A randomized controlled trial of telephone-delivered cognitive-behavioral therapy for late-life anxiety disorders. *The American Journal of Geriatric Psychiatry, 20*, 707-716. http://dx.doi.org/10.1097/JGP.0b013e31822ccd3e

*Busseri, M. A. & Tyler, T. D. (2003). Interchangeability of the Working Alliance Inventory and Working Alliance Inventory, Short Form. *Psychological Assessment, 15*, 193-197. http://dx.doi.org/10.1037/1040-3590.15.2.193

*Byrd, K. R., Patterson, C. L., & Turchik, J. A. (2010). Working alliance as a mediator of client attachment dimensions and psychotherapy outcome. *Psychotherapy: Theory, Research, Practice, Training, 47*, 631-636. http://dx.doi.org/10.1037/a0022080

*Cahill, J., Stiles, W. B., Barkham, M., Hardy, G. E., Stone, G. Agnew-Davies, R., & Unsworth, G. (2012). Two short forms of the Agnew Relationship Measure: The ARM-5 and ARM-12. *Psychotherapy Research, 22*, 241-255. http://dx.doi.org/10.1080/10503307.2011.643253

*Chisholm, S. M. A. (1998). *A comparison of the therapeutic alliances of premature*

terminators versus therapy completers (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 1998-95024-077)

*Cieslak, E. N. (2009). *Hope in psychotherapy process and outcome* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 2009-99060-439)

*Coleman, D. (2005). Psychodynamic and cognitive mechanisms of change in adult therapy: A pilot study. *Bulletin of the Menninger Clinic, 69*, 206-219. http://dx.doi.org/10.1521/bumc.2005.69.3.206

*Coleman, D. (2006). Client personality, working alliance and outcome: A pilot study. *Social Work in Mental Health, 4*, 83-98. http://dx.doi.org/10.1300/J200v04n04_06

*Constantine, M. G. (2007). Racial microaggressions against African American clients in cross-racial counseling relationships. *Journal of Counseling Psychology, 54*, 1-16. http://dx.doi.org/10.1037/0022-0167.54.1.1

*Constantino, M. J., Schwaiger, E. M., Smith, J. Z., DeGeorge, J. McBride, C., Ravitz, P., & Zuroff, D. C. (2010). Patient interpersonal impacts and the early therapeutic alliance in interpersonal therapy for depression. *Psychotherapy: Theory, Research, Practice, Training, 47*, 418-424. http://dx.doi.org/10.1037/a0021169

*Cook, J. E. & C. Doyle (2002). Working alliance in online therapy as compared to face-to-face therapy: Preliminary results. *CyberPsychology & Behavior, 5*, 95-105. http://dx.doi.org/10.1089/109493102753770480

*Coutinho, J., Ribeiro, E., Fernandes, C., Sousa, I., & Safran, J. D. (2014). The development of the therapeutic alliance and the emergence of alliance ruptures. *Anales de Psicologia, 30*, 985-994. http://dx.doi.org/10.6018/analesps.30.3.168911

* Crameri, A., von Wyl, A., Koemeda, M., Schulthess., & Tschuschke, V. (2015). Sensitivity analysis in multiple imputation in effectiveness studies of psychotherapy. *Frontiers in Psychology, 6,* 1-11. http://doi:10.3389/fpsyg.2015.01042

*Davis, L. W., Eicher, A. C., & Lysaker, P. H. (2011). Metacognition as a predictor of therapeutic alliance over 26 weeks of psychotherapy in schizophrenia. *Schizophrenia Research, 129*, 85-90. http://dx.doi.org/10.1016/j.schres.2011.02.026

*De Bolle, M., Johnson, J. G., & De Fruyt, F. (2010). Patient and clinician perceptions of therapeutic alliance as predictors of improvement in depression. *Psychotherapy and Psychosomatics, 79*, 378-385. http://dx.doi.org/10.1159/000320895

*Delsignore, A., Rufer, M., Moergeli, H., Emmerich, J., Schlesinger, J., Milos, G., … Weidt, S. (2013). California Psychotherapy Alliance Scale (CALPAS): Psychometric properties of the German version for group and individual therapy patients. *Comprehensive Psychiatry, 55*, 736-742. http://dx.doi.org/10.1016/j.comppsych.2013.11.020

*Derisley, J., & S. Reynolds (2000). The transtheoretical stages of change as a predictor of premature termination, attendance and alliance in psychotherapy." *British Journal of Clinical Psychology, 39*, 371-382. http://dx.doi.org/10.1348/014466500163374

*de Roten, Y., Fischer, M., Drapeau, M., Beretta, V., Kramer, U., Favre, N., & Despland, J. (2004). Is one assessment enough? Patterns of helping alliance development and outcome. *Clinical Psychology & Psychotherapy, 11*, 324-331. http://dx.doi.org/10.1002/cpp.420

*Diaz, N. M. (2004). *Development of the therapeutic alliance in cross-cultural psychotherapy dyads* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 2004-99004-154)

Dimitrov, D. M. (2002). Reliability: Arguments for multiple perspectives and potential problems with generalization across studies. *Educational and Psychological Measurement, 62,* 783-801. doi: 10.1177/001316402236878

Elvins, R., and Green, J. (2008). The conceptualization and measurement of therapeutic alliance: an empirical review. *Clinical Psychological Review, 28,* 1167–1187. doi:10.1016/j.cpr.2008.04.002

*Emmerling, M. E., & W. J. Whelton (2009). Stages of change and the working alliance in psychotherapy. *Psychotherapy Research, 19*, 687-698. http://dx.doi.org/10.1080/10503300902933170

*Errazuriz, P., Constantino, M. J., & Calvo, E. (2014). The relationship between patient object relations and the therapeutic alliance in a naturalistic psychotherapy sample. *Psychology and Psychotherapy: Theory, Research and Practice, 20*, 254-269.

Flückiger, C., Del Re, A. C., Wampold, B. E., Symonds, D. & Horvath, A. O. (2012). How central is the alliance in psychotherapy? A multilevel longitudinal meta-analysis. *Journal of Counseling Psychology*. *59*, 10-17. doi: 10.1037/a0025749

*Fuertes, J. N., Stracuzzi, T. I., Bennett, J., Scheinholtz, J., Mislowack, A., Hersh, M., & Cheng, D. (2006). Therapist multicultural competency: A study of therapy dyads. *Psychotherapy: Theory, Research, Practice, Training, 43*, 480-490. http://dx.doi.org/10.1037/0033-3204.43.4.480

*Fuertes, J. N., Mislowack, A., Brown, S., Gur-Arie, S., Wilkinson, S., & Gelso, C. J. (2007).

Correlates of the real relationship in psychotherapy: A study of dyads. *Psychotherapy Research, 17*, 423-430. http://dx.doi.org/10.1080/10503300600789189

*Gaston, L. (1991). Reliability and criterion-related validity of the California Psychotherapy Alliance Scales-patient version. *Psychological Assessment: Journal of Consulting and Clinical Psychology, 3*, 68-74. http://dx.doi.org/10.1037/1040-3590.3.1.68

Gaston, L., & Marmar, C. R. (1994). The California Psychotherapy Alliance Scales. In A. O. Horvath & L. S. Greenberg (Eds.), *The working alliance: Theory, research, and practice* (pp. 85-108). New York: Wiley.

Geisinger, K. F. (2013). Reliability. In Geisinger, K. F., Bracken, B., Carlson, J. F., Hansen, J. C., Kuncel, N. R., Reise, S. P., & Rodriguez, M. C. (Eds.), *APA Handbook of Testing and Assessment in Psychology, Vol. 1: Test Theory and Testing and Assessment in Industrial and Organizational Psychology* (pp. 21-42). American Psychological Association. Washington, DC.

*Geller, S. M., Greenberg, L. S., & Watson, J. C. (2010). Therapist and client perceptions of therapeutic presence: The development of a measure. *Psychotherapy Research, 20*, 599-610. http://dx.doi.org/10.1080/10503307.2010.495957

*Goldman, E. D. (2009). *Chicken or egg, alliance or outcome: An attempt to answer an age-old question* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 2009-99120-159)

Graham, J. M., & Christiansen, K. (2009). The reliability of romantic love: A reliability generalization meta-analysis. *Personal Relationships, 16,* 49-66. doi: 10.1111/j.1475-6811.2009.01209.x

Graham, J. M., Diebels, K. J, & Barnow, Z.B. (2011). The reliability of relationship satisfaction: A reliability generalization meta-analysis. *Journal of Family Psychology, 25*, 39-48. http://psycnet.apa.org/doi/10.1037/a0022441

*Gullo, S., Lo Coco, G., & Gelso, C. (2012). Early and later predictors of outcome in brief therapy: The role of real relationship. *Journal of Clinical Psychology, 68*, 614-619. http://dx.doi.org/10.1002/jclp.21860

Hanson, W. E., Curry, K. T., & Bandalos, D. L. (2002). Reliability generalization of Working Alliance Inventory scale scores. *Educational and Psychological Measurement*, 62, 659-673. doi: 10.1177/001316402128775076

*Hara, K. M., Westra, H. A., Aviram, A., Button, M. L., Constantino, M. J., & Antony, M. M. (2015). Therapist awareness of client resistance in cognitive-behavioral therapy for generalized anxiety disorder. *Cognitive Behaviour Therapy, 44*, 162-174. http://dx.doi.org/10.1080/16506073.2014.998705

Hatcher, R. L. (2010). Clinical studies of the therapeutic alliance. In Muran, J. C. & Barber, J. P. (Eds). *The therapeutic alliance: An evidence-based guide to practice* (pp.7-28). New York, NY: Guilford

*Hatcher, R. L., Barends, A., Hansell, J., & Gutfreund, M. J. (1995). Patients' and therapists' shared and unique views of the therapeutic alliance: an investigation using confirmatory factor analysis in a nested design. *Journal of Consulting and Clinical Psychology, 63*, 636-643.

Hatcher, R. L., & Barends, A. W. (1996). Patients' view of the alliance in psychotherapy: exploratory factor analysis of three alliance measures. *Journal of Consulting and Clinical Psychology, 64*, 1326–1336. doi: 10.1037/0022-

006X.64.6.1326

*Hayes, J. A., Yeh, Y., & Eisenberg, A. (2007). Good grief and not-so-good grief: Countertransference in bereavement therapy. *Journal of Clinical Psychology, 63*, 345-355. http://dx.doi.org/10.1002/jclp.20353

*Hayes-Skelton, S. A., Roemer, L., & Orsillo, S. M. (2013). A randomized clinical trial comparing an acceptance-based behavior therapy to applied relaxation for generalized anxiety disorder. *Journal of Consulting and Clinical Psychology, 81*, 761-773. http://dx.doi.org/10.1037/a0032871

Henchy, A. M. (2013). *Review and evaluation of reliability generalization research.* (Unpublished dissertation). University of Kentucky, U.S.A.

Henson, R. K. (2001). Understanding internal consistency reliability estimates: A conceptual primer on coefficient alpha. *Measurement and Evaluation in Counseling and Development*, *34*, 177-189.

*Heinonen, E., Lindfors, O., Harkanen, T., Virtala, E., Jaaskelainen, T., & Knekt, P. (2014). Therapists' professional and personal characteristics as predictors of working alliance in short-term and long-term psychotherapies. *Clinical Psychology and Psychotherapy, 21*, 475-494. doi: 10 1002/cpp 1852

Higgins, J. P. T., & Green, S. (Eds). (2011). *Cochrane handbook for systematic reviews of interventions* Version 5.1.0 [updated March 2011]. The Cochrane Collaboration, 2011. Retrieved from http://handbook.cochrane.org/chapter_10/10_4_3_1_recommendations_on_testing_for_funnel_plot_asymmetry.htm

*Hill, C. E., Baumann, E., Shafran, N., Gupta, S., Morrison, A., Rojas, A. E., … Gelso, C. J. (2015).

Is training effective? A study of counseling psychology doctoral trainees in a psychodynamic/interpersonal training clinic. *Journal of Counseling Psychology, 62*, 184-201. http://dx.doi.org/10.1037/cou0000053

*Himelhoch, S., Mohr, D., Maxfield, J., Clayton, S., Weber, E., Medoff, D., & Dixon, L. (2011). Feasibility of telephone-based cognitive behavioral therapy targeting major depression among urban dwelling African-American people with co-occurring HIV. *Psychology, Health & Medicine, 16*, 156-165. http://dx.doi.org/10.1080/13548506.2010.534641

*Hoglend, P., Hersoug, A., Bogwald, K., Amlo, S., Marble, A., 7 Sorbye, O., … Crits-Christoph, P. (2011). Effects of transference work in the context of therapeutic alliance and quality of object relations. *Journal of Consulting and Clinical Psychology, 79*, 697-706. http://dx.doi.org/10.1037/a0024863

* Holmqvist, R., Philips, B., & Mellor-Clark, J. (2016). Client and therapist agreement about the client's problems: Associations with treatment alliance and outcome. *Psychotherapy Researcher, 26,* 399-409. http://dx.doi.org/10.1080/10503307.2015.1013160

*Hooper, K. E. (2006). Physically ill clients' optimism and hope as predictors of the psychotherapeutic alliance (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 2006-99001-077)

Horvath, A. O., & Bedi, R. P. (2002). The alliance (pp. 37-69). In J. C. Norcross. *Psychotherapy relationships that work.* New York, NY: Oxford University Press.

Horvath, A., Del Re, A. C., Flückiger, C., & Symonds, D. (2011). The alliance. In J. C. Norcross (Ed.). *Relationships that work* (pp. 25-69). New York, NY: Oxford University

Press.

Horvath, A. O., & Symonds, B. D. (1991). Relation between working alliance and outcome in psychotherapy: A meta-analysis. *Journal of Counselling Psychology, 38,* 139-149. http://dx.doi.org/10.1037/0022-0167.38.2.139

*Huang, T. C., Hill, C. E., Strauss, N., Heyman, M., & Hussain, M. (2016). Corrective relational experiences in psychodynamic-interpersonal psychotherapy: Antecedents, types, and consequences. *Journal of Counseling Psychology, 63,* 183-197. http://dx.doi.org/10.1037/cou0000132

Huedo-Medina, T. B., Sanchez-Meca, J., Marin-Martinez, F., & Botella, J. (2006). Assessing heterogeneity in meta-analysis: *Q* statistic or I^2 index? *Psychological Methods, 11,* 193-206. http://psycnet.apa.org/doi/10.1037/1082-989X.11.2.193

Hunsley, J. & Mash E. J. (2008). *A guide to assessments that work.* New York: Oxford University Press.

*Jarry, J. L. (2010). Core conflictual relationship theme-guided psychotherapy: Initial effectiveness study of a 16-session manualized approach in a sample of six patients. *Psychology and Psychotherapy: Theory, Research and Practice, 83*, 385-394. http://dx.doi.org/10.1348/147608310X486093

*Jordan, K. (2003). Relating therapeutic working alliance to therapy outcome. *Family Therapy,* 30, 95-108.

*Kahn, W. L. (1995). Patient-induced inspiration in the therapist and its relationship to the therapeutic alliance, patient-therapist similarities and patient attributes (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 1995-95022-132)

*Karlin, B. E., Walser, R. D., Yesavage, J., Zhang, A., Trockel, M., & Taylor, C. B. (2013). Effectiveness of acceptance and commitment therapy for depression: comparison among older and younger veterans. *Aging Mental Health, 17*, 555-563. doi: 10 1080/13607863 2013 789002

*Kelley, F. A. (2015). The therapy relationship with lesbian and gay clients. *Psychotherapy, 52*, 113-118. http://dx.doi.org/10.1037/a0037958

Kieffer, K. M., & MacDonald, G. (2011). Exploring factors that affect score reliability and variability in the Ways of Coping Questionnaire reliability coefficients: A meta-analytic reliability generalization study. *Journal of Individual Differences, 32*, 26-38. doi: 10.1027/1614-0001/a000031

*Kivlighan, D. M., Jr, & P. Shaughnessy (2000). Patterns of working alliance development: A typology of client's working alliance ratings. *Journal of Counseling Psychology,* 47, 362-371. http://dx.doi.org/10.1037/0022-0167.47.3.362

*Kivlighan, D. M., Jr. (2010). Changes in trainees' intention use and volunteer clients' evaluations of sessions during early skills training. *Psychotherapy: Theory, Research, Practice, Training, 47*, 198-210. http://dx.doi.org/10.1037/a0019760

*Koenig, H. G., Pearce, M., Nelson, B., Shaw, S., Robins, C., Daher, N., … King, M. B. (2016). Effects of religious vs. standard cognitive behavioural therapy on therapeutic alliance: A randomized clinical trial. *Psychotherapy Research, 26,* 365-376. http://dx.doi.org/10.1080/10503307.2015.1006156

*Kramer, U., de Roten, Y., Beretta, V., Michel, L., & Despland, J. (2009). Alliance patterns

over the course of short-term dynamic psychotherapy: The shape of productive relationships. *Psychotherapy Research, 19*, 699-706. http://dx.doi.org/10.1080/10503300902956742

*Kramer, U., Kolly, S., Berthoud, L., Keller, S., Preisig, M., Caspar, F., … Despland, J. N. (2014). Effects of motive-oriented therapeutic relationship in a ten-session general psychiatric treatment of borderline personality disorder: A randomized controlled trial. *Psychotherapy and Psychosomatics, 83*, 176-186. doi/10.1159/000358528

Krause, M., Altimir, C., & Horvath, A. (2011). Reconstructing the therapeutic alliance: Reflections on the underlying dimentions of the concept. *Clinica y Salud, 22,* 267-283. http://dx.doi.org/10.5093/cl2011v22n3a7

*Lawson, D. M., & Brossart, D. F. (2003). The relationship between counselor trainee family-of-origin structure and counseling effectiveness. *The Clinical Supervisor, 22*, 21-36. http://dx.doi.org/10.1300/J001v22n02_03

*Leibert, T. W., Smith, J. B., & Agaskar, V. R. (2011). Relationship between the working alliance and social support on counseling outcome. *Journal of Clinical Psychology, 67*, 709-719. http://dx.doi.org/10.1002/jclp.20800

*Leibert, T. W., & A. Dunne-Bryant (2015). Do common factors account for counseling outcome? *Journal of Counseling & Development, 93*, 225-235. http://dx.doi.org/10.1002/j.1556-6676.2015.00198.x

*Lewin, J. K. (2011). The importance of emotional-reflexive patterns for productive therapy: A narrative process analysis of emotion-focused and client-centered psychotherapy (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 2011-99220-388)

*Lo Coco, G., Gullo, S., Prestano, C., & Gelso, C. J. (2011). Relation of the real relationship and the working alliance to the outcome of brief psychotherapy. *Psychotherapy, 48*, 359-367. http://dx.doi.org/10.1037/a0022426

López-López, J. A., Botella, J., Sánchez-Meca, J., & Marin-Martinez,F. (2012). Alternatives for mixed-effects meta-regression models in the reliability generalization approach: A simulation study. *Journal of Educational and Behavioral Statistics, 38*, 443-469. doi: 10.3102/1076998612466142

Luborsky, L. (1976). Helping alliances in psychotherapy. In J. L. Claghorn (Ed.), *Successful psychotherapy* (pp. 92–116). New York: Brunner Mazel.

*Lukin, M. E. (1996). Effect of client motivation and counselor social influence on the development of the working alliance (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No 1997-95012-243)

*Lysaker, P. H., Davis, L., Outcalt, S. D., Gelkopf, M., & Roe, D. (2011). Therapeutic alliance in cognitive behavior therapy for schizophrenia: Association with history of sexual assault. *Cognitive Therapy and Research, 35*, 456-462. http://dx.doi.org/10.1007/s10608-010-9328-9

*Manne, S. L., Kashy, D. A., Rubin, S., Hernandez, E., & Bergman, C. (2012). Therapist and patient perceptions of alliance and progress in psychological therapy for women diagnosed with gynecological cancers. *Journal of Consulting and Clinical Psychology, 80*, 800-810. http://dx.doi.org/10.1037/a0029158

*Marcus, D. K., Kashy, D. A., & Baldwin, S. A. (2009). Studying psychotherapy using the one-with-many design: The therapeutic alliance as an exemplar. *Journal of Counseling Psychology, 56*, 537-548. http://dx.doi.org/10.1037/a0017291

*Marmarosh, C. L., Gelso, C. J., Markin, R. D., Majors, R., Mallery, C., & Choi, J. (2009).The real relationship in psychotherapy: Relationships to adult attachments, working alliance, transference, and therapy outcome. *Journal of Counseling Psychology, 56*, 337-350. http://dx.doi.org/10.1037/a0015169

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

*Matos, M., Santos, A., Goncalves, M., & Martins, C. (2009). Innovative moments and change in narrative therapy. *Psychotherapy Research, 19*, 68-80. http://dx.doi.org/10.1080/10503300802430657

Miller, M. D. (2010). Classical test theory reliability. *International encyclopedia of education (3rd ed.),* 27-30. doi 10.1016/B978-0-08-044894-7.00235-9

Minium, E. W., Clarke, R. C., & Coladarci, T. (1999). Elements of statistical reasoning (2nd ed.). New York: John Wiley & Sons, Inc

*Mohr, J. J., Fuertes, J. N., & Stracuzzi, T. I. (2015). Transference and insight in psychotherapy with gay and bisexual male clients: The role of sexual orientation identity integration. *Psychotherapy, 52*, 119-126. http://dx.doi.org/10.1037/a0036510

*Morgan, R., Luborsky, L., Crits-Christoph, P., Curtis, H., & Solomon, J. (1982). Predicting the outcomes of psychotherapy by the Penn Helping Alliance Rating Method. *Archives of General Psychiatry, 39*, 397-402. http://dx.doi.org/10.1001/archpsyc.1982.04290040013002

*Muench, J. L. (1996). "Negative indicators in psychotherapy: The assessment of client

difficulty and its relationship to therapist behavior and the working alliance (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 1996-95021-226)

*Multon, K. D., Kivlighan, D. M., Jr. & Gold, P. B. (1996). Changes in counselor adherence over the course of training. *Journal of Counseling Psychology, 43*, 356-363. http://dx.doi.org/10.1037/0022-0167.43.3.356

*Munder, T., Wilmers, F., Leonhart, R., Linster, H. W., & Barth, J. (2010). Working Alliance Inventory-Short Revised (WAI-SR): Psychometric properties in outpatients and inpatients. *Clinical Psychology & Psychotherapy, 17*, 231-239.

*Nissen-Lie, H. A., Monsen, J. T., & Ronnestad, M. H. (2010). Therapist predictors of early patient-rated working alliance: A multilevel approach. *Psychotherapy Research, 20*, 627-646. http://dx.doi.org/10.1080/10503307.2010.497633

*Nofzinger, D. M. (2003). A qualitative study of the inner experiences of therapists at different training levels (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 2003-95014-014)

Orlinsky, D. E., & Rønnestad, M. H. (2000). Ironies in the history of psychotherapy research: Rogers, Bordin, and the shape of things that came. *Journal of Clinical Psychology, 56,* 841-851. doi: 10.1002/1097-4679(200007)56:7<841::AID-JCLP3>3.0.CO;2-V

Orwin, R. G. (1983). A fail-safe N for effect size meta-analysis. *Journal of Educational Statistics, 8,* 157–159. doi: 10.3102/10769986008002157

*Osika, T. S. (1995). "The effectiveness of the early memories procedure in facilitating the

development of the working alliance (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 1996-95011-186)

*Owen, J., Imel, Z., Adelson, J., & Rodolfa, E. (2012). 'No-show': Therapist racial/ethnic disparities in client unilateral termination. *Journal of Counseling Psychology, 59*, 314-320. http://dx.doi.org/10.1037/a0027091

*Owen, J., Quirk, K., Hilsenroth, M. J., & Rodolfa, E. (2012). Working through: In-session processes that promote between-session thoughts and activities. *Journal of Counseling Psychology, 59*, 161-167. http://dx.doi.org/10.1037/a0023616

*Owen, J. J., Tao, K., Leach, M. M., & Rodolfa, E. (2011). "Clients' perceptions of their psychotherapists' multicultural orientation." *Psychotherapy, 48*, 274-282. http://dx.doi.org/10.1037/a0022065

*Owen, J., Tao, K., & Rodolfa, E. (2010). Microaggressions and women in short-term psychotherapy: Initial evidence. *The Counseling Psychologist, 38*, 923-946. http://dx.doi.org/10.1177/0011000010376093

*Patterson, C.L., Anderson, T., & Wei, C. (2014). Clients' pretreatment role expectations, the therapeutic alliance, and clinical outcomes in outpatient therapy. *Journal of Clinical Psychology, 70,* 673-680. http://doi: 10.1002/jclp.22054

*Patterson, C. L., Uhlin, B., & Anderson, T. (2008). Clients' pretreatment counseling expectations as predictors of the working alliance. *Journal of Counseling Psychology, 55*, 528-534. http://dx.doi.org/10.1037/a0013289

*Pinto-Coelho, K. G., Hill, C. E. & Kivlighan, D. M. Jr. (2016). Therapist self-disclosure in

psychodynamic psychotherapy: A mixed methods investigation. *Counselling Psychology Quarterly, 29,* 29-52. http://dx.doi.org/10.1080/09515070.2015.1072496

*Pos, A. E. (2006). Experiential treatment for depression: A test of the experiential theory of change, differential effectiveness, and predictors of maintenance of gains (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 2007-99012-040)

*Preschl, B., Maercker, A., & Wagner, B. (2011). The working alliance in a randomized controlled trial comparing online with face-to-face cognitive-behavioral therapy for depression. *BMC Psychiatry, 11*, 189. http://dx.doi.org/10.1186/1471-244X-11-189

*Price, P. B., & E. E. Jones (1998). Examining the alliance using the Psychotherapy Process Q-Set. *Psychotherapy: Theory, Research, Practice, Training, 35*, 392-404. http://dx.doi.org/10.1037/h0087654

*Rader, J., & L. A. Gilbert (2005). The egalitarian relationship in feminist therapy. *Psychology of Women Quarterly, 29*, 427-435. http://dx.doi.org/10.1111/j.1471-6402.2005.00243.x

*Razzhavaikina, T. I. (2007). Mandatory counseling: A mixed methods study of factors that contribute to the development of the working alliance (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 2008-99230-500)

Reinhardt, B. (1996). Factors affecting coefficient alpha: A mini Monte Carlo study. In B. Thompson (Ed.), *Advances in social science methodology* (Vol. 4, pp. 3-20). Greenwich,CT: JAI Press.

*Reis, S., & B. F. Grenyer (2004). Fearful Attachment, Working Alliance and Treatment Response for Individuals with Major Depression. *Clinical Psychology and Psychotherapy, 11*, 414-424. http://dx.doi.org/10.1002/cpp.428

*Renner, F., Jarrett, R. B., Vittengl, J. R., Barrett, M. S., Clark, L. A., & Thase, M. E. (2012). Interpersonal problems as predictors of therapeutic alliance and symptom improvement in cognitive therapy for depression. *Journal of Affective Disorders, 138*, 458-467. http://dx.doi.org/10.1016/j.jad.2011.12.044

*Richards, P., & S. Simpson (2015). Beyond the therapeutic hour: An exploratory pilot study of using technology to enhance alliance and engagement within face-to-face psychotherapy. *British Journal of Guidance & Counselling, 43*, 57-93. http://dx.doi.org/10.1080/03069885.2014.936824

*Rozmarin, E., Muran, J. C., Safran, J., Gorman, B., Nagy, J., & Winston, A. (2008). Subjective and intersubjective analyses of the therapeutic alliance in a brief relational therapy. *American Journal Psychotherapy, 62*, 313-328.

*Ryum, T., Stiles, T. C., Svartberg, M., & McCullough, L. (2010). The role of transference work, the therapeutic alliance, and their interaction in reducing interpersonal problems among psychotherapy patients with Cluster C personality disorders. *Psychotherapy: Theory, Research, Practice, Training, 47*, 442-453. http://dx.doi.org/10.1037/a0021183

Safran, J. D., & Muran, J. C. (2000). *Negotiating the therapeutic alliance: a relational treatment guide.* New York: Guilford Press

*Santiago, N. J., Klein, D. N., Vivian, D., Vocisano, C., Dowling, F., Arnow, B. A., … Keller, M. B.

(2002). Pretreatment correlates of the therapeutic alliance in the chronically depressed. *Journal of Contemporary Psychotherapy, 32*, 281-290. http://dx.doi.org/10.1023/A:1020524910971

*Sauer, E. M., Anderson, M. Z., Gormley, B., Richmond, C. J., & Preacco, L. (2010). Client attachment orientations, working alliances, and responses to therapy: A psychology training clinic study. *Psychotherapy Research, 20*, 702-711. http://dx.doi.org/10.1080/10503307.2010.518635

*Saunders, S. M., Howard, K., & Orlinsky, D. E. (1989). The Therapeutic Bond Scales: Psychometric characteristics and relationship to treatment effectiveness. *Psychological Assessment: A Journal of Consulting and Clinical Psychology, 1*, 323-330.

*Saunders, S. M. (2000). Examining the relationship between the therapeutic bond and the phases of treatment outcome. *Psychotherapy: Theory, Research, Practice, Training, 37*, 206-218.

Schmidt, F. L., & Hunter, J. E. (1977). Development of a general solution to the problem of validity generalization. *Journal of Applied Psychology, 62*, 529-540. doi: 10.1037/0021-9010.62.5.529

Sharf, J., Primavera, L. H., & Diener, M. J. (2010). Dropout and therapeutic alliance: A meta-analysis of adult individual psychotherapy. *Psychotherapy: Theory, Research, Practice, Training, 47,* 637-645. doi: 10.1037/a0021175

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin, 86*, 420-428. http://psycnet.apa.org/doi/10.1037/0033-2909.86.2.420

*Spinhoven, P., Giesen-Bloo, J., van Dyck, R., Kooiman, K., & Arntz, A. (2007). The

therapeutic alliance in schema-focused therapy and transference-focused psychotherapy for borderline personality disorder. *Journal of Consulting and Clinical Psychology, 75*, 104-115. http://dx.doi.org/10.1037/0022-006X.75.1.104

*Stapor, B. S. (1999). The effect of working alliance on termination status at a college counseling center (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 1999-95009-170)

*Stiles, W. B., Agnew-Davies, R., Barkham, M., Culverwell, A., Goldfried, M. R., & Halstead, J., … Shapiro, D. A. (2002). Convergent validy of the Agnew Relationship Measure and the Working Alliance Inventory. *Psychological Assessment, 14*, 209-220. http://dx.doi.org/10.1037/1040-3590.14.2.209

*Stiles-Shields, C., Kwasny, M. J., Cai, X., & Mohr, D. C. (2014). Therapeutic alliance in face-to-face and telephone-administered cognitive behavioral therapy. *Journal of Consulting and Clinical Psychology, 82*, 349-354. http://dx.doi.org/10.1037/a0035554

*Stough, R. (1999). The impact of the reflecting team on the working alliance and therapy outcome (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 2000-95010-192)

*Stracuzzi, T. I., et al. (2011). Gay and bisexual male clients' perceptions of counseling: The role of perceived sexual orientation similarity and counselor universal-diverse orientation. *Journal of Counseling Psychology, 58*, 299-309. http://dx.doi.org/10.1037/a0023603

*Strauss, J. L., Hayes, A. M., Johnson, S. L., Newman, C. F., Brown, G. K., Barber, J. P., … Beck,

A. T. (2006). Early alliance, alliance ruptures, and symptom change in a nonrandomized trial of cognitive therapy for avoidant and obsessive-compulsive personality disorders. *Journal of Consulting and Clinical Psychology, 74*, 337-345. http://dx.doi.org/10.1037/0022-006X.74.2.337

*Stringer, J. V., Levitt, H. M., Berman, J. S., & Mathews, S. S. (2010). A study of silent disengagement and distressing emotion in psychotherapy. *Psychotherapy Research, 20*, 495-510. http://dx.doi.org/10.1080/10503301003754515

*Taylor, P. J., Rietzschel, J., Danquah, A., & Berry, K. (2015). The role of attachment style attachment to therapist, and working alliance in response to psychological therapy. *Psychology and Psychotherapy: Theory, Research and Practice, 88,* 240-253. http://doi:10.1111/papt.12045

Therrien, Z., & Hunsley, J. (2013). Assessment of anxiety in older adults: A reliability generalization meta-analysis of commonly used measures. *Clinical Gerontologist 36*, 171-194. doi:10.1080/07317115.2013.767871

Thompson, R. (1999, January). *Reliability generalization: An important meta-analytic method, because it is incorrect to say, "The test is unreliable."* Paper presented at the Annual Meeting of the Southwest Educational Research Association (San Antonio, TX, U.S.A.). Retrieved from http://files.eric.ed.gov/fulltext/ED434121.pdf

*Tichenor, V., & C. E. Hill (1989). A comparison of six measures of working alliance. *Psychotherapy: Theory, Research, Practice, Training, 26*, 195-199. http://dx.doi.org/10.1037/h0085419

*Tokar, D. M., Hardin, S. I., Adams, E. M., & Brandel, I. W. (1996). Clients' expectations about

counseling and perceptions of the working alliance. *Journal of College Student Psychotherapy, 11*, 9-26. http://dx.doi.org/10.1300/J035v11n02_03

Vacha-Haase, T. (1998). Reliability generalization: Exploring variance in measurement error affecting score reliability across studies. *Educational and Psychological Measurement*, *58,* 6-20. doi: 10.1177/0013164498058001002

Vacha-Haase, T., Kogan, L. R. & Thompson, B. (2000). Sample composition and variabilities in published studies versus those in test manuals: Validity of score reliability inductions. *Educational and Psychological Measurement*, *60*, 509–552. doi: 10.1177/00131640021970682

Vacha-Haase, T., & Thompson, B. (2011). Score reliability: A retrospective look back at 12 years of reliability generalization studies. *, 44,* 159-168. doi: 10.1177/0748175611409845

* Vogel, P. A., Hansen, B., Stiles, T. C., & Gotestam, K. G. (2006). Treatment motivation, treatment expectancy, and helping alliance as predictors of outcome in cognitive behavioral treatment of OCD. *Journal of Behavior Therapy and Experimental Psychiatry, 37*, 247-255. http://dx.doi.org/10.1016/j.jbtep.2005.12.001

*Vogel, P. A., Launes, G., Moen, E. M., Solem, S., Hansen, B., Haland, A. T., & Himle, J. A. (2012). Videoconference- and cell phone-based cognitive-behavioral therapy of obsessive-compulsive disorder: A case series. *Journal of Anxiety Disorders, 26*, 158-164. http://dx.doi.org/10.1016/j.janxdis.2011.10.009

*Vogel, P. A., Solem, S., Hagen, K., Moen, E. M., Launes, G., Haland, A. T., … Himle, J. A.

(2014). A pilot randomized controlled trial of videoconference-assisted treatment for obsessive-compulsive disorder. *Behaviour Research and Therapy, 63*, 162-168. http://dx.doi.org/10.1016/j.brat.2014.10.007

Von Hippel, P. T. (2015). The heterogeneity statistic I^2 can be biased in small meta-analyses. *BMC Medical Research Methodology*, *15*, 35. http://doi.org/10.1186/s12874-015-0024-z

*Wagner, B., Brand, J., Schulz, W., Knaevelsrud, C. (2012). Online working alliance predicts treatment outcome for posttraumatic stress symptoms in Arab war-traumatized patients. *Depression and Anxiety, 29*, 646-651. http://dx.doi.org/10.1002/da.21962

*Watson, J. C., & S. M. Geller (2005). The relation among the relationship conditions, working alliance, and outcome in both process-experiential and cognitive-behavioral psychotherapy. *Psychotherapy Research, 15*, 25-33. http://dx.doi.org/10.1080/10503300512331327010

*Westmacott, R., Hunsley, J., Best, M., Rumstein-McKean, O., & Schindler, D. (2010). Client and therapist views of contextual factors related to termination from psychotherapy: A comparison between unilateral and mutual terminators. *Psychotherapy Research, 20*, 423-435. http://dx.doi.org/10.1080/10503301003645796

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

*Whelton, W. J., Paulson, B., & Marusiak, C. W. (2007). Self-criticism and the therapeutic

relationship. *Counselling Psychology Quarterly, 20*, 135-148. http://dx.doi.org/10.1080/09515070701412423

Wilkinson, L., & Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*, 594-604. http://dx.doi.org/10.1037/0003-066X.54.8.594

*Wongpakaran, T., & N. Wongpakaran (2012). How the interpersonal and attachment styles of therapists impact upon the therapeutic alliance and therapeutic outcomes. *Journal of the Medical Association Thailand, 95*, 1583-1592.

*Xu, H., & Tracey, T. J. G. (2015). Reciprocal influence model of working alliance and therapeutic outcome over individual therapy course. *Journal of Counseling Psychology, 62,* 351-359. http://dx.doi.org/10.1037/cou0000089

Yin, P., & Fan, X. (2000). Assessing the reliability of Beck Depression Inventory scores: Reliability generalization across studies. *Educational and Psychological Measurement, 60*, 201-223. doi: 10.1177/00131640021970466

*Zickgraf, H. F., et al. (2015). Interpersonal factors are associated with lower therapist adherence in cognitive-behavioural therapy for panic disorder. *Clinical Psychology and Psychotherapy*. Advance online publication. doi: 10.1002/cpp.1955

*Zilcha-Mano, S., & Errázuriz, P. (2015). One size does not fit all: Examining heterogeneity and identifying moderators of the alliance-outcome association. *Journal of Counseling Psychology, 62,* 579-591. http://dx.doi.org/10.1037/cou0000103

*Zilcha-Mano, S., Muran, C. J., Huang, C., Eubanks, C. F., & Safran, J. D. (2016). The

relationship between alliance and outcome: Analysis of a two-person perspective alliance and session outcome. *Journal of Consulting and Clinical Psychology, 84,* 484-496. http://dx.doi.org/10.1037/ccp0000058

Table 3

List of Measure-related Search Terms

	Measure as identified in Elvins and Green (2008)	Measure's full name used in search	Abbreviations/Alternate name of measure used in search
1	Barrett-Lennard's Relationship Inventory	Barrett-Lennard's Relationship Inventory	BLRI
2	Counselling Evaluation Inventory	Counselling Evaluation Inventory	CEI
3	Therapy Session Report Scales	Therapy Session Report Scale	Therapy Session Report; TSRS; TRS
4	The Counselor Rating Form	The Counselor Rating Form	CRF
5	The Penn Alliance Scales	The Penn Helping Alliance Scales	The Helping Alliance Questionnaire; The Penn; HAq
6	Vanderbilt Scales	Vanderbilt Therapeutic Scales	Vanderbilt Therapeutic Alliance Scale; VTAS; Vanderbilt Psychotherapy Process Scale; VPPS
7	Toronto Scales	Therapeutic Alliance Rating Scales	TARS
8	Menninger Alliance Rating Scale or Collaboration Scale	Menninger Alliance Rating Scale or Collaboration Scale	Menninger
9	Psychotherapy Status Report	Psychotherapy Status Report	PSR
10	Patient Collaboration Scale	Patient Collaboration Scale	PCS
11	California Scales	California Psychotherapy Alliance Scales	CALPAS
12	Therapeutic bond Scales	Therapeutic Bond Scale	TBS
13	Working Alliance Inventory	Working Alliance Inventory	WAI
14	Treatment Alliance Scales	Treatment Alliance Scales	TAS

15	Adapted psychotherapy Process Inventory	Adapted Psychotherapy Process Inventory	PPI
16	Helping Alliance Scales	Helping Alliance Scales	Helping Alliance Scale; HAS
17	Empathy and Understanding Questionnaire	Empathy and Understanding Questionnaire	EUQ
18	Barriers to Treatment Participation Scale	Barriers to Treatment Participation Scale	BTPS
19	Therapist Alliance Focus Scale	Therapist Alliance Focus Scale	TAFS
20	Agnew Relationship Measure	Agnew Relationship Measure	ARM
21	Kim Alliance Scale	Kim Alliance Scale	KAS
22	The Therapy Process Observational Coding System — Alliance Scale	The Therapy Process Observational Coding System — Alliance Scale	TPOCS
23	Scale to Assess Therapeutic Relationship	Scale to Assess Therapeutic Relationship	STAR

Table 4

Descriptive Statistics for the Mean Reliability Coefficients

			95% confidence interval						
Measure	*k*	Mean α	Lower	Upper	Min.	Max.	*Q*	I^2	Orwin's *N*
ARM-28-C†	2	.90	.83	.94	.87	.91	.56	.00	n/a††
ARM-5-C†	4	.78	.67	.86	.64	.96	9.70*	69.07	3
ARM-5-T†	2	.87	.83	.90	.86	.88	.30	.00	n/a††
CALPAS-24-C	8	.88	.83	.91	.70	.93	29.05*	75.91	13
CALPAS-O†	3	.96	.93	.97	.90	.96	1.04	.00	10
CRF-S-C	6	.94	.93	.96	.90	.96	8.00	37.54	18
HAq-I-C	8	.86	.85	.88	.83	.89	5.1	.00	12
HAq-II-C†	4	.84	.73	.91	.71	.91	26.06*	88.49	5
HAq-II-T†	2	.89	.80	.95	.86	.93	3.48	71.23	n/a††
HAq-O†	2	.95	.90	.98	.93	.96	.32	.00	n/a††
TBS-C†	2	.74	.48	.88	.62	.82	10.30*	90.29	n/a††
WAI-C	37	.93	.91	.94	.67	.96	326.86*	88.97	89
WAI-O†	2	.97	.85	.99	.89	.98	2.18	54.11	n/a††
WAI-S-C	49	.90	.89	.91	.76	.99	216.64*	77.84	90
WAI-SR-C	17	.93	.91	.94	.88	.97	164.04*	90.25	47
WAI-S-T	21	.90	.88	.92	.78	.95	100.05*	80.01	35*
WAI-T	18	.92	.88	.95	.59	.97	227.57*	92.53	40

Note. k = number of reliability coefficients used in analysis; Min. = Minimum reliability coefficient reported in literature; Max. = Maximum reliability coefficient reported in literature; See Table 1 for alliance measure full name *$p \leq 0.05$ †Did not meet minimum number of 5 studies need to be included in moderator analysis. ††Did not meet minimum number of 3 studies needed to run publication bias for CMA program.

Table 5

Meta-Regression for Mean Reliability Coefficients and Sample Characteristics for Significant Continuous Moderators

				95% Confidence Interval	
Measure, Covariate, & *k*	***b***	***z***	***p***	**Lower**	**Upper**
CALPAS-24-C					
SD of client measure *(k=5)*	.069	3.69	.000	.032	.107
WAI-C					
Year of publication *(k=37)*	-.019	-2.17	.03	-.035	-.002
WAI-S-C					
SD of client measure *(k=31)*	.03	2.15	.032	.002	.05
Mean age of therapist *(k=20)*	-.018	-2.60	.009	-.032	-.005
Mean years of experience *(k=20)*	-.03	-2.11	.04	-.057	-.002
WAI-SR-C					
Mean client of measure *(k=13)*	.018	4.42	.000	.009	.025
SD of client measure *(k=13)*	.061	5.16	.000	.038	.085
WAI-S-T					
Mean age of therapist *(k=10)*	-.023	-3.2	.001	-.036	-.009
Percentage of female clients *(k=19)*	-.005	-2.03	.04	-.009	-.000
Client's sample size *(k=21)*	-.001	-2.4	.017	-.003	-.000
Therapist's sample size *(k=13)*	.004	1.96	.049	.000	.007
WAI-T					
Mean age of therapist *(k=7)*	.019	2.34	.019	.003	.034

SD age of therapist *(k=4)*	.138	2.97	.003	.047	.229
Therapist's sample size *(k=17)*	-.004	-2.49	-.013	-.008	-.001

Note. *SD* = standard deviation

Table 6

Analysis of Variance between Mean Reliability Coefficients and Sample Characteristics

Measure	Type of Population	Type of Treatment	Type of Mental Disorder/ Problem	Language of Measure	Mode of Treatment	Type of Report
CALPAS-24-C	.595†	4.585†	4.274†	.772†	--	--
WAI-C	3.495†	1.39†	7.724*	1.655†	.134†	.002
WAI-S-C	3.459	12.815*	3.625	.336†	.703†	1.815†
WAI-SR-C	.0015†	2.374†	.0133†	.0118†	--	.0011†
WAI-S-T	2.336†	.044	4.686†	.271†	2.138†	.469†
WAI-T	11.938†*	8.41†	2.018†	.343†	.837†	.468

† Estimates of tau-squared pooled

*$p \leq 0.05$

Blank cells indicate reliability estimates for only one category within the moderator variable were obtained, and therefore, the Q-value between the subgroups could not be calculated

Figure 1 Flow Chart of Study Selection

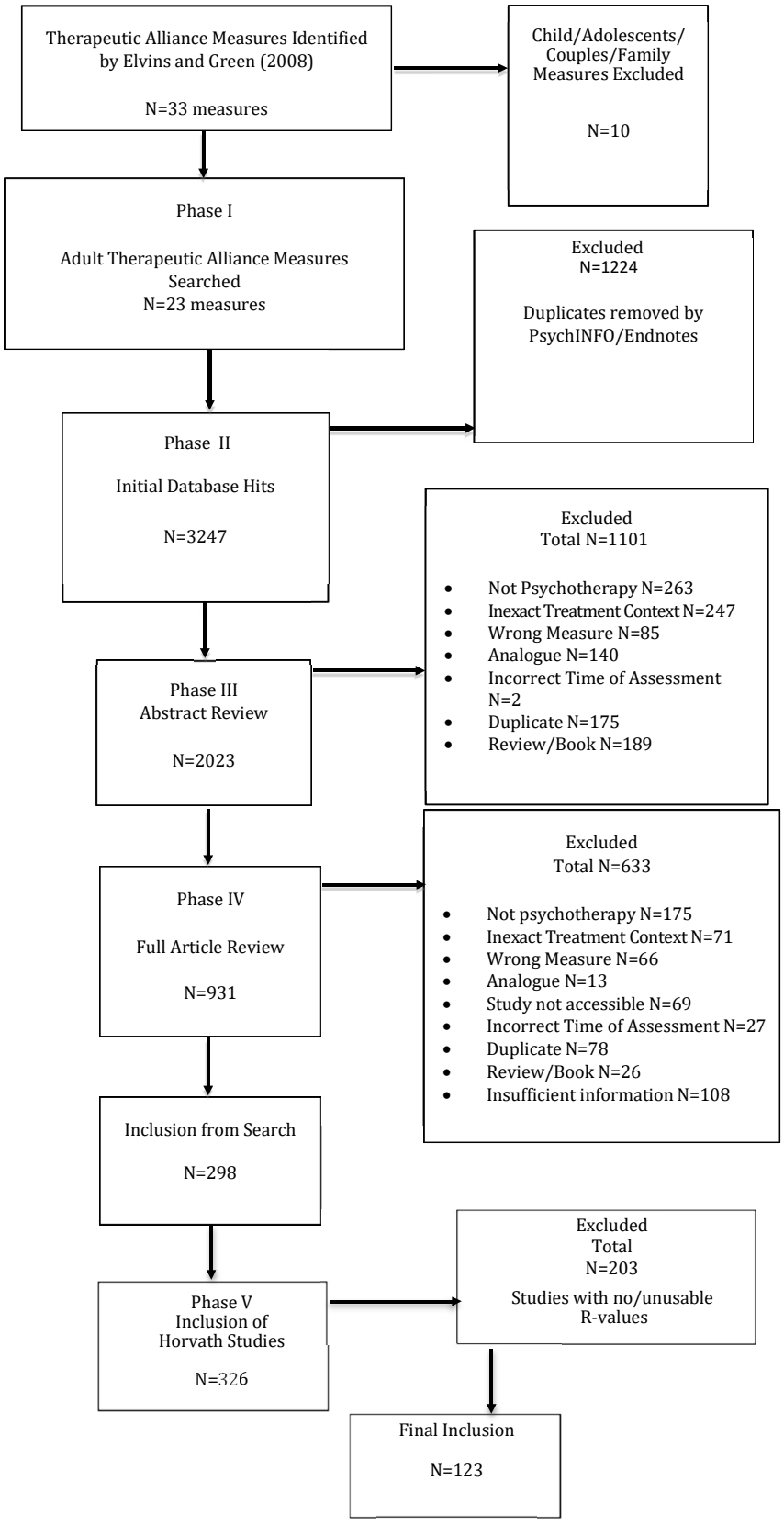


Figure 2 Funnel Plot for CALPAS-24-C

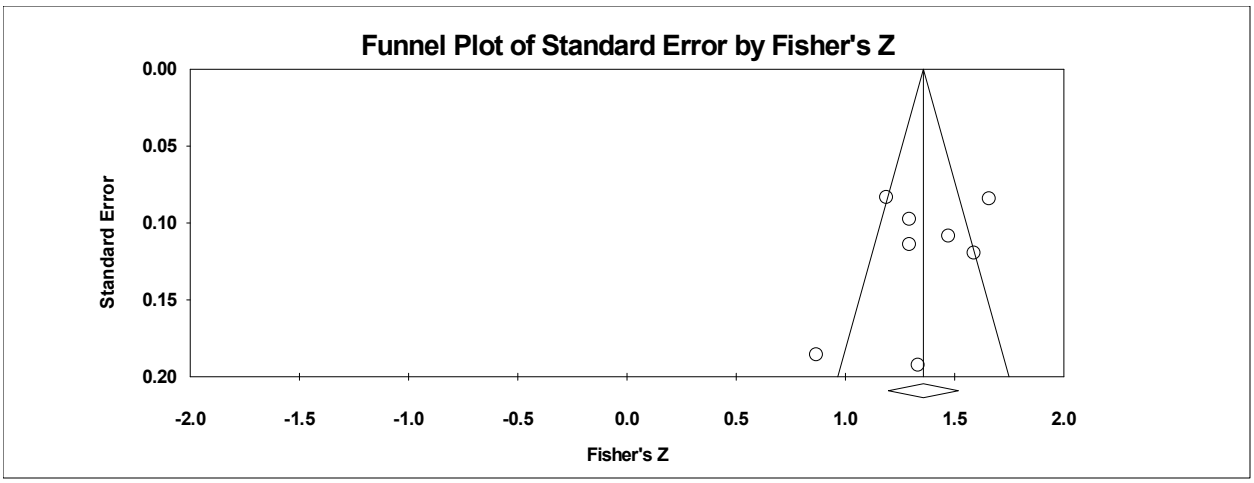


Figure 3 Funnel Plot for WAI-T

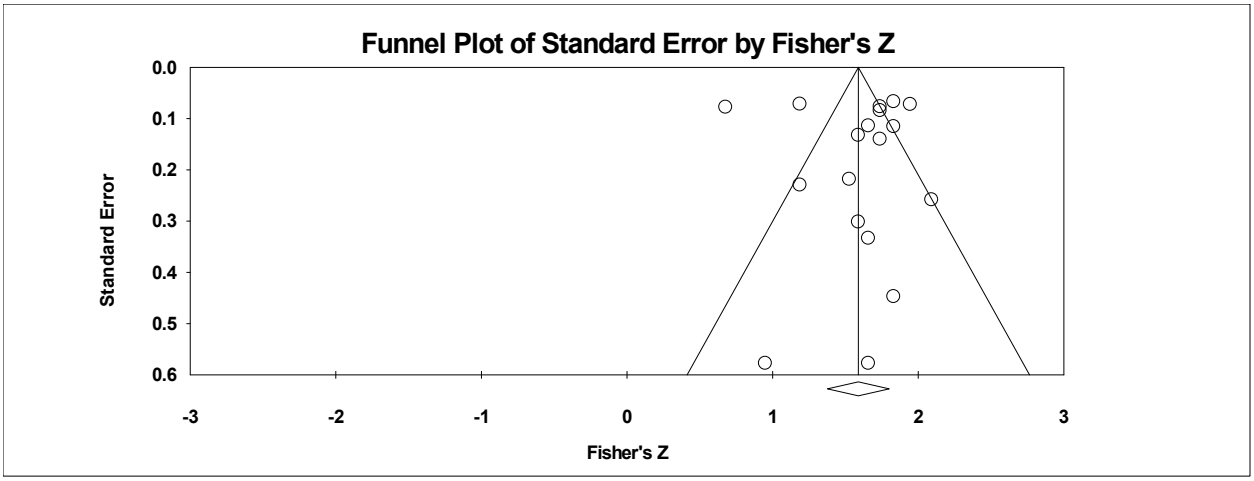


Figure 4 Funnel Plot for WAI-S-T

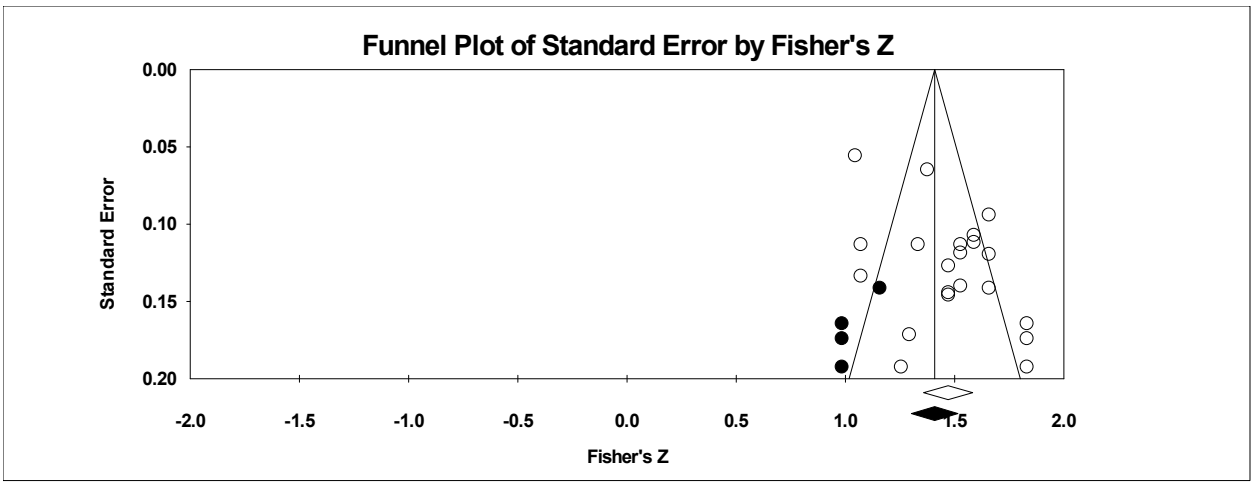


Figure 5 Funnel Plot for WAI-C

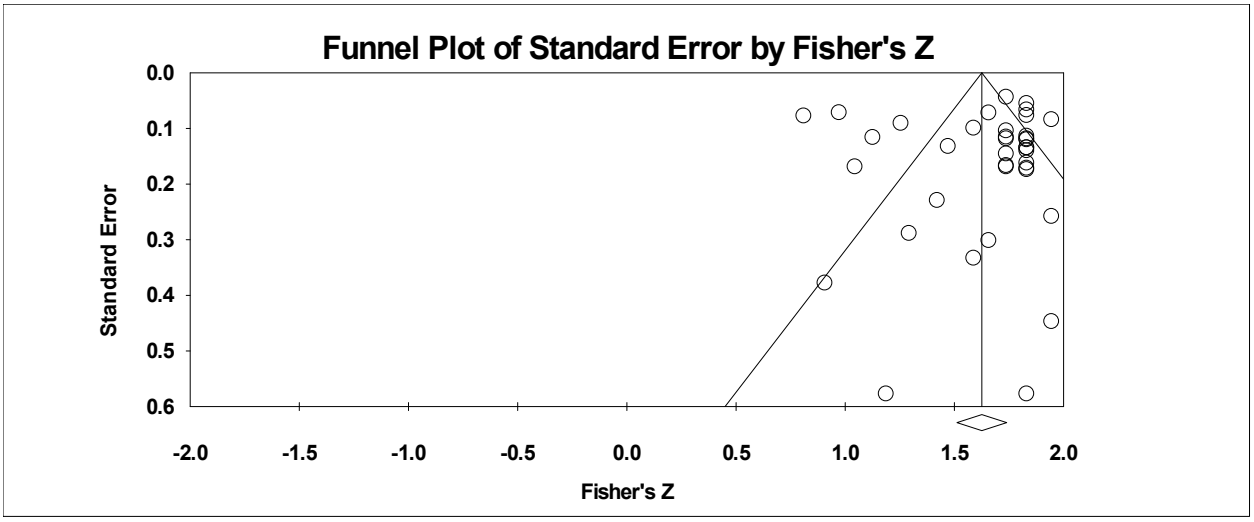


Figure 6 Funnel Plot for WAI-S-C

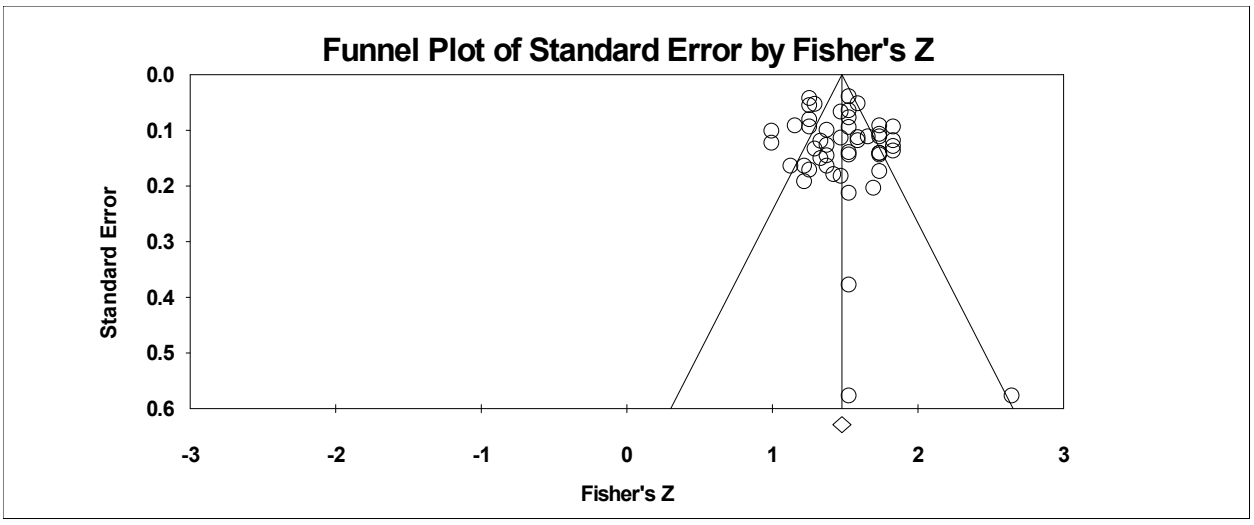
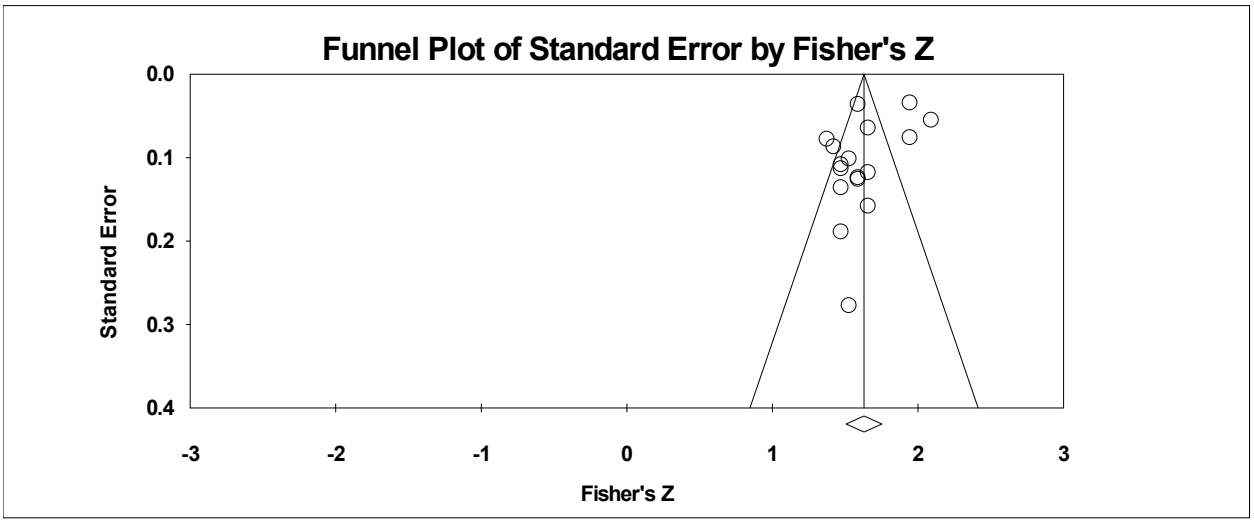


Figure 7 Funnel Plot for WAI-SR-C



Appendix A

Lists of studies included in RG and Horvath et al., 2011 study.

Studies obtained in RG search; included in both RG and Horvath	Studies obtained in RG search; included in Horvath but not in RG	Studies not obtained in RG search; retrieved from Horvath *(* Studies included into RG after phase V)*
Cloitre et al. (2004)	Fakhoury et al. (2007)	Constantino et al. (2005)*
Coleman (2006)	Dearing et al. (2005)	Deu et al. (2009)
Davis et al. (2007)	Florsheim,et al. (2000)	Dorsch et al. (2002)
Knaevelsrud et al. (2007)	Forbes et al. (2008)	Eaton et al. (1988)
Mallinckrodt (1993)	Gallop et al. (1994)	Freitas (2001)
Marmarosh et al. (2009	Hervé et al. (2008)	Forman (1990)
Missirlian et al. (2005)	Howard et al. (2006)	Greenberg et al. (1982)
Muran et al. (2009)	Ilgen et al. (2006a)	Greenberg et al. (2002)
Pos (2007)	Ilgen et al. (2006b)	Hopkins (1988)*
Reis et al. (2004)	Jumes (1995)	Jacob (2003)
Stevens et al. (2007)	Kabuth et al. (2005)	Katz (1999)*
Spinhoven et al. (2007)	Karver et al. (2008)	Kivlighan et al. (2000)*
Wettersten et al. (2005)	Kelly et al. (2009)	Kivlighan et al. (1995)*
Hayes et al. (2007)	Kokotovic et al. (1990)	Klein et al. (2003)
Baldwin (2007)	Meier et al. (2006 b)	Mallinckrodt (1996) *
Biscoglio (2005)	Meier et al. (2006 a	Multon et al. (2001)
Busseri et al. (2003)	Pos, et al. (2009)	Moseley (1983)
Dunn et al. (2006)	Ramnerö et al. (2007)	Pugh (1991)
Gaiton (2004)	Solomon et al. (1995)	Rogers et al. (2008)
Gaston et al. (1991)	Tyron et al. (1993)	Santiago et al. (2005)
Tichenor (1989)	Tyron et al. (1995)	Schönberger et al. (2006c)
Bethea et al. (2008)	Schönberger et al. (2006a)	Sexton (1996)
de Roten et al. (2004)	Schönberger et al. (2006b)	Vronmans (2007)
Gerstley et al. (1989)	Schönberger et al. (2007)	Wettersten (2000)
Kramer et al. (2009)	Botella (2008)	Adler (1988)
Morgan et al. (1982)	Burns et al. (2007)	Arnow et al. (2003)
O'Malley et al. (1983)	Gaston et al. (1994)	Bieschke et al. (1995)
Marziali et al. (1981)	Geiser et al. (2002)	Brotman (2004)
Stiles et al. (2004)	Hatcher et al. (1996)	Busseri et al. (2004)
Rounsaville et al. (1987)	Sherer et al. (2007)	Castonguay et al. (1996)
Emmerling et al. (2009)	Trepka et al. (2004)	Chilly (2004)*
	Bassler, et al. (1995)	Barber et al. (2006)
	Gunderson et al. (1997)	Barber et al. (1999)
	Kramer et al. (2008)	Barber et al. (2001)
	Mohl et al. (1991)	Barber et al. (2008)*
	Schauenburg et al. (2005)	Barber et al. (2000)*
	Schleussner (2005)	Castonguay et al. (1996)*
	Konzag et al. (2004)	Ferleger (1993)*
	Gomes-Swartz (1978)	Gaston et al. (1998)
	Hawley et al. (2006)	Geider (1997)

	Loneck et al. (2002)	Gunther (1991)*
	Klee et al. (1990)	Gutfreund (1992)
	Windholtz et al. (1988)	Hardy et al. (2001)
	Horvath (1981)	Card (1991)
	Irelan (2004)	Marmar et al. (1989 b)
		Marmar et al. (1989 a)*
		Muran et al. (1995)
		Riley (1992)*
		Strauss et al. (2006)*
		Strauss (2001)
		Tunis, et al. (1995)
		Yeomans et al. (1994)*
		Cislo (1998)
		Crits-Cristoph et al. (1988)
		Feeley (1993)*
		Janecke (2003)
		Johansson et al. (2006)
		Luborsky et al. (1985)*
		Luborsky et al. (1983)*
		Marshel (1986)
		Pyne (1991)
		Tyron et al. (1990)
		Vogel et al. (2006)*
		Van et al. (2008)*
		Hartley et al. (1983)
		Krupnick et al. (1994)
		Krupnick et al. (1996)
		Meyer et al. (2002)*
		Zuroff et al. (2000)
		Marziali (1984)*
		Marziali et al. (1999)
		Horowitz et al. (1984)*
		Kolden (1996)*
		Reiner (1987)
		Saunders (2000)*
		Saunders et al. (1989)*
		Zuroff et al. (2006)*
Exclusion of studies obtained from Horvath not in RG search Insufficient information *n*=8 Wrong measure *n*=4 Not individual adult psychotherapy *n*= 2 Study not accessible n=20 Duplicate *n*= 8		

Not psychotherapy *n*=6		

NOTE: Dissertations obtained in original search were retained in place of the published articles obtained from Horvath's studies as the dissertations reported more sample characteristics needed for coding. These articles were still counted for under the exclusion criteria Dissertation. Articles included in Horvath et al., 2011 study that used alliance measures other than those listed in the Elvins & Green's 2008 study were excluded from this RG.

Appendix B

Coding manual for RG

Study's characteristics for moderator analysis	Specifications on what to code	Codes for data entry
A: Study title	Write out the first half of Study's title	
B: Author's name	Write author's name in full (Last name first)	
C: Number of sample coded from article	Rank the number of samples being recorded for the article. Each sample will be coded on a separate row. (e.g., Study 1, row 1, sample 2; Study 1, row 2, sample 2)	Code in rank order 1, 2, 3, etc…
D: Year of Publication	Indicate the year the article was published **NOTE***: if there if a discrepancy b/w Endnotes and Article code articles date*	Code as continuous variable
E: Language of article	Indicate language of article	0= English 1=French
F: Type of report	Indicate the type of article (e.g., journal article)	0=journal article 1=dissertation
G: Language of measure	Indicate language of measure **NOTE**: *If not clear if the measure was translated (i.e., specifically indicated) code as 999*	0=English 1=other (i.e., translated version)
H: Alliance informant (therapist)	Indicate if assessed the alliance	0=if therapist **did not** assess alliance in study 1=if therapist **did** assess alliance
I: Alliance informant (client)	Indicate if client assessed the alliance	0=client **did not** assess alliance in study 1=if client **did** assess alliance
J: Alliance informant (observer)	Indicate if observer assessed the alliance	0=if observer **did not** assess alliance in study 1=if observer **did** assess alliance
	Indicate what type of population received treatment.	0=clinical sample (i.e., inpatient & outpatient from hospital)

K: Type of population	**NOTE:** *This may be reported as the location the clients were recruited from*	1=community (i.e., university training centers, community centers, medical centers, veteran centers) 2=university student sample (i.e., if sample is entirely comprised of college or university students) 3=mixed sample 4=private practice 5=corrections 999=missing data
L: Type of mental disorder(s) or problem(s)	Indicate specific mental disorder(s) or problem(s) treated **NOTE***: Diagnostic criteria may be from the(DSM-III, DSM-IV, & DSM-IV-R) depending on date of publication* *Criteria for type of mental disorder classification for this VG was derived, in part, from the DSM-IV-R criteria .*	0=depression or other mood disorders (e.g., bipolar, dysthymia) 1=anxiety disorders (e.g., social, GAD, PTSD) 2=psychotic disorders (e.g., schizophrenia, schizoaffective disorder) 3=personality disorders 4=eating disorders 5= health-related problems (e.g., chronic pain) 6=samples with multiple disorders 7=other (e.g., substance abuse, relationship difficulties not indicated as PD, etc…) 999=missing data
M: Type of treatment provided	Indicate what type of treatment was provided to client	0=family of cognitive behavioural therapies 1=family of experiential therapies 2=family of psychodynamic therapies 3=family of interpersonal therapies 4=combined treatments (e.g., medication plus CBT) 5=mixed sample of different psychotherapies 6=integrative(all therapists indicate that they are integrative) 7=other 999=missing data
N: Mode of treatment	Indicate manner in which treatment was provided	0=face to face 1=distant (e.g., texting, email, videoconferencing, phone) 2=mixed

O, P & Q: Sample size	Indicate sample size (i.e., o) number of clients the therapist(s) rated alliance for; p) number of clients; q) number of clients observer(s) rated alliance on for observer column **NOTE**: *if multiple sample sizes reported code the sample size that corresponds to the alpha value extracted from study (e.g., r-value is from third therapy session, take sample size that corresponds closest to this time period)*	Code as continuous variable 999=missing data
R: Number of Therapists	Indicate the number of therapists that completed alliance measures	Code as a continuous variable 999=missing data
S: Number of Observers	Indicate the number of observers that completed alliance measures	Code as a continuous variable 999=missing data
T: Mean age of Therapists	Indicate the mean age of the study's therapists **NOTE***: Combined means if reported for multiple groups (e.g., therapists for the completers and dropouts) in the cases where the alpha value was reported as a combined value. Use equations provided to calculate*	Code as a continuous variable (report to two decimal points) 999=missing data
U: Standard deviation of age of Therapists	Indicate the standard deviation of the therapists' age **NOTE***: Combined standard deviations if reported for multiple groups (e.g., therapists for the completers and dropouts) in the cases where the alpha value was reported as a combined value. Use equations provided to calculate*	Code as a continuous variable (report to two decimal points) 999=missing data
V: Percentage of female Therapists	Indicate the percentage of female therapists of the sample	Code as a continuous variable (report to two decimal points) 999=missing data

W: Mean of years of experience of therapists	Indicate the mean of the years of experience of the therapists	Code as a continuous variable (report to two decimal points) 999=missing data
X: Standard deviation of the years of experience of therapists	Indicated the SD of the years of experience of the therapists	Code as a continuous variable (report to two decimal points) 999=missing data
Y: Mean age of Clients	Indicate the mean age of the study's clients **NOTE***: Combined means if reported for multiple groups (e.g., completers and dropouts) in the cases where the alpha value was reported as a combined value. Use equations provided to calculate*	Code as a continuous variable (report to two decimal points) 999=missing data
Z: Standard deviation of age of client	Indicate the standard deviation of the clients' age **NOTE***: Combined standard deviations if reported for multiple groups (e.g., completers and dropouts) in the cases where the alpha value was reported as a combined value. Use equations provided to calculate*	Code as a continuous variable (report to two decimal points) 999=missing data
AA: Percentage of female Clients	Indicate the percentage of female clients of the sample	Code as a continuous variable 999=missing data
AB: Name of measure (WAI) Working Alliance Inventory	Indicate what measure is being coded. Include the version of that measure	0=**did not** use WAI 1=did use WAI
AC: Name of measure (WAI-S) Working Alliance Inventory Short	Indicate what measure is being coded	0=**did not** use WAI-S 1=did use WAI-S
AD: Name of measure (WAI-SR) Working Alliance Inventory Short Revised	Indicate what measure is being coded	0=**did not** use WAI-SR 1=did use WAI-SR
AE: Name of measure (CRF-S) Counselor Rating Form (short)	Indicate what measure is being coded	0=**did not** use CRF-S 1=did use CRF-S
AF: Name of measure (CALPAS 24) California Psychotherapy Alliance Scale	Indicate what measure is being coded	0=**did not** use CALPAS 1=did use CALPAS

AG: Name of measure (HAq-I) The Penn Helping Alliance Questionnaire	Indicate what measure is being coded	0=**did not** use HAq-I 1=did use HAq-I
AH: Name of measure (HAq-II) The Revised Penn Helping Alliance Questionnaire	Indicate what measure is being coded	0=**did not** use HAq-II 1=did use HAq-II
AI: Name of measure (ARM-28) The Agnew Relationship Measure	Indicate what measure is being coded	0=**did not** use ARM-28 1=did use ARM-28
AJ: Name of measure (ARM-5) The Agnew Relationship Measure-5	Indicate what measure is being coded	0=**did not** use ARM-5 1=did use ARM-5
AK: Name of measure (VTAS) The Vanderbilt Therapeutic Alliance Scale	Indicate what measure is being coded	0=**did not** use VTAS 1=did use VTAS
AL: Name of measure (TBS) The Therapeutic Bond Scales	Indicate what measure is being coded	0=**did not** use TBS 1=did use TBS
AM-AO: Reliability coefficient, mean score and standard deviation WAI therapist	Indicate the alpha value, the mean and SD for the WAI-T for each independent sample reported in the article. Use separate rows for each sample **NOTE**: *If total item mean and SD is reported convert values to total scale mean using formulas provided.* *If r-values reported in range – code the range* *If r-value is reported for combined sample (e.g., completers and dropouts) calculate the combined mean and SD for this r-value using the formulas provided* *If multiple r-values are reported extract the r-value closest to the third session* *Continue to apply these rules to the following measures.*	Code as a continuous variable 999=missing data
AP-AR: Reliability coefficient, mean score and standard deviation WAI client	Indicate the alpha value, the mean and SD for the WAI-C for each independent sample reported in the article. Use separate rows for each sample	Code as a continuous variable 999=missing data

AS-AU: Reliability coefficient, mean score and standard deviation WAI-S therapist	Indicate the alpha value, the mean and SD for the WAI-S-T for each independent sample reported in the article. Use separate rows for each sample	Code as a continuous variable 999=missing data
AV-AX: Reliability coefficient, mean score and standard deviation WAI-S client	Indicate the alpha value, the mean and SD for the WAI-S-C for each independent sample reported in the article. Use separate rows for each sample	Code as a continuous variable 999=missing data
AY-BA: Reliability coefficient, mean score and standard deviation WAI-SR client	Indicate the alpha value, the mean and SD for the WAI-SR-C for each independent sample reported in the article. Use separate rows for each sample	Code as a continuous variable 999=missing data
BB-BD: Reliability coefficient, mean score and standard deviation CRF-S client	Indicate the alpha value, the mean and SD for the CRF-S-C for each independent sample reported in the article. Use separate rows for each sample	Code as a continuous variable 999=missing data
BB-BD: Reliability coefficient, mean score and standard deviation CALPAS client	Indicate the alpha value, the mean and SD for the CALPAS-C for each independent sample reported in the article. Use separate rows for each sample	Code as a continuous variable 999=missing data
BH-BJ: Reliability coefficient, mean score and standard deviation CALPAS observer	Indicate the alpha value, the mean and SD for the CALPAS-O for each independent sample reported in the article. Use separate rows for each sample	Code as a continuous variable 999=missing data
BK-BM: Reliability coefficient, mean score and standard deviation Haq-I client	Indicate the alpha value, the mean and SD for the HAq-I C for each independent sample reported in the article. Use separate rows for each sample	Code as a continuous variable 999=missing data
BN-BP: Reliability coefficient, mean score and standard deviation	Indicate the alpha value, the mean and SD for the HAq-I O for each independent sample	Code as a continuous variable 999=missing data

Haq-I observer	reported in the article. Use separate rows for each sample	
BN-BS: Reliability coefficient, mean score and standard deviation Haq-II therapist	Indicate the alpha value, the mean and SD for the HAq-II T for each independent sample reported in the article. Use separate rows for each sample	Code as a continuous variable 999=missing data
BT-BV: Reliability coefficient, mean score and standard deviation Haq-II client	Indicate the alpha value, the mean and SD for the HAq-II C for each independent sample reported in the article. Use separate rows for each sample	Code as a continuous variable 999=missing data
BW-BY: Reliability coefficient, mean score and standard deviation ARM-28 client	Indicate the alpha value, the mean and SD for the ARM-28 C for each independent sample reported in the article. Use separate rows for each sample	Code as a continuous variable 999=missing data
BZ-CB: Reliability coefficient, mean score and standard deviation ARM-5 therapist	Indicate the alpha value, the mean and SD for the ARM-5 T for each independent sample reported in the article. Use separate rows for each sample	Code as a continuous variable 999=missing data
CC-CE: Reliability coefficient, mean score and standard deviation ARM-5 client	Indicate the alpha value, the mean and SD for the ARM-5 C for each independent sample reported in the article. Use separate rows for each sample	Code as a continuous variable 999=missing data
CF-CH: Reliability coefficient, mean score and standard deviation TBS client	Indicate the alpha value, the mean and SD for the TBS C for each independent sample reported in the article. Use separate rows for each sample	Code as a continuous variable 999=missing data

CHAPTER 3: Assessing Therapeutic Alliance Scales: Meta-Analytic Evaluation of Adult Individual Psychotherapy

Assessing Therapeutic Alliance Scales: A Validity Generalization Meta-Analytic Evaluation of Outcomes in Adult Individual Psychotherapy

The therapeutic alliance is widely understood to be an essential ingredient to the therapeutic process. Given this, it is not surprising that one of the most extensively researched areas using this construct focuses on the relation between alliance and treatment outcome. To date, there have been hundreds of studies and numerous meta-analyses conducted on the predictive validity of the alliance on treatment outcomes (i.e., the alliance-outcome relation). Of these many meta-analyses, five have focused on therapy results with either adults or adults and older adolescents, with four focused on the alliance and general treatment outcomes and one focused on the relation between alliance and treatment dropout. These five meta-analyses all found comparable summary effect sizes (General treatment outcomes: Horvath & Symonds, 1991, r = .26; Martin et al., 2000, r = .22; Horvath & Bedi, 2002, r = .21; Horvath et al., 2011, r = .275. Dropout: Sharf et al., 2010, d = .55 [approximately r = .27]), indicating that the therapeutic alliance is a predictor of outcomes across therapeutic approaches.

The findings from alliance-outcome research appear conclusive, however, they may not be as definitive as they first seem. The alliance literature has been fraught with unaddressed measurement issues that may impact these results. In addition, the range of the studies included in these analyses may have been overly broad, thus leading to potentially inaccurate conclusions about the alliance-outcome relation. Finally, the combination of data based on slightly different measurement models (i.e., correlation, partial correlation, and multiple regression) may also have influenced the results. These measurement and methodological issues are further discussed below, with the goal of this

meta-analysis to address these potential confounds that may have influenced the alliance-outcome research.

Measurement Considerations

Although the research on the alliance and the alliance-outcome relation is extensive, there are a number of measurement issues in this literature that have not been fully addressed or resolved. As a result, there is little information on the extent to which these issues influence the estimated alliance-outcome relation effect size and or the variability of this relation across studies. First, the manner in which alliance is operationalized and measured is potentially problematic, as the therapeutic alliance construct is often vaguely defined within much of both the professional and research literature. Current conceptualizations of the therapeutic alliance are said to share a number of common features, such as being pan-theoretical and emphasizing the collaboration and consensus between therapist and client. However, many researchers do not clearly define the alliance construct they are studying (Horvath et al., 2011). Consequently, this has resulted in the development of alliance measures that are based on a range of differing conceptualizations of the therapeutic alliance. Indeed, over 30 alliance measure have been identified within the psychotherapy literature, with none of the measures based on a commonly accepted definition of this psychological construct (Hatcher & Barends, 1996; Hatcher, 2010). Despite substantial differences in how alliance is conceptualized in these measures, they are frequently treated as being conceptually and functionally equivalent.

Second, because of this confusion about the precise nature of these alliance measures, and the considerable variability in the thoroughness of efforts taken to develop and validate the various alliance measures, a number of measurement problems have

arisen within the alliance literature (Ardito & Rabellino, 2011; Elvins & Green, 2008; Horvath et al., 2011; Krause et al., 2011). For example, the measures vary considerably on which dimensions of the alliance construct they address or emphasis (e.g., bond, collaboration, consensus, the therapist's involvement in therapy, or the client's capacity and motivation to engage in therapeutic work) and the number of items used to operationalize these dimensions. Findings from factor analytic analyses also suggest that, for some measures, items designed to tap a specific alliance dimension do not necessarily load on the factor representing that dimension (Hatcher & Barends, 1996; Hatcher, 2010). For instance, it has been noted that the WAI bond scale, which has its conceptual roots embedded in Bordin's (1979) working alliance theory, does a good job of capturing the therapist's contribution to the personal relationship between client and therapist. However, the WAI places less emphasis on the client's capacity and motivation to engage in therapeutic work and fails to capture Bordin's central theme of active purposive mutual work (Elvins & Green, 2008; Hatcher & Barends, 1996).

Third, there are concerns about inconsistencies in content between different informant versions (e.g., client versus therapist) of the same measure (Gaston & Marmar, 1994; Horvath & Bedi, 2002; Horvath et al., 2011; Krause et al., 2011). This variability in content means that the various informants may not be evaluating same underlying construct. Finally, there is a tendency for some researchers to modify alliance measures without conducting any evaluations to determine the impact of the modifications on the reliability or validity of scores on the measures. For example, both Acheson et al., (2015) and Sherer et al., (2007) modified the alliance measures they used (i.e., WAI and CALPAS, respectively) to better suit to the treatment provided in their study. Details on what

specific modifications where made to the measure were not reported by either author. More importantly, there was no attempt in either study to evaluate the psychometric properties of scores on the new measures. Because of these untested modifications, the results of such studies may not accurately provide information on the therapeutic alliance or the alliance-outcome relation.

The impact of these measurement issues has rarely been evaluated, so there is little information on the extent to which these potential problems exert a meaningful influence on the measurement of alliance or on evaluation of the alliance-outcome relation. Results from Horvath et al.'s (2011) meta-analysis provided some evidence of meaningful differences between alliance measures. In their examination of potential moderators of the alliance-outcome relation, they examined the alliance-outcome correlations reported in research using the four most commonly used alliance measures (i.e., the California Psychotherapy Alliance Scales (CALPAS), the Helping Alliance Questionnaire (HAq), the Vanderbilt Psychotherapy Process Scales (VPPS), and the Working Alliance Inventory (WAI). For two of these measures (the HAq and the WAI), there was significant heterogeneity across studies, meaning that different versions (i.e., different informant versions, full versions vs. short form versions) of the same measure did not yield the same results in predicting treatment outcome. Horvath and colleagues also found evidence of measurement effects when they examined the interactions between alliance measures used in a study and who completed the measures (i.e., client, therapist, other rater). They reported that the interaction between which alliance measure was used and who rated the alliance accounted for 23% of the variance in the alliance-outcome relation. In other words,

in the studies that Horvath and colleagues included in their meta-analyses, *who* completed *which* alliance measure had a clear impact on the size of the alliance-outcome relation.

Study Inclusion Considerations

Beyond the concerns related to the measurement of the alliance, there are a number of issues about the scope of studies that have been included in previous meta-analyses. These issues include (a) inclusion of studies of mental health services other than psychotherapy (or mental health counselling services), (b) studies that assessed the alliance from multiple sources, (c) studies where it was unclear if those delivering services had any official psychological training, and (d) studies that included youth in their sample. In its resolution of the recognition of psychotherapy effectiveness, the American Psychological Association (APA) Council of Representatives approved a working definition of Psychotherapy as "…the informed and intentional application of clinical methods and interpersonal stances derived from established psychological principles for the purpose of assisting people to modify their behaviors, cognitions, emotions, and/or other personal characteristics in directions that the participants deem desirable," (APA, 2013, p. 102). Therefore, studies that evaluate the alliance in a context that falls outside of this definition should not be considered "true" psychotherapy, and as such, should not be included when evaluating the alliance-outcome relation in psychotherapy literature. This is particularly relevant for the conclusions derived from previous meta-analyses where alliance and outcome data were included from studies of case management, simple psychoeducational programs, and other forms of mental health services that are not psychotherapy. For example, Solomon et al. (1995) measured the alliance within the context of case management and not psychotherapy.

Even within the context of studies examining the effect of the alliance on psychotherapy outcome, many meta-analyses have included data that are problematic for evaluating the alliance-outcome relation. When psychotherapy was provided in the context of a larger array of mental health services, some studies reported therapeutic alliance data based on involvement with all members of the health care team, not just the clinician who provided psychotherapy. For example, Schonberger and colleagues (2006) assessed the alliance between the client and their primary therapist of the rehabilitation team. In a number of these cases, the primary therapist was someone other than the treating psychologist (e.g., physiotherapist, speech pathologist). Furthermore, according to the APA resolution on psychotherapy effectiveness, services must be provided by "a bona fide health provider" (2013, p. 102). In the alliance-outcome literature there many studies where it was unclear if those providing the treatment had any psychotherapy training. As an illustration, many studies examining the alliance-outcome relation in the treatment of substance abuse involved the service being delivered by recovering alcoholics (Meier & Donmall, 2006). Finally, previous meta-analyses examining the alliance-outcome relation have included studies in which there was insufficient information reported in the study to determine if the service was actually psychotherapy or if the service was delivered by an appropriately trained clinician.

A further confound that could influence the results of previous meta-analyses is that the meta-analytic studies often included primary studies in which substantial numbers of participants were adolescents, even though the focus of the meta-analyses was on the alliance-outcome relation in the literature on psychotherapy for adults. For example, Botella (2008), Karver et al., (2008), Tyron (1995) all had clients under the age of 18.

Beyond the issue of whether the alliance-outcome relation is the same for adolescents and adults, it is important to recognize that most alliance measures were developed and validated for use with adult samples. Developmental differences between youth and adults (and the challenges this brings to therapy), makes the alliance especially critical for youth treatment, (Shirk & Karver, 2012). As such, the use of adult-focused alliance measure with adolescent samples is appropriate and may well yield inaccurate information. Thus, including studies with both adolescents and adults in meta-analyses designed to evaluate the alliance-outcome relation in adult psychotherapy services may misrepresent the true alliance-outcome relation and limit what can be concluded concerning psychotherapy for adults.

A final, but separate, confound that may yield inaccurate conclusions from meta-analyses is the inclusion of correlations, partial correlations, and data from regression analyses in the same analyses, despite the fact that the measurement models underlying these different statistics are distinct. Even if the mean effect sizes found across these three types of analyses are comparable in size, it is still inappropriate to combine them in one analysis, as they are based on different conceptual approaches to assessing the alliance-outcome relation. A simple correlation between, for example, alliance measured early in therapy and eventual treatment outcome provides information on whether early alliance predicts client functioning at the end of treatment. In contrast, the partial correlation model involves the early assessment of alliance and the assessment of client functioning both early in treatment and at the end of treatment. Correlating early alliance with client functioning at the termination of treatment, after controlling for the influence of early client functioning on this association, provides information on whether early alliance

predicts *changes* in client functioning over the course of treatment. As for regression analyses, it is common to find studies in which the size of the alliance-outcome association is provided only after controlling for the effects of other variables, such as gender, age, and various psychological characteristics. The results from such analyses provide very different information than what is available from correlational or partial correlational analyses. By combining, in a meta-analysis, data from simple correlational analyses with those of more complex regression analyses, there is considerable risk of obtaining inaccurate results.

The Present Meta-Analysis

Because the findings from the alliance-outcome research inform psychotherapy training and practice, it is important to address the above issues in a meta-analysis of the alliance-outcome literature on adult psychotherapy. To clearly define where any potential differences may lie within the varying alliance measures identified above, in the present study I will conduct meta-analyses in order to synthesize the alliance-outcome effect sizes that have been reported for the most commonly used therapeutic alliance measures in the adult psychotherapy literature. I will use meta-analytic procedures to determine (a) summary effect sizes (both mean values and confidence intervals) for treatment outcomes and (b) the potential moderating influence of sample and study characteristics on these effect sizes, including the potential impact of which alliance measure was used.

Although the meta-analytic procedures outlined above are similar to the previous alliance-outcome meta-analyses, there are several important differences to this meta-analysis that should be noted. First, I will take a more conservative approach in this meta-analysis to analyzing the alliance-outcome research by including only those studies that could be identified as providing “true psychotherapy” as defined by the APA (2013).

Second, in this meta-analysis I will only include studies that focused solely on adult individual psychotherapy so that appropriate conclusions can be drawn regarding the alliance-outcome relation in psychotherapy for adults. Third, as most alliance-outcome studies have used only a small subset of the available alliance measures, I will only include studies that used an alliance measure that has been identified as being commonly used in the literature (Elvins & Green, 2008). This will allow for a detailed analysis of the potential moderating effects due to the type of alliance measure used in a study because, by ensuring that there are sufficient data associated with each measure, there will be sufficient power for such moderation analyses to be conducted. Fourth, I will conduct separate analyses for correlational data and partial correlational data when examining the alliance-outcome relation. Fifth, given the emphasis in the psychotherapy literature on the ability of early alliance to predict client functioning at the end of treatment, I will only include studies in which alliance was measured early in treatment. This will ensure that any results are not confounded by the inclusion of data from studies in which alliance was measured in later stages of treatment. Because of these substantial differences in both inclusion criteria and the manner in which analyses will be conducted, I am not able to make predictions on whether the findings of this study (i.e., mean effect sizes, results of moderator analyses) will be comparable to those in previous alliance-outcome meta-analyses.

Method

Literature Search

The search for studies using therapeutic alliance measures was conducted through searches of three electronic databases: PsycINFO (1806 – January 09, 2017; OVID platform), PubMed (1800 – January 09, 2017; Medline platform), and Dissertations and

Theses (1861 – January 09, 2017; ProQuest platform) (see figure 2 for flowchart regarding study selection). As PsycINFO and PubMed are key databases for the field of psychology, the vast majority of the therapeutic alliance research would be accessible through these two databases. Therefore, SCOPUS, Web of Science, CINHAL, or Social Work abstracts were not included in the search. In addition, as Dissertations and Theses contain unpublished dissertations, no other grey literature was searched. In general, if results from a study were reported in both a dissertation and a published article, results from the published article were retained for the meta-analysis to ensure that the data are only used once. However, in the cases where the dissertation provided more usable data (e.g., sample characteristics data) than the published study, the dissertation was retained instead.

The terms used in the searches were based on Elvins and Green's (2008) review of the literature on the conceptualization and measurement of the therapeutic alliance. Thirty-three alliance measures were identified in this review. As the intent for the present study was to examine measures of alliance in the adult psychotherapy literature (i.e., clients/participants aged 18 years and older), 10 alliance measures designed to assess child, adolescent, or family populations were not included in the database searches. Therefore, the literature was searched using the 23 of the 33 measures. Search terms included the measure's full name, any or all the acronyms for the measure (if applicable), and the acronym(s) of the measure in conjunction with the term alliance (see Table 1 for a list of these search terms). Finally, the results of this search was compared to the studies containing one (or more) of the 23 alliance measures that were included in the Horvath et al. (2011) meta-analysis. Studies included in that meta-analysis, but not identified with the current search strategy, were then reviewed for possible inclusion in the current study (see

Appendix A for differences between studies used in the Horvath et al. (2011) meta-analysis and the current VG study).

Inclusion/Exclusion Criteria

The following were used as inclusion criteria for this VG study: (a) The study was clinical as opposed to analogue (i.e., the study had to have taken place within treatment conditions, not simply a simulation of psychotherapy). This exclusion criterion has been labelled *Analogue*; (b) Alliance was measured in a psychotherapy context, as opposed to the alliance being measured in other service provision situations such as career counselling. This criterion has been labelled *Not Psychotherapy*; (c) The study was an empirical study and not a review or critique of the literature, a book chapter or a book review. This criterion was labelled *Review/Book*; (d) The study had a minimum of 5 adult participants (this requirement was based on the inclusion criteria used in Horvath et al.'s (2011) meta-analysis). As most of these studies consisted of case studies presented in the context of an in-depth examination of another topic, this criterion was included with the criterion labelled *Review/Book*; (e) The psychotherapy provided in the study had been conducted by a licensed service provider. Service providers who were in a graduate level professional training program were also included, as they were under the supervision of a licensed service provider. This criterion was included in order to ensure a certain level of competency on the part of those providing the treatment. Studies were excluded if it was not clear who provided the psychotherapy services or if it was a mixed sample of providers providing the psychotherapy (e.g., multidisciplinary team – not all providers licensed) and the outcome data were derived from the group of providers as whole. Studies in this category were included with those in the criterion of *Not Psychotherapy*; (f) The client must

have direct contact with a therapist. Studies where psychotherapy was provided via a computer program were removed. This criterion was included as part of the criterion *Not Psychotherapy*; (g) The study contained an adult sample only (i.e., 18 years of age and older). However, exceptions were made if the study contained a small number of 17-year olds in the sample. This criterion was labelled *Inexact Treatment Context*; (h) The alliance was measured within the context of individual therapy, rather than group, couple, or family therapy. In the event a study contained a mixed sample (e.g., couples and individuals, licensed versus non-licensed provider), the data were included if the authors reported separate outcome data for those groups that met inclusion criteria (e.g., outcome data for a licensed provider subgroup was provided). This criterion was included as part of the criterion *Inexact Treatment Context*; (i) The study was written in English or French. This criterion was labelled *Study Not Accessible;* (j) A print or electronic version of the study report had to be accessible. Studies that were retrievable via the university library or through interlibrary loan were included. However, studies that could not be accessed by these means were excluded. This criterion was also included under the heading *Study Not Accessible;* (k) Descriptive information needed to evaluate the study's eligibility for the VG study was reported in the article (e.g., age of participants, type of treatment used). If this information was not clearly presented in the article, the study was excluded from further review. In the case where the authors reported that the descriptive information was recorded elsewhere (i.e., made reference to another article), this source was searched for the missing information. This criterion was labelled *Insufficient Information;* (l) The therapeutic alliance had to have been measured during early stages of therapy. Studies where the alliance was assessed only following an intake interview, assessment feedback,

or consultation were eliminated, as were studies where the alliance was assessed only following the midpoint of treatment. The midpoint of treatment was selected as the cut-off for study inclusion as the therapeutic alliance has been shown to be established by the third session (Horvath & Symonds, 1991). This criterion was labelled *Incorrect Time of Assessment*; (m) The study had to have used the specific alliance measure that was entered in the literature search. The specific alliance measure also had to be in its complete form. Studies that reported using a modified version of the measure (e.g., items were removed or reworded) were eliminated. This criterion was labeled *Wrong Measure*; (n) As described previously, in the case of a study that was described in both a dissertation and a published article, the dissertation was eliminated from the pool of studies unless it contained more complete information than the article, in which the article was removed. This criterion was labelled *Duplicates*; (o) Any duplicate publications were eliminated from the pool of studies. Included here were studies in which the data reported in a particular study were also reported in another article. This was done to ensure the independence of data used for the meta-analysis. The study that had the least amount of data pertinent to the VG analyses was removed. This criterion was labelled *Duplicates;* and (p) The studies had to have alliance outcome data. This criterion was labelled *No Alliance Outcome Data*.

Selection of Studies

Phase I of the literature review consisted of excluding the alliance measures that were intended to assess child, adolescent, or family samples (see Figure 1). There were 3,682 reports identified in this initial database search of the remaining 23 therapeutic alliance measures. Phase II of the literature review involved the removal of duplicate studies retrieved from the multiple databases. This resulted in the elimination of 1,367

studies. Phase III of the search consisted of reading the abstracts in order to remove any study that did not fit the inclusion criteria. This evaluation resulted in eliminating a total of 1,277 studies. Phase IV included a full article review to determine fit with study inclusion criteria; this step resulted in the elimination of 879 studies. Phase V involved both abstract and full article review of the articles used by Horvath et al. (2011) but not identified in the current search. As a result of this review, an additional 28 studies were added. At the end of these five phases, a total of 187 studies are available to be included in the meta-analysis.

With regards to the alliance measures, 11 out of the 23 (i.e., ARM, BLRI, CALPAS, CEI, CRF, HAq, TARS, TBS, TSRS, Vanderbilt scales, WAI) were used in the available studies. Analyses conducted for this study were based on validity coefficients for these alliance measures within the 187 studies.

Prior to the analyses being conducted the data were screened to evaluate the suitability of the alliance outcome data reported. Studies that contained alliance-outcome data were removed from this VG for the following reasons: (a) Regression analyses were conducted to examine the association between alliance and outcome, but sociodemographic or psychological variables (other than initial levels of distress/adjustment) were entered in the regression equation prior to the alliance variable; (b) Analyses were not conducted on outcome variables assessed at termination of therapy (e.g., some studies used only treatment follow-up data in analyses); (c) Alliance and outcome data were reported only for an alliance measure subscale, not the full measure; (d) Alliance measure scores were transformed prior to statistical analyses being conducted; (e) Outcome was not assessed in a standardized and reliable manner (e.g., raters conducted chart reviews and provided improvement scores based on therapist

session notes, but no attempts were made to standardize the scoring or examine reliability across raters); (f) Alliance could not have been measured more than once, per rater, during a therapy session; and (g) Sample sizes were not available to allow conversion of the results of statistical analyses into effect sizes. The process of determining suitability resulted in the removal of study data for 6 alliance measures. Thus, 5 different alliance measures (CALPAS, CRF, HAq, WAI, VTAS), in varying formats, were included in the planned analyses.

To control for potential non-independence of data in the planned analyses, each study was allowed to contribute only one alliance-outcome coefficient per study. For studies that reported multiple alliance outcome data, the data were aggregated, as outlined by Borenstein et al. (2009), to obtain an average effect size for use in the analyses.

As noted by Martin et al. (2000), most researchers in the area typically report their results as correlations between alliance and treatment outcomes. As a result, the planned analyses focused on effect size estimates using a strength of association index: Pearson's product-moment correlation (Pearson's *r*) and its non-parametric equivalent Spearman's rank correlation (Spearman's *rho*). This type of estimate focuses on the variance that is shared between two or more variables (Ferguson, 2009). For studies in which the alliance and outcome association was not reported as correlation, the data were converted to *r* prior to analysis (Billiet, 2003; Borenstein et al., 2009; DeFife, 2009; Fritz, Morris, & Richler, 2012)

Coding Sample Characteristics and Potential Moderators

A number of study and sample characteristics, along with a number of variables related to the alliance-outcome data, were coded for each study. These variables were then

tested for heterogeneity of variance. As previously noted, Horvath et al. (2011) found a number of moderator effects within the alliance-outcome data. However, as this VG study has taken a more conservative approach to evaluating the alliance-outcome data (i.e., including only those studies that included the provision of psychotherapy and not those that included the provision of other types of mental health services) these variables were tested as potential moderators even in the cases where heterogeneity of variance was not found to be present in the calculation of mean effect sizes (see Appendix B for the coding manual).

The coded variables were: (a) *Year of publication* (coded as continuous); (b) *Type of report* (coded as 0 for journal, 1 for dissertation); (c) *Language of measure* (coded 0 for English, 1 for other); (d) *Alliance informant therapist* (coded 0 if alliance was not assessed by therapist, 1 if assessed by therapist); (e) *Alliance informant client* (coded 0 if alliance was not assessed by client, 1 if assessed by client); (f) *Alliance informant observer* (coded 0 if alliance was not assessed by observer, 1 if assessed by observer); (g) *Type of population* (coded as 0 for clinical sample, 1 for community sample, 2 for university student sample, 3 for mixed); (h) *Type of mental disorder treated* (coded as 0 for depression, 1 for anxiety, 2 for psychosis, 3 for personality disorder, 4 for mixed sample of disorders, 5 for other); (i) *Type of treatment provided* (coded as 0 for forms of cognitive-behavioural therapy [CBT], 1 for forms of experiential therapy, 2 for forms of psychodynamic therapy, 3 for forms of interpersonal psychotherapy [IPT], 4 for a mixed sample of different psychotherapies, and 5 for other forms of therapy); (j) *Mode of treatment* (coded 0 for face-to-face, 1 for distance communication, 2 for mixed); (k) *Sample size* (coded as continuous); (l) *Mean age and standard deviation of age of therapist and client* (both variables coded separately as

continuous); (m) *Percentage of female therapists and percentage of female clients* (coded separately as continuous); (n) *Mean and standard deviation of years of experience of therapists* (coded separately as continuous); (o) *Name of alliance and outcome measures* (coded separately as reported in article); (p) *Classification of outcome measure* (coded as one of the following outcome variable grouping categories and subcategories: General Distress: Anxiety and related measures, depression and related measures, other symptoms of distress, psychotic symptoms; General Psychosocial Functioning: interpersonal functioning/problems, general functioning; Satisfaction with Life and Self: self-esteem; Substance Use; Dropout - see Appendix C); and (q) *Type of correlation* (coded Full if zero-order correlation reported, Partial if either partial or residualized gain score reported, CS if change score reported, PreTer if outcome variable was premature termination).

In order to examine the reliability of the coding process, a second rater coded 23% (n=14) of the articles coded by the primary investigator (n=60). Inter-rater reliability was assessed for continuous variable codes using intraclass correlation coefficients (ICC), (Shrout & Fleiss, 1979). Benchmarks recommended by Hunsley and Mash (2008) were used (i.e., ICC values of .70-.79 as adequate, ICC values of .80-.89 as good, and ≥ .90 as excellent). Reliability values for the continuous moderator variables were excellent, with only one variable below a value of 1. The rating of percentage of female therapists obtained an ICC value of .91. Disagreements for this moderator occurred within two studies and were the results of extracting the data at different locations in the article (i.e., under results table versus in method section). Inter-rater reliability for categorical codes was assessed using the kappa benchmarks recommended by Hunsley and Mash (2008) (i.e., k values of .60-.74 as adequate, k values of .75 - .84 as good, and k values of ≥ .85 excellent). Reliability

for the categorical moderators were all excellent with the exception of one variable. Reliability for the type of population had a value below the good range (i.e., Kappa = .64). The disagreement in coding for this variable originated from only one study where there was a difference in interpreting the type of setting as either clinical or community. The likely reasons for kappa being so affected by a single study stems from the small sample size (i.e., n=14) and the fact that almost all the studies coded fell into the community category. Discrepancies in coding were resolved through joint discussion and review of the articles until a unanimous agreement was reached.

Results

Analyses were conducted using Comprehensive Meta-Analysis, Version 3 (CMA; Borenstein, 2016). The analytic procedures outlined by Borenstein, Hedges, Higgins, and Rothstein (2009) were followed, with all analyses weighted by the inverse of their variance. Most meta-analyses do not use *r* values in their computations, as these values tend not to be normally distributed. Therefore, Fisher-*z* transformations were performed to normalize the sampling distribution of the *r* values (Borenstein et al., 2009). After the Fisher's *z* score and the standard deviation of the score were used in the analyses, the transformation process was reversed and the meta-analytic results were presented in their original metrics (i.e., *r*). CMA was used to calculate the summary effect size for each of the alliance-outcome coefficients based on the effect size's category (i.e., full versus partial correlation), along with the 95% confidence intervals. Table 8 presents the descriptive statistics for the alliance-outcome relation for both full and partial correlation data. As previously noted, *r*-values slightly different than Cohen's benchmarks (1992) may be more suitable for interpreting effects in psychology (i.e., < .20 is a small effect, between .20 and

.30 is a moderate effect, and < .30 is a large effect; Hemphill, 2003). As seen in Table 8, the full correlation alliance-outcome data overall effect size was r=.24, p<.001, with 95% confidence interval of .20 to .28. The partial correlation data (e.g., controlling for initial levels of distress) overall effect size was r=.23, p<.001, with a 95% confidence interval of .19 to .27.

Q statistic and I^2 index

Heterogeneity of effects was calculated using the *Q* statistic and the I^2 index. The *Q* statistic and its *p* value were used to determine significance (Huedo-Medina, Sanchez-Meca, Marin-Martinez, & Botella, 2006), as a significant *Q* statistic provides evidence for true effect size variance among the different mean alliance-outcome coefficients for the full and partial data. Next, the I^2 index was used to determine how much of the observed variance is due to true differences in effect sizes across studies (Borenstein et al., 2009). This statistic is expressed as a percentage (i.e., from 0% to 100%), with values moving closer to 100% indicating a likelihood the observed variance is non-random and may be explained by covariates. As presented in Table 8, the omnibus *Q* statistics for both the full and partial correlation data were not significant and the I^2 value was 0 for both types of data. As these results suggest that any observed variance for both full and partial data may be explained by random sampling error alone, the likelihood of obtaining significant results for the planned moderator analyses is low.

Analysis of Moderator Variables

A series of planned moderator analyses were performed to evaluate the potential effects of sample characteristics and a number of variables tested by Horvath et al. (2011) as possible moderators. To assess the relation between the obtained effect sizes and

sample characteristics that are continuous in nature (i.e., mean and standard deviation of both client and therapist ages, percentage of female clients and therapists, mean and standard deviation of year of therapist experience, sample size, and year of publication), a series of simple meta-regressions were conducted. None of these moderators, for either full or partial correlation data, was found to be significant.

Following methods described by Borenstein et al. (2009), a variant of analysis of variance (ANOVA) was used to compare the subgroup means for the sample characteristics that were categorical in nature (i.e., population, type of treatment provided, primary mental disorder in the study sample, language of measure, and mode of treatment). When there are five or few studies within a specific subgroup, it is likely that estimates of tau-squared for the analyses will be imprecise (Borenstein, Hedges, Higgins, & Rothstein, 2010). Therefore, it has been suggested to conduct these analyses using the tau-squared pooled estimate as the increase in accuracy gained by pooling more studies is likely to be greater than true differences between the subgroups (Borenstein et al., 2010). This approach was taken in analyses where (a) there were uneven numbers between the subgroups, (b) most subgroups had three or fewer studies contributing data, or (c) there were only two subgroups and one subgroup had data from three or fewer studies. Again, none of these moderators, for either full or partial correlation data, was found to be significant.

The next set of moderator analyses examined the potential effects of the alliance-outcome variables identified by Horvath et al. (2011). Subgroup means were compared for data on: (a) general treatment outcome and dropout, (b) alliance measure variant (as only measures with $k \geq 3$ studies were included, data on the VTAS and CRF-S were not

analyzed), and (c) alliance informant (i.e., client, therapist, other). No significant differences were found for any of these analyses.

The final set of moderator analyses involved evaluating the alliance-outcome relation within two different levels of specificity for treatment outcomes. The first subgroup analyses were conducted in which measures of (a) anxiety symptoms, depressive symptoms, and other symptoms of distress were collapsed into the category of general distress and measures of (b) interpersonal psychosocial functioning and general functioning were collapsed into the category of general psychosocial functioning. The second subgroup analyses were conducted at the level of each type of outcome measure (i.e., separate analyses for data from anxiety symptoms, depressive symptoms, other symptoms of distress, interpersonal functioning, general functioning, satisfaction with life and self, and premature termination of treatment). As with previous subgroup analyses, only those outcome groupings that had a $k \geq 3$ were included, thus resulting in the removal of data on psychotic symptoms and substance use being eliminated from this set of analyses. No significant moderating effects were found for any analyses.

Publication Bias

Rosenthal's fail-safe *N*

Although this VG study included data from both published articles and dissertation studies, it is important to assess whether publication bias may have influenced study results. To address this concern, Rosenthal's fail-safe *N* (1979) was calculated to obtain an estimate of how many studies with smaller effects would impact on the results of this meta-analysis. The number of missing studies needed to render the VG's overall summary effect size *p*-value nonsignificant (i.e., $ES = 0$) was calculated. As seen in Table 8, both the full and

partial correlation alliance-outcome effect sizes would need over 1,000 studies with an effect size of zero to bring the overall summary effect *p*-value nonsignificant. These findings add further support for how robust the observed alliance-outcome relation is.

Funnel Plots

Funnel plot diagrams help to assess the distribution of the effect sizes of the studies included in a meta-analysis and, thus, provide another way to examine possible publication bias effects (Borenstein et al., 2010). A funnel plot that has its point distributed symmetrically around the mean indicates the likely absence of publication bias as the sampling error is considered random. This study's funnel plots have the standard error on the vertical axis (*y*-axis) and the effect sizes (i.e., Fisher's *z*; higher scores representing higher reliability coefficients) on the horizontal (*x*-axis). The use of the standard error on the *y*-axis, as opposed to the sample size or variance, makes it easier to identify asymmetry. By plotting the data in this fashion, it allows for the points to be spread out on the bottom half of the diagram. This is where the studies with smaller sample sizes are plotted, typically due to having more sampling error variation (Borenstein et al., 2010). As shown in Figures 9 and 10, the distribution of the effect sizes for both full and partial data are symmetrical. Egger's tests were then conducted to provide further evidence of asymmetry (Borenstein, 2005). All Egger's test were found to be nonsignificant, indicating there was no significant asymmetry for these effect sizes. Both the funnel plot and Egger's test results indicated that any missing data from the literature would not have an impact on the obtained effects sizes for either the full or partial correlation analyses.

Discussion

The purpose of this study was to (a) explore the alliance-outcome association using predictive validity data reported for the most commonly used therapeutic alliance measure in the adult psychotherapy literature, using only those studies that involved the provision of psychotherapy by professionals trained to provide this service, (b) to assess the alliance-outcome relation for full and partial correlation data separately, and (c) to identify specific study characteristics that might moderate the alliance-outcome relation. After reviewing the available literature, five different alliance measures (CALPAS, CRF, HAq, WAI, VTAS) in various formats and rater versions were included in the analyses, resulting in a total of 15 alliance measure variants being included in this study. Consistent with previous meta-analyses (Horvath & Symonds, 1991; Martin et al., 2000; Horvath & Bedi, 2002; Horvath et al., 2011), an overall moderate effect for the alliance-outcome relation was found. Surprisingly, the magnitude of this effect was almost identical for studies using full correlation data (mean effect size of r = .24) and partial correlation data (mean effect size of r = .23). These results suggest that there is effectively no difference in the early alliance's ability to predict client functioning at the end of treatment versus how much the client's functioning will change over the course of treatment. Put another way, the strength of the association between early alliance and treatment outcome is unaffected by whether or not statistical procedures are used to control for the influence of clients' pre-treatment level of functioning.

An important finding of the research was the homogeneity of variance found across the results of studies included in the meta-analysis. More specifically, no statistically significant variability was observed in effect sizes across the different alliance measures or

the different informants who rated the alliance. That is, *who* completed *which* alliance measure did not affect the alliance-outcome relation. Likewise, there were no observable differences found among the various ways treatment outcomes were assessed. Whether treatment outcomes were measured by assessing client symptoms, level of psychosocial functioning, a combination of symptoms and psychosocial functioning, or premature termination of therapy had no impact on the size of the alliance-outcome association. Overall, the size of the alliance-outcome relation was not influenced by any study, sample, treatment, or measure characteristic.

These results contrast with those reported by Horvath et al. (2011) who found a number of measurement effects on the alliance-outcome relation. It is likely this difference in findings is due to different study inclusion criteria that were used by Horvath and colleagues. These differences include a more exact definition of psychotherapy, an adult only sample, the use of data in which the alliance assessed at mid-point or earlier, and the use of a subset of the most commonly used alliance measures. With this in mind, this study's results strongly suggest that, when assessing the alliance-outcome relation based on studies of adult psychotherapy in which services were provided by appropriately trained mental health professionals, the association between therapeutic alliance and treatment outcome is consistent. The fact that there were no statistically significant moderating effects in this study is a strong indicator of how robust the alliance-outcome relation is. Because there was no evidence to claim significant differences across all forms of treatment, client presenting problems, measurements of alliance, and outcome variables, it can be suggested that the alliance may have a generalized effect on treatment outcome. Although, the correlational nature of the studies included in the meta-analysis does not

allow one to draw conclusions regarding causality, one can speculate this generalized effect on treatment most likely increases the likelihood that clients will remain in therapy and will actively work on their problems.

Limitations

This study has a number of limitations. First, as with all meta-analyses, the main limitation relates to the identification and retrieval of studies that used the targeted therapeutic alliance measures. To address this challenge, study searches were conducted with the two databases most likely to include psychotherapy research (i.e., PsycInfo and PubMed). It is likely, however, that a number of unpublished studies were not located. In addition, a number of studies identified in the initial search that were not retrievable. Of those studies identified as "not accessible" approximately 60% being another language other than English or French (most frequently, German, Spanish, and Chinese), 20% unpublished dissertations, and 20% the journal not available. Attempts were made to retrieve the unpublished dissertations identified in previous meta-analyses but many of these studies were not accessible. In addition, as there are databases that were not searched (e.g., CINHAL), it is conceivable that a number of counselling, nursing, or social work articles may have been missed. However, articles identified from the PsycInfo and PubMed databases did include a number of studies that were specific to these three disciplines and/or were in journals with a focus on these disciplines. To address these issues, several statistical analyses were conducted.

Rosenthal's fail-safe *N* takes into account the possibility of missing data (i.e., unpublished studies), and the very high fail-safe *N* values obtained for the meta-analytic results (i.e., > 1,000 studies) suggest that it is unlikely that addition of data from studies not

included in the meta-analysis would meaningfully affect the observed effect sizes. Funnel plots and Egger's test were also used to address the possibility of bias existing in the set of study results that were used in the meta-analysis. The lack of evidence of an asymmetrical distribution in the data sets used in the meta-analysis suggest that results of unpublished studies not included in the meta-analysis would be unlikely to have an impact on the observed size of the alliance-outcome association.

In addition to identification and retrieval issues, it is possible that the exclusion criteria were too broad and that, therefore, some relevant studies of psychotherapy for adults were excluded from the meta-analysis. Many of the studies deemed not usable were eliminated because they did not provide sufficient information to determine what type of treatment was provided (i.e., whether it was truly psychotherapy or some other form of mental health service) or the qualifications of the treatment provider (i.e., was the clinician qualified to deliver psychotherapy). It is conceivable the inclusion of these studies could have influenced the results of the meta-analysis. Although it is unlikely that the inclusion of these studies would have affected the size of the observed alliance-outcome association (based on the analyses summarized in the previous paragraph), but it is possible that they might have affected the degree of heterogeneity in the effect sizes. It is critical, therefore, that researchers report full details of their study's sample characteristics, including details on who provided the treatment and the precise nature of the treatment that was provided.

Another limitation of this study is that it only included studies that assessed the alliance at, or prior to, the mid-point of treatment. The purpose of this inclusion criterion was to capture the alliance rating closest to the third session, the time point at which most researchers measure the early therapeutic alliance. However, it has been noted that the

alliance-outcome relation can be complex and should be assessed at multiple times points throughout treatment to best predict outcome (Doran, 2016). With this in mind, it is possible that the inclusion of the studies that assessed alliance past the mid-point of therapy could have impacted this study's results, particularly with respect to potential moderator effects.

Undoubtedly, the WAI is the most utilized alliance measure within the literature. It is no surprise then that the WAI is by far the most represented measure within the data set used in this study's analyses. The WAI, in all its variants, accounted for 74% of the effect sizes analyzed within this study, with the CALPAS and the HAq representing the second and third highest proportion of the effect sizes at 18% and 6%, respectively. Relatedly, as the client's rating of the alliance is the most frequently measured within the literature, it is no surprise that it accounted for 71% of the effect sizes, whereas both the therapist's and observer's rating of the alliance accounted for only 14.5% of the effect sizes. Although the findings from this study seemed to be robust, one must keep in mind that they are largely derived from the WAI and the client's perspective on the alliance and, therefore, may not be truly represent outcome-related associations with alliance measures not included in the analyses. With this in mind, it is conceivable that, as additional data for other sources and measures accumulates, estimates of the alliance-outcome relations and its moderators could change.

Conclusions and Recommendations

Early alliance was shown to moderately predict clients' functioning at the end of treatment as well as how much the clients' functioning changed over the course of

treatment. This magnitude of effect was found across studies, regardless of contextual differences in sample, treatment, or measure characteristics, or in the different ways treatment outcomes were assessed. Despite the potential limitations to this meta-analysis, these findings indicate that, when appropriately trained professionals provide psychotherapy to adults, the early therapeutic alliance is a robust predictor of therapy outcome.

With this mind, it is becoming increasingly clear of the importance for clinicians to establish and maintain a good alliance with their clients throughout therapy. In order to help facilitate a good alliance, clinicians would benefit from discussing the importance of the therapeutic relationship with their clients from early on. By explicitly eliciting feedback on the therapeutic relationship, clinicians can help establish the appropriateness of discussing the alliance with their clients. This may help to minimize the likelihood of misunderstandings between client and clinician and to increase the likelihood that clients will maintain their efforts in treatment.

References (General Introduction and Conclusion)

References with an asterisk (*) indicate the studies that were included in the VG analysis.

Acheson, D. T., Feifel, D., Kamenski, M., McKinney, R. & Risbrough, V. B. (2015). Intranasal oxytocin administration prior to exposure therapy for arachnophobia impedes treatment response. *Depression and Anxiety*, *32*, 400–407. doi:10.1002/da.22362

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

American Psychological Association. (2013). Recognition of psychotherapy effectiveness. *Psychotherapy, 50,* 102-109. http://dx.doi.org/10.1037/a0030276

*Andersson, G., Paxling, B., Wiwe, M., Vernmark, K., Felix, C. B., Lundborg, L., … Carlbring, P. (2012). Therapeutic alliance in guided internet-delivered cognitive behavioural treatment of depression, generalized anxiety disorder and social anxiety disorder. *Behaviour Research and Therapy, 50*, 544-550. doi: 10.1016/j.brat.2012.05.003

*Andrade-Gonzalez, N., & A. Fernandez-Liria, A. (2015). Spanish adaptation of the Revised Helping Alliance Questionnaire (HAq-II). *Journal of Mental Health, 24,* 155-161. doi: 10 3109/09638237 2015 1036975

Ardito, R. B., & Rabellino, D. (2011). Therapeutic alliance and outcome of psychotherapy: Historical excursus, measurements, and prospects for research. *Frontiers in Psychology*, *2*, 270. http://doi.org/10.3389/fpsyg.2011.00270

*Auszra, L., Greenberg, L. S., & Herrmann, I. (2013). Client emotional productivity-optimal client in-session emotional processing in experiential therapy. *Psychotherapy Research, 23*, 732-746. http://dx.doi.org/10.1080/10503307.2013.816882

*Baldwin, S. A., Wampold, B. E., & Imel, Z. E. (2007). Untangling the alliance-outcome correlation: Exploring the relative importance of therapist and patient variability in the alliance. *Journal of Consulting and Clinical Psychology, 75*, 842-852. http://dx.doi.org/10.1037/0022-006X.75.6.842

*Barber, J. P., Connolly, M. B., Crits-Christoph, P., Gladis, L., & Siqueland, L. (2000). Alliance predicts patients' outcome beyond in-treatment change in symptoms. *Journal of Consulting and Clinical Psychology, 68*, 1027-1032. http://dx.doi.org/10.1037/0022-006X.68.6.1027

Billiet, P., (2003). The Mann-Whitney U-test: Analysis of 2 between group data with quantitative response variable. Retrieved from http://www.saburchill.com

*Biscoglio, R. L., (2005). *Patient and therapist personality, therapeutic alliance, and overall outcome in brief relational therapy*. (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3159200)

Borenstein, M., (2005). Software for Publication Bias. In H. R. Rothstein, H. R., Sutton, A. J., & Borenstein, M. (Eds), *Publication Bias in Meta-Analysis: Prevention, Assessment and Adjustments* (pp. 193-220). John Wiley & Sons, Ltd, Chichester, UK. doi: 10.1002/0470870168

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

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

Borenstein, M., Hedges, L., Higgins, J. P. T., & Rothstein, H. R. (2016). Comprehensive

Meta-Analysis Version 3. Englewood, NJ: Biostat.

Botella, L., Corbella, S., Belles, L., Pacheco, M., María Gómez, A., Herrero, O., … & Pedro, N. (2008). Predictors of therapeutic outcome and process. *Psychotherapy Research, 18*, 535-542. doi: 10.1080/10503300801982773

*Burns, J. W., Neilson, W. R., Jensen, M. P., Heapy, A., Czlapinski, R., & Kerns, R. D. (2015). Specific and general therapeutic mechanisms in cognitive behavioral treatment of chronic pain. *Journal of Consulting and Clinical Psychology, 83,* 1-11. http://dx.doi.org/10.1037/a0037208

*Busseri, M. A. & Tyler, T. D. (2003). Interchangeability of the Working Alliance Inventory and Working Alliance Inventory, Short Form. *Psychological Assessment, 15*, 193-197. http://dx.doi.org/10.1037/1040-3590.15.2.193

*Byrd, K. R., Patterson, C. L., & Turchik, J. A. (2010). Working alliance as a mediator of client attachment dimensions and psychotherapy outcome. *Psychotherapy: Theory, Research, Practice, Training, 47*, 631-636. http://dx.doi.org/10.1037/a0022080

*Cailhol, L., Rodgers, R., Burnand, Y., Brunet, A., Damsa, C., & Andreoli, A. (2009). Therapeutic alliance in short-term supportive and psychodynamic psychotherapies: A necessary but not sufficient condition for outcome? *Psychiatry Research, 170,* 229-233. doi:10.1016/j.psychres.2008.09.005

*Carryer, J. (2006). *A new jugsaw puzzle: understanding the relationship between psychotherapeutic processes and outcome* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 2008-99020-192)

*Castonguay, L. G., Goldfried, M. R., Wiser, S., Raue, P. J., & Hayes, A. M. (1996). Predicting

the effect of cognitive therapy for depression: A study of unique and common factors. *Journal of Consulting and Clinical Psychology, 64*, 497-504. http://dx.doi.org/10.1037/0022-006X.64.3.497

*Chisholm, S. M. A. (1998). *A comparison of the therapeutic alliances of premature terminators versus therapy completers* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 1998-95024-077)

*Cloitre, M., Stovall-McClough, K. C., Miranda, R., & Chemtob, C. M. (2004). Therapeutic alliance, negative mood regulation and treatment outcome in child abuse-related posttraumatic stress disorder. *Journal of Consulting and Clinical Psychology, 72,* 411-416. doi: 10.1037/0022-006X.72.3.411

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112,* 155–159.

*Coleman, D. (2006). Client personality, working alliance and outcome: A pilot study. *Social Work in Mental Health, 4*, 83-98. http://dx.doi.org/10.1300/J200v04n04_06

DeFife, J. (2009). Emory University. Effect size calculator retrieved from http://web.cs.dal.ca/~anwar/ds/Excel4.xlsx

*de Roten, Y., Fischer, M., Drapeau, M., Beretta, V., Kramer, U., Favre, N., & Despland, J. (2004). Is one assessment enough? Patterns of helping alliance development and outcome. *Clinical Psychology & Psychotherapy, 11*, 324-331. http://dx.doi.org/10.1002/cpp.420

DeVon, H. A., Block, M. E., Moyle-Wright, P., Ernst, D. M., Hayden, S. J., Lazzara, D. J., ... & Kostas-Polston, E. (2007). A psychometric toolbox for testing validity and reliability. *Journal of Nursing Scholarship*, *39*, 155-164.

Doran, J. M. (2016). The working alliance: Where have we been, where are we going?

Psychotherapy Research, 26, 146-163. doi.org/10.1080/10503307.2014.954153

Dozois, D. J. A., Mikail, S. F., Alden, L. E., Bieling, P. J., Bourgon, G., Clark, D. A., … Johnson, C., (2014). The CPA presidential task force on evidence-based practice of psychological treatments. *Canadian Psychology, 55,* 153-160. doi: 10.1037/a0035767

*Dunn, H., Morrison, P., & Bentall, R. P. (2006). The relationship between patient suitability, therapeutic alliance, homework compliance and outcome in cognitive therapy for psychosis. *Clinical Psychology and Psychotherapy, 13,* 145-152. doi: 10.1002/cpp.481Elvins, R., and Green, J. (2008). The conceptualization and measurement of

therapeutic alliance: an empirical review. *Clinical Psychological Review, 28,* 1167–1187. doi:10.1016/j.cpr.2008.04.002

*Fenton, L. R., Cecero, J. J., Nich, C., Frankforter, T. L., & Carroll, K. M. (2001). Perspective is everything: The predictive validity of six working alliance instruments. *The Journal of Psychotherapy Practice and Research, 10,* 262-268.

Ferguson, C. J. (2009). An effect size primer: A guide for clinicians and researchers. *Professional Psychology: Research and Practice, 40,* 532-538. doi: 10.1037/a0015808

Fritz, C., Morris, P., & Richler, J. J., (2012). Effect size estimates: Current Use, Calculations, and interpretation. *The Journal of Experimental Psychology, 141,* 2-18. doi: 10.1037/a0024338

Furr, R. M., & Bacharach, V. R. (2008). *Psychometrics: An introduction* (2nd ed.). Los Angeles, CA: SAGE Publications, Inc.

*Gaiton, L. R. (2004). *Investigation of therapeutic alliance in a treatment study with*

substance-abusing women with Ptsd (Doctoral dissertation) Available from ProQuest Dissertations and Theses database. (UMI No. 2005-99006-313)

Gaston, L., & Marmar, C. R. (1994). The California Psychotherapy Alliance Scales. In B. O. Horvath & L. S. Greenberg (Eds.), *The working alliance: Theory, research, and practice* (pp. 85-108). New York: Wiley.

*Goldman, E. D. (2008). *Chicken or egg, alliance or outcome: An attempt to answer an age-old question* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 2009-99120-159)

*Goldman, R. N., Greenberg, L. S., & Pos, A. E. (2005). Depth of emotional experience and outcome. *Psychotherapy Research, 15,* 248-260. doi: 10.1080/10503300512331385188

Graziano, A. M., & Raulin, M. L. (2004). *Research methods: A process of inquiry (5th ed.).* Boston, MA: Pearson Education Group.

*Grimes, W. R., & Murdock, N. L. (1989). Social influence revisited: Effects of counselor influence on outcome variables. *Psychotherapy: Theory, Research, Practice, Training*, *26*, 469-474. doi.org/10.1037/h0085465

*Hara, K. M., Westra, H. A., Aviram, A., Button, M. L., Constantino, M. J., & Antony, M. M. (2015). Therapist awareness of client resistance in cognitive-behavioral therapy for generalized anxiety disorder. *Cognitive Behaviour Therapy, 44*, 162-174. http://dx.doi.org/10.1080/16506073.2014.998705

Hatcher, R. L. (2010). Clinical studies of the therapeutic alliance. In Muran, J. C. & Barber, J. P. (Eds). *The therapeutic alliance: An evidence-based guide to practice* (pp.7-28). New York, NY: Guilford

Hatcher, R. L., & Barends, A. W. (1996). Patients' view of the alliance in psychotherapy: exploratory factor analysis of three alliance measures. *Journal of Consulting and Clinical Psychology, 64*, 1326–1336. doi: 10.1037/0022-006X.64.6.1326

*Hayes, J. A., Yeh, Y., & Eisenberg, A. (2007). Good grief and not-so-good grief: Countertransference in bereavement therapy. *Journal of Clinical Psychology, 63*, 345-355. http://dx.doi.org/10.1002/jclp.20353

Hemphill, J. (2003). Interpreting the magnitudes of correlation coefficients. *American Psychologist, 58*(1),78-79. doi.org/10.1037/0003-066X.58.1.78

*Herrmann, I. R., Greenberg, L. S., & Auszra, L. (2016). Emotion categories and patterns of change in experiential therapy for depression. *Psychotherapy Research, 26,* 178-195. doi.org/10.1080/10503307.2014.958597

*Hill, C. E., Baumann, E., Shafran, N., Gupta, S., Morrison, A., Rojas, A. E., … Gelso, C. J. (2015). Is training effective? A study of counseling psychology doctoral trainees in a psychodynamic/interpersonal training clinic. *Journal of Counseling Psychology, 62*, 184-201. http://dx.doi.org/10.1037/cou0000053

*Hopkins, W. E. (1988). *The effects of conceptual level matching on the working alliance and outcome in time-limited counseling*. (Doctoral dissertation) Available from ProQuest Dissertations and Theses database. (UMI No. 1988-8818405)

Horvath, A. O., & Bedi, R. P. (2002). The alliance (pp. 37-69). In J. C. Norcross. *Psychotherapy relationships that work.* New York, NY: Oxford University Press.

Horvath, A., Del Re, A. C., Flückiger, C., & Symonds, D. (2011). The alliance. In J. C.

Norcross (Ed.). *Relationships that work* (pp. 25-69). New York, NY: Oxford University Press.

Horvath, A. O., & Symonds, B. D. (1991). Relation between working alliance and outcome in psychotherapy: A meta-analysis. *Journal of Counselling Psychology, 38,* 139-149. http://dx.doi.org/10.1037/0022-0167.38.2.139

Huedo-Medina, T. B., Sanchez-Meca, J., Marin-Martinez, F., & Botella, J. (2006). Assessing heterogeneity in meta-analysis: *Q* statistic or I^2 index? *Psychological Methods, 11,* 193-206. http://psycnet.apa.org/doi/10.1037/1082-989X.11.2.193

Hunsley, J. & Mash E. J. (2008). *A guide to assessments that work.* New York: Oxford University Press.

Karver, M., Shirk, S., Handelsman, J. B., Fields, S., Crisp, H., Gudmundsen, G., & McMakin, D. (2008). Relationship processes in youth psychotherapy: Measuring alliance, alliance-building behaviors, and client involvement. *Journal of Emotional and Behavioral Disorders*, *16*(1), 15-28. doi: 10.1177/1063426607312536

*Katz, J. (1999). *Self-handicapping, the working alliance, Intrepersonal tendencies and the prediction of drop-out*. (Doctoral dissertation) Available from ProQuest Dissertations and Theses database. (UMI No. 1999-9980033)

*Koeing, H. G., Pearce, M., Nelson, B., Shaw, S., Robins, C., Daher, N., … King, M. B. (2016). Effects of religious vs. standard cognitive behavioral therapy on therapeutic alliance: A randomized clinical trial. *Psychotherapy Research, 26,* 365-376. doi: 10.1080/10503307.2015.1006156

*Kramer, U., Kolly, S., Berthoud, L., Keller, S., Preisig, M., Caspar, F., … Despland, J. N. (2014).

Effects of motive-oriented therapeutic relationship in a ten-session general psychiatric treatment of borderline personality disorder: A randomized controlled trial. *Psychotherapy and Psychosomatics, 83*, 176-186. doi/10.1159/000358528

Krause, M., Altimir, C., & Horvath, A. (2011). Reconstructing the therapeutic alliance: Reflections on the underlying dimentions of the concept. *Clinica y Salud, 22,* 267-283. http://dx.doi.org/10.5093/cl2011v22n3a7

*Levy, E. G. (1999). *Therapeutic process in a managed care type setting: The working alliance, pretreatment characteristics and outcome*. (Doctoral dissertation) Available from ProQuest Dissertations and Theses database. (UMI No. 1999-9905784)

*Lo Coco, G., Gullo, S., Prestano, C., & Gelso, C. J. (2011). Relation of the real relationship and the working alliance to the outcome of brief psychotherapy. *Psychotherapy, 48*, 359-367. http://dx.doi.org/10.1037/a0022426

*Lorenzo-Luaces, L., DeRubeis, R. J., & Webb, C. A. (2014). Client characteristics as moderators of the relation between the therapeutic alliance and outcome in cognitive therapy for depression. *Journal of Consulting and Clinical Psychology, 82,* 368-373. doi.org/10.1037/a0035994

*Mallinckrodt, B. (1996). Change in working alliance, social support and psychological symptoms in brief therapy. *Journal of Counseling Psychology, 43*, 448-455. doi.org/10.1037/0022-0167.43.4.448

*Marmar, C. R., Gaston, L., Gallagher, D., & Thompson, L. W. (1989). Alliance and outcome in late-life depression. *The Journal of Nervous and Mental Disease, 177,* 464-472. doi: 10.1097/00005053-198908000-00003

*Marmarosh, C. L., Gelso, C. J., Markin, R. D., Majors, R., Mallery, C., & Choi, J. (2009).The real

relationship in psychotherapy: Relationships to adult attachments, working alliance, transference, and therapy outcome. *Journal of Counseling Psychology, 56*, 337-350. http://dx.doi.org/10.1037/a0015169

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

Meier, P. S., & Donmall, M. C. (2006). Differences in client and therapist views of the working alliance in drug treatment. *Journal of Substance Use, 11*(1), 73-80. doi.org/10.1080/14659890500137004

*Missirlian, T. M., Toukmanian, S. G., Warwar, S. H., & Greenberg, L. S. (2005). Emotional arousal, client perceptual processing, and the working alliance in experiential psychotherapy for depression. *Journal of Consulting and Clinical Psychology, 73*, 861-871. doi.10.1037/0022-006X.73.5.861

*Muran, J. C., Safran, J. D., Gorman, B. S., Samstag, L. W., Eubanks-Carter, C., & Winston, A. (2009). The relationship of early alliance ruptures and their resolution to process and outcome in three time-limited psychotherapies for personality disorders. *Psychotherapy: Theory, Research, Practice, Training, 46*, 233-248. doi.org/10.1037/a0016085

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill

*Nysaeter, T. E., Nordahl, H. M., & Havik, O. E. (2010). A preliminary study of the naturalistic course of non-manualized psychotherapy for outpatients with borderline personality disorder: Patient characteristics, attrition and outcome. *Nordic Journal of Psychiatry, 64,* 87-93. doi: 10.3109/08039480903406731

*Paivio, S. C.,& Bahr, L. M. (1998). Interpersonal problems, working alliance, and outcome in short-term experiential therapy. *Psychotherapy Research, 8,* 392-407. doi.org/10.1093/ptr/8.4.392

*Paivio, S. C., Jarry, J. L., Chagigiorgis, H., Hall, I., & Ralston, M. (2010). Efficacy of two versions of emotion-focused therapy for resolving child abuse trauma. *Psychotherapy Research, 20,* 353-366. doi: 10.1080/10503300903505274

*Patterson, C.L., Anderson, T., & Wei, C. (2014). Clients' pretreatment role expectations, the therapeutic alliance, and clinical outcomes in outpatient therapy. *Journal of Clinical Psychology, 70,* 673-680. http://doi: 10.1002/jclp.22054

*Pos, A. E. (2006). Experiential treatment for depression: A test of the experiential theory of change, differential effectiveness, and predictors of maintenance of gains (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 2007-99012-040)

*Preschl, B., Maercker, A., & Wagner, B. (2011). The working alliance in a randomized controlled trial comparing online with face-to-face cognitive-behavioral therapy for depression. *BMC Psychiatry, 11*, 189. http://dx.doi.org/10.1186/1471-244X-11-189

*Price, P. B., & E. E. Jones (1998). Examining the alliance using the Psychotherapy Process Q-Set. *Psychotherapy: Theory, Research, Practice, Training, 35*, 392-404. http://dx.doi.org/10.1037/h0087654

*Raue, P., Castonguay, L., & Goldfried, M. (1993). The working alliance: A comparison of two therapies. *Psychotherapy Research, 3,* 197-207. doi: 10.1080/10503309312331333789

*Reis, S., & B. F. Grenyer (2004). Fearful Attachment, Working Alliance and Treatment

Response for Individuals with Major Depression. *Clinical Psychology and Psychotherapy, 11*, 414-424. http://dx.doi.org/10.1002/cpp.428

*Richards, D., Timulak, L., & Hevey, D. (2013). A comparison of two online cognitive-behavioural interventions for symptoms of depression in a student population: The role of therapist responsiveness. *Counselling & Psychotherapy Research, 13,* 184-193. doi: 10.1080/14733145.2012.733715

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*, 638. doi/10.1037/0033-2909.86.3.638

*Safran, J. D., & Wallner, L. K. (1991). The relative predictive validity of two therapeutic alliance measures in cognitive therapy. *Psychological Assessment, 3*, 188-195. doi.org/10.1037/1040-3590.3.2.188

*Sauer, E. M., Anderson, M. Z., Gormley, B., Richmond, C. J., & Preacco, L. (2010). Client attachment orientations, working alliances, and responses to therapy: A psychology training clinic study. *Psychotherapy Research, 20*, 702-711. http://dx.doi.org/10.1080/10503307.2010.518635

Schönberger, M., Humle, F., & Teasdale, T. W. (2006). Subjective outcome of brain injury rehabilitation in relation to the therapeutic working alliance, client compliance and awareness. *Brain Injury, 20,* 1271-1282. doi:10.1080/02699050601049395

Sharf, J., Primavera, L. H., & Diener, M. J. (2010). Dropout and therapeutic alliance: A meta-analysis of adult individual psychotherapy. *Psychotherapy: Theory, Research, Practice, Training, 47,* 637-645. doi: 10.1037/a0021175

Sherer, M., Evans, C.C., Leverenz, J.T., Stouter, J., Irby, J.W., Lee, J.E., & Yablon, S.A. (2007).

Therapeutic alliance in post-acute brain injury rehabilitation: predictors of strength of alliance and impact of alliance on outcome. *Brain injury, 21,* 663-72. doi.org/10.1080/02699050701481589

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin, 86*, 420-428. http://psycnet.apa.org/doi/10.1037/0033-2909.86.2.420

Solomon, P., Draine, J. & Delaney, M.A. (1995). The working alliance and consumer case management. *The Journal of Mental Health Administration, 22*: 126-134. https://doi.org/10.1007/BF02518753

*Stapor, B. S. (1999). The effect of working alliance on termination status at a college counseling center (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 1999-95009-170)

*Stracuzzi, T. I., et al. (2011). Gay and bisexual male clients' perceptions of counseling: The role of perceived sexual orientation similarity and counselor universal-diverse orientation. *Journal of Counseling Psychology, 58*, 299-309. http://dx.doi.org/10.1037/a0023603

*Strauss, J. L., Hayes, A. M., Johnson, S. L., Newman, C. F., Brown, G. K., Barber, J. P., … Beck, A. T. (2006). Early alliance, alliance ruptures, and symptom change in a nonrandomized trial of cognitive therapy for avoidant and obsessive-compulsive personality disorders. *Journal of Consulting and Clinical Psychology, 74*, 337-345. http://dx.doi.org/10.1037/0022-006X.74.2.337

*Stringer, J. V., Levitt, H. M., Berman, J. S., & Mathews, S. S. (2010). A study of silent disengagement and distressing emotion in psychotherapy. *Psychotherapy Research,*

20, 495-510. http://dx.doi.org/10.1080/10503301003754515

Shirk, S. R., & Karver, M. S. (2011). The alliance. In J. C. Norcross (Ed.). *Relationships that work* (pp. 70-108). New York: Oxford University Press.

*Stevens, C. L., Muran, J. C., Safran, J. D., Gorman, B. S., & Winston, A. (2007). Levels and Patterns of the Therapeutic Alliance in Brief Psychotherapy. *American Journal of Psychotherapy, 61*, 109-129.

Suen, H. K. (1990). *Principles of Test Theories*. Hillsdale: Lawrence Erlbaum.

Swift, J. K., & Greenberg, R. P. (2012). Premature discontinuation in adult psychotherapy: A meta-analysis. *Journal of Consulting and Clinical Psychology*, *80*(4), 547. doi/10.1037/a0028226

*Taylor, P. J., Rietzschel, J., Danquah, A., & Berry, K. (2015). The role of attachment style attachment to therapist, and working alliance in response to psychological therapy. *Psychology and Psychotherapy: Theory, Research and Practice, 88,* 240-253. http://doi:10.1111/papt.12045

Tyron, G. S., & Kane, A. S. (1995). Client involvement, working alliance and type of psychotherapy termination. *Psychotherapy Research, 5,* 189-198. doi.org/10.1080/10503309512331331306

Tyron, W. W., & Bernstein, D. (2003). Understanding measurement. In J. C. Thomas and M. Hersen (Eds.). *Understanding research in clinical and counseling psychology: A textbook* (pp. 27-68). Mahwah: Earlbaum.

*Turner, H., Bryant-Waugh, R., & Marshall, E. (2015). The impact of early symptom change

and therapeutic alliance on treatment outcome in cognitive-behavioural therapy for eating disorders. *Behavior Research and Therapy, 73,* 165-169. doi.org/10.1016/j.brat.2015.08.006

Vacha-Haase, T. (1998). Reliability generalization: Exploring variance in measurement error affecting score reliability across studies. *Educational and Psychological Measurement*, *58,* 6-20. doi: 10.1177/0013164498058001002

*Van, H. L., Hendriksen, M., Schoevers, R. A., Peen, J., Abraham, R. A., & Dekker, J. (2008). Predictive value of object relations for therapeutic alliance and outcome in psychotherapy for depression: an exploratory study. *Journal of Nervous & Mental Disease, 196*, 655-662. doi: 10.1097/NMD.0b013e318183f8c2

*Wagner, B., Brand, J., Schulz, W., Knaevelsrud, C. (2012). Online working alliance predicts treatment outcome for posttraumatic stress symptoms in Arab war-traumatized patients. *Depression and Anxiety, 29*, 646-651. http://dx.doi.org/10.1002/da.21962

*Westmacott, R., Hunsley, J., Best, M., Rumstein-McKean, O., & Schindler, D. (2010). Client and therapist views of contextual factors related to termination from psychotherapy: A comparison between unilateral and mutual terminators. *Psychotherapy Research, 20*, 423-435. http://dx.doi.org/10.1080/10503301003645796

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

*Xu, H., & Tracey, T. J. G. (2015). Reciprocal influence model of working alliance and therapeutic outcome over individual therapy course. *Journal of Counseling*

Psychology, 62, 351-359. http://dx.doi.org/10.1037/cou0000089

Table 7

List of Measure-related Search Terms

	Measure as identified in Elvins and Green (2008)	Measure's full name used in search	Abbreviations/Alternate name of measure used in search
1	Barrett-Lennard's Relationship Inventory	Barrett-Lennard's Relationship Inventory	BLRI
2	Counselling Evaluation Inventory	Counselling Evaluation Inventory	CEI
3	Therapy Session Report Scales	Therapy Session Report Scale	Therapy Session Report; TSRS; TRS
4	The Counselor Rating Form	The Counselor Rating Form	CRF
5	The Penn Alliance Scales	The Penn Helping Alliance Scales	The Helping Alliance Questionnaire; The Penn; HAq
6	Vanderbilt Scales	Vanderbilt Therapeutic Scales	Vanderbilt Therapeutic Alliance Scale; VTAS; Vanderbilt Psychotherapy Process Scale; VPPS
7	Toronto Scales	Therapeutic Alliance Rating Scales	TARS
8	Menninger Alliance Rating Scale or Collaboration Scale	Menninger Alliance Rating Scale or Collaboration Scale	Menninger
9	Psychotherapy Status Report	Psychotherapy Status Report	PSR
10	Patient Collaboration Scale	Patient Collaboration Scale	PCS
11	California Scales	California Psychotherapy Alliance Scales	CALPAS
12	Therapeutic bond Scales	Therapeutic Bond Scale	TBS
13	Working Alliance Inventory	Working Alliance Inventory	WAI
14	Treatment Alliance Scales	Treatment Alliance Scales	TAS
15	Adapted psychotherapy Process Inventory	Adapted Psychotherapy Process Inventory	PPI

16	Helping Alliance Scales	Helping Alliance Scales	Helping Alliance Scale; HAS
17	Empathy and Understanding Questionnaire	Empathy and Understanding Questionnaire	EUQ
18	Barriers to Treatment Participation Scale	Barriers to Treatment Participation Scale	BTPS
19	Therapist Alliance Focus Scale	Therapist Alliance Focus Scale	TAFS
20	Agnew Relationship Measure	Agnew Relationship Measure	ARM
21	Kim Alliance Scale	Kim Alliance Scale	KAS
22	The Therapy Process Observational Coding System — Alliance Scale	The Therapy Process Observational Coding System — Alliance Scale	TPOCS
23	Scale to Assess Therapeutic Relationship	Scale to Assess Therapeutic Relationship	STAR

Table 8

Descriptive Statistics for Mean Alliance-Outcome Coefficients

				95% Confidence Interval				
Type of correlation	*k*	Mean	*p*	Lower	Upper	*Q*	I^2	Fail-safe *N* (# of studies)
Full	34	.24	<.001	.20	.28	32.41	--	1,069
Partial	46	.23	<.001	.19	.27	38.02	--	1,129

Figure 8 Flow Chart of Studies Selection

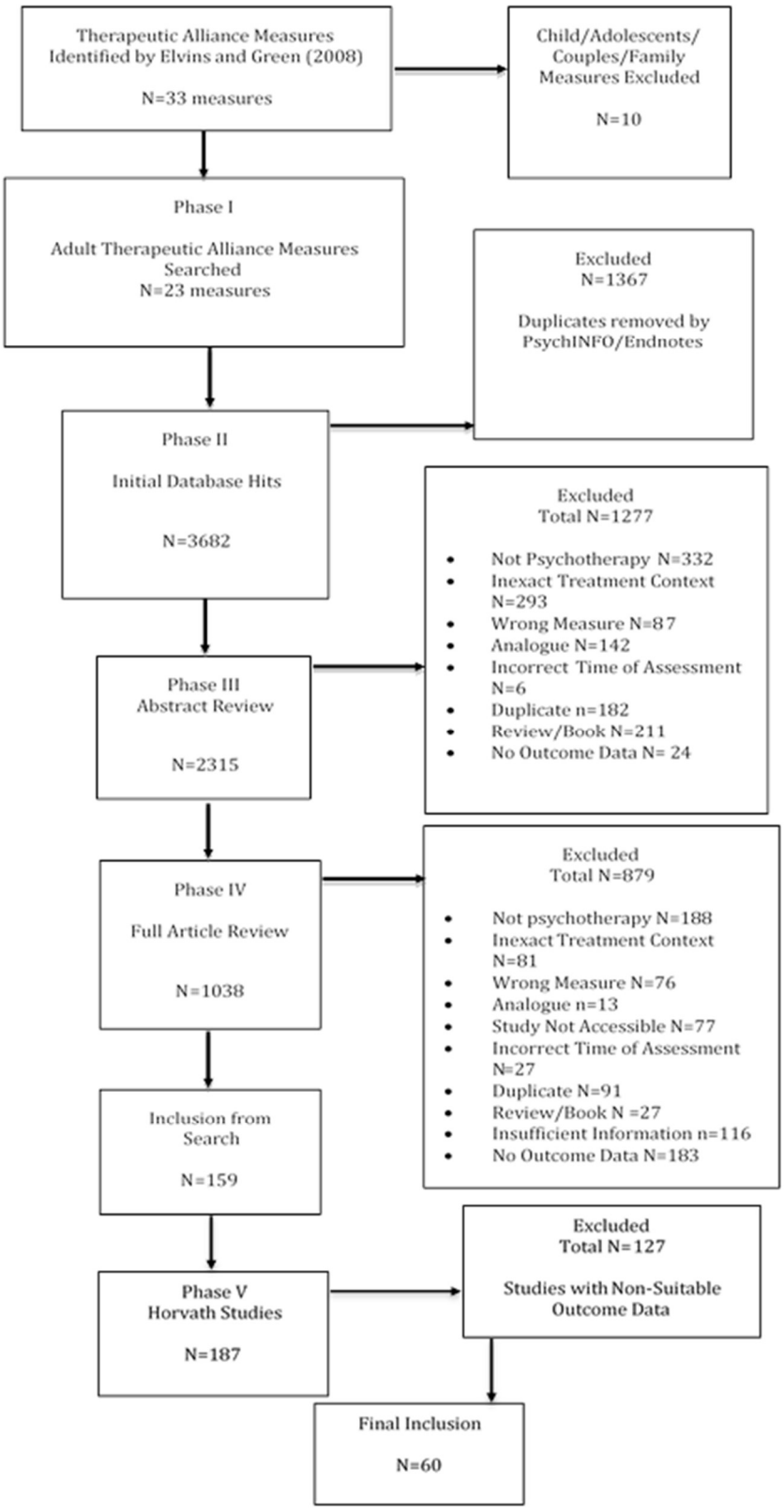


Figure 9 Funnel Plot Publication Bias (full correlation)

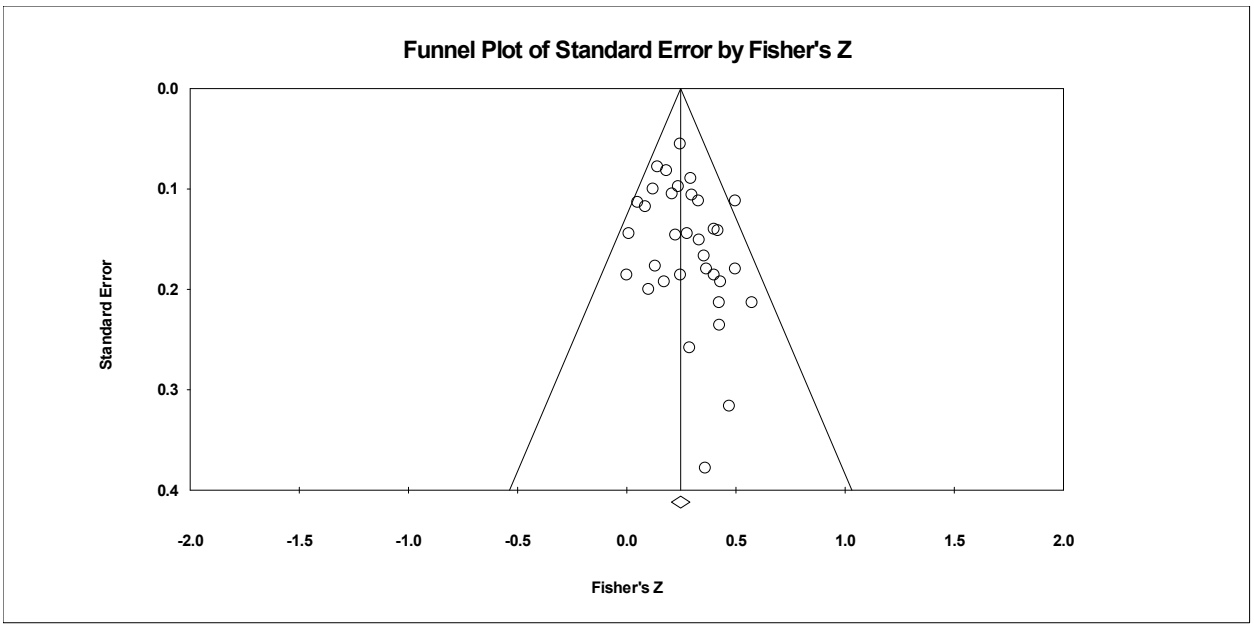
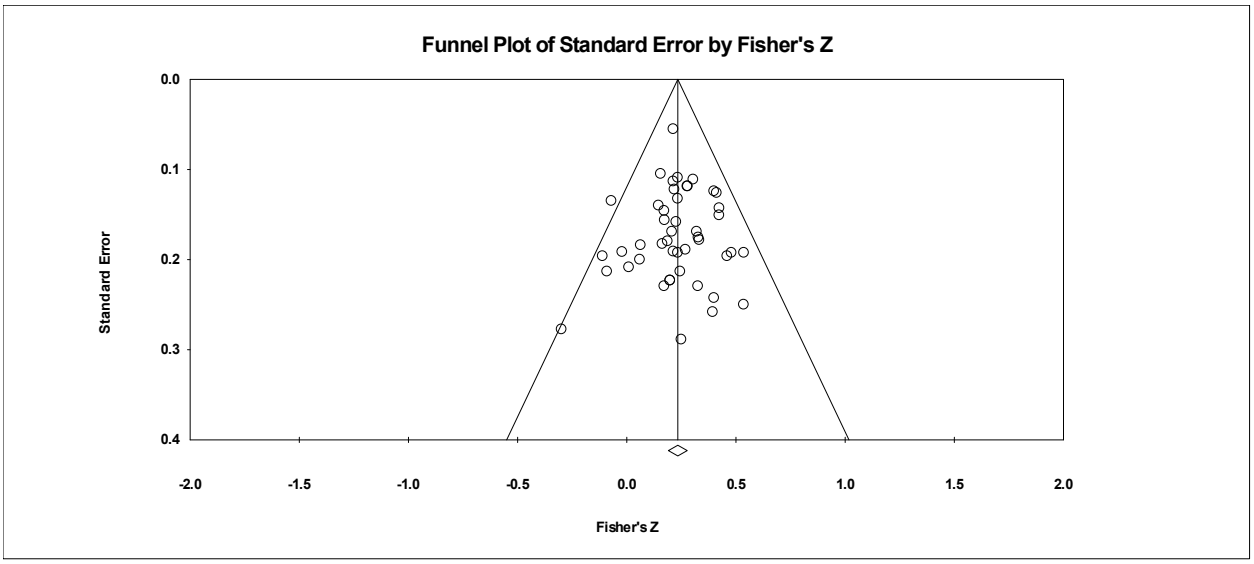


Figure 10 Funnel Plot Publication Bias (partial correlation)



Appendix A

Lists of studies included in VG and Horvath et al., 2011 study.

Studies obtained in VG search; included in both VG and Horvath	Studies obtained in VG search; included in Horvath but not in vG	Studies not obtained in VG search; retrieved from Horvath *(* Studies included into VG after phase V)*
Cloitre et al. (2004)	Fakhoury et al. (2007)	Constantino et al. (2005)*
Coleman (2006)	Dearing et al. (2005)	Deu et al. (2009)
Davis et al. (2007)	Florsheim et al. (2000)	Dorsch et al. (2002)
Knaevelsrud et al. (2007)	Forbes et al. (2008)	Eaton et al. (1988)
Mallinckrodt (1993)	Gallop et al. (1994)	Freitas (2001)
Marmarosh et al. (2009	Hervé et al. (2008)	Forman (1990)
Missirlian et al. (2005)	Howard et al. (2006)	Greenberg et al. (1982)
Muran et al. (2009)	Ilgen et al. (2006a)	Greenberg et al. (2002)
Pos (2007)	Ilgen et al. (2006b)	Hopkins (1988)*
Reis et al. (2004)	Jumes (1995)	Jacob (2003)
Stevens et al. (2007)	Kabuth et al. (2005)	Katz (1999)*
Spinhoven et al. (2007)	Karver et al. (2008)	Kivlighan et al. (2000)*
Wettersten et al. (2005)	Kelly et al. (2009)	Kivlighan et al. (1995)*
Hayes et al. (2007)	Kokotovic et al. (1990)	Klein et al. (2003)
Baldwin (2007)	Meier et al. (2006 b)	Mallinckrodt (1996) *
Biscoglio (2005)	Meier et al. (2006 a	Multon et al. (2001)
Busseri et al. (2003)	Pos et al. (2009)	Moseley (1983)
Dunn et al. (2006)	Ramnerö et al. (2007)	Pugh (1991)
Gaiton (2004)	Solomon et al. (1995)	Rogers et al. (2008)
Gaston et al. (1991)	Tyron et al. (1993)	Santiago et al. (2005)
Tichenor (1989)	Tyron et al. (1995)	Schönberger et al. (2006c)
Bethea et al. (2008)	Schönberger et al. (2006a)	Sexton (1996)
de Roten et al. (2004)	Schönberger et al. (2006b)	Vronmans (2007)
Gerstley et al. (1989)	Schönberger et al. (2007)	Wettersten (2000)
Kramer et al. (2009)	Botella (2008)	Adler (1988)
Morgan et al. (1982)	Burns et al. (2007)	Arnow et al. (2003)
O'Malley et al. (1983)	Gaston et al. (1994)	Bieschke et al. (1995)
Marziali et al. (1981)	Geiser et al. (2002)	Brotman (2004)
Stiles et al. (2004)	Hatcher et al. (1996)	Busseri et al. (2004)
Rounsaville et al. (1987)	Sherer et al. (2007)	Castonguay et al. (1996)
Emmerling et al. (2009)	Trepka et al. (2004)	Chilly (2004)*
	Bassler et al. (1995)	Barber et al. (2006)
	Gunderson et al. (1997)	Barber et al. (1999)
	Kramer et al. (2008)	Barber et al. (2001)
	Mohl et al. (1991)	Barber et al. (2008)*
	Schauenburg et al. (2005)	Barber et al. (2000)*
	Schleussner (2005)	Castonguay et al. (1996)*
	Konzag et al. (2004)	Ferleger (1993)*

	Gomes-Swartz (1978)	Gaston et al. (1998)
	Hawley et al. (2006)	Geider (1997)
	Loneck et al. (2002)	Gunther (1991)*
	Klee et al. (1990)	Gutfreund (1992)
	Windholtz et al. (1988)	Hardy et al. (2001)
	Horvath (1981)	Card (1991)
	Irelan (2004)	Marmar et al. (1989 b)
		Marmar et al. (1989 a)*
		Muran et al. (1995)
		Riley (1992)*
		Strauss et al. (2006)*
		Strauss (2001)
		Tunis et al. (1995)
		Yeomans et al. (1994)*
		Cislo (1998)
		Crits-Cristoph et al. (1988)
		Feeley (1993)*
		Janecke (2003)
		Johansson et al. (2006)
		Luborsky et al. (1985)*
		Luborsky et al. (1983)*
		Marshel (1986)
		Pyne (1991)
		Tyron et al. (1990)
		Vogel et al. (2006)*
		Van et al. (2008)*
		Hartley et al. (1983)
		Krupnick et al. (1994)
		Krupnick et al. (1996)
		Meyer et al. (2002)*
		Zuroff, et al. (2000)
		Marziali (1984)*
		Marziali et al. (1999)
		Horowitz et al. (1984)*
		Kolden (1996)*
		Reiner (1987)
		Saunders (2000)*
		Saunders et al. (1989)*
		Zuroff et al. (2006)*
Exclusion of studies obtained from Horvath not in RG search Insufficient information *n*=8 Wrong measure *n*=4		

Not individual adult psychotherapy *n*= 2 Study not accessible *n*=20 Duplicate *n*= 8 Not psychotherapy *n*=6		

NOTE: Dissertations obtained in original search were retained in place of the published articles obtained from Horvath's studies as the dissertations reported more sample characteristics needed for coding. These articles were still counted for under the exclusion criteria Dissertation. Articles included in Horvath et al.'s 2011 study that used alliance measures other than those listed in the Elvins & Green's 2008 study were excluded from this VG.

Appendix B

Coding Manual for VG

Study's characteristics for moderator analysis	Specifications on what to code	Codes for data entry
A: Study title	Write out the first half of Study's title	
B: Author's name	Write author's name in full (Last name first)	
C: Number of samples coded from article	Rank the number of samples being recorded for the article. Each sample will be coded on a separate row. (e.g., Study 1, row 1, sample 2; Study 1, row 2, sample 2)	Code in rank order 1, 2, 3, etc…
D: Year of Publication	Indicate the year the article was published **NOTE***: if there if a discrepancy b/w Endnotes and Article code articles date*	Code as continuous variable
F: Type of report	Indicate the type of article (e.g., journal article)	0=journal article 1=dissertation
G: Language of measure	Indicate language of measure **NOTE**: *If not clear if the measure was translated (i.e., specifically indicated) code as 999*	0=English 1=other (i.e., translated version)
H: Alliance informant (therapist)	Indicate if assessed the alliance	0=if therapist **did not** assess alliance in study 1=if therapist **did** assess alliance
I: Alliance informant (client)	Indicate if client assessed the alliance	0=client **did not** assess alliance in study 1=if client **did** assess alliance
J: Alliance informant (observer)	Indicate if observer assessed the alliance	0=if observer **did not** assess alliance in study 1=if observer **did** assess alliance
K: Type of population	Indicate what type of population received treatment.	0=clinical sample (i.e., inpatient & outpatient from hospital) 1=community (i.e., university training centers, community

	NOTE**:** *This may be reported as the location the clients were recruited from. If location indicated in article is unclear, search for site in the internet to help clarify sample*	centers, medical centers, veteran centers) 2=university student sample (i.e., if sample is entirely comprised of college or university students) 3=mixed sample 999=missing data
L: Type of mental disorder(s) or problem(s)	Indicate specific mental disorder(s) or problem(s) treated ***NOTE****: Diagnostic criteria may be from the DSM-III, DSM-IV, DSM-IV-R, or DSM-V depending on date of publication; Criteria for type of mental disorder classification for this VG was derived, in part, from the DSM-IV-R criteria .* *Type of mental disorder or problem treated may be described as presenting problem(s) of the sample in the article. Use this description as bases for coding if indicated as such in article*	0=depression or other mood disorders (e.g., bipolar, dysthymia) 1=anxiety disorders (e.g., social, GAD, PTSD) 2=Psychotic disorders (e.g., schizophrenia, schizoaffective disorder) 3=personality disorders 4=mixed sample (i.e., multiple disorders treated across individuals; may include co-morbid disorders) 5=other (e.g., substance abuse, relationship difficulties not indicated as PD, etc…) 999=missing data
M: Type of treatment provided	Indicate what type of treatment was provided to client ***NOTE:*** *Treatment provided may be identified by the therapists' therapeutic orientation in the article*	0=family of cognitive behavioural therapies 1=family of experiential therapies 2=family of psychodynamic therapies 3=family of interpersonal therapies 4=mixed sample of different psychotherapies 5=other 999=missing data
N: Mode of treatment	Indicate manner in which treatment was provided	0=face to face 1=distant (e.g., texting, email, videoconferencing, phone) 2=mixed
O, P & Q: Sample size	Indicate sample size (i.e., o) number of clients the therapist(s) rated alliance for; p) number of clients; q) number of clients observer(s)	Code as continuous variable 999=missing data

	rated alliance on for observer column ***NOTE**: code the sample size value that corresponds to the specific alliance/outcome correlation*	
R: Mean age of Therapists	Indicate the mean age of the study's therapists	Code as a continuous variable (report to two decimal points) 999=missing data
S: Standard deviation of age of therapists	Indicate the standard deviation of the therapists' age	Code as a continuous variable (report to two decimal points) 999=missing data
T: Percentage of female therapists	Indicate the percentage of female therapists of the sample	Code as a continuous variable (report to two decimal points) 999=missing data
U: Mean of years of experience of therapists	Indicate the mean of the years of experience of the therapists	Code as a continuous variable (report to two decimal points) 999=missing data
V: Standard deviation of the years of experience of therapists	Indicated the SD of the years of experience of the therapists	Code as a continuous variable (report to two decimal points) 999=missing data
W: Mean age of Clients	Indicate the mean age of the study's clients	Code as a continuous variable (report to two decimal points) 999=missing data
X: Standard deviation of age of client	Indicate the standard deviation of the clients' age	Code as a continuous variable (report to two decimal points) 999=missing data
Y: Percentage of female Clients	Indicate the percentage of female clients of the sample	Code as a continuous variable 999=missing data
Name of alliance measure	Indicate each alliance measure used for the alliance/outcome correlation. This may include different informants of the same measure (e.g., WAI-C and WAI-T) ***NOTE:** If multiple version of the alliance measure has been used code only the long version of the measure* *(e.g., If WAI-C and WAI-S-C were used code only WAI-C)*	Write name of the alliance measures used in study. Use list below to identity the alliance measure WAI-C WAI-S-C WAI-SR-C WAI-T WAI-S-T WAI-O CALPAS-C-24 CALPAS-C-31 CALPAS-T

		CALPAS-O HAq-I-C HAq-II-C HAq-II-T CRF-S-C
Classify outcome measure grouping	Categorize the outcome measure into one of the outcome measure group categories. Use the outcome measure grouping table provided for list of categories ***NOTE:*** *A composite score may have been calculated from multiple outcome measures. Use only those composite scores that have been identified in the outcome grouping table.* *Dropout data will be indicated by a definition provided by authors and is not measured with a standard outcome measure. (e.g., premature termination was defined as client completing less than a certain number of sessions)*	Write name of outcome category grouping. The abbreviated title may be used to identify the group. The abbreviated title has been identified in the table. (e.g., General Distress may be identified with the initials GD) Use the following acronym to identify Premature Termination (or dropout) (**PreTer**) followed by the definition used by author (e.g., PreTer (DO/CMN) Codes for definition of dropout i) dropout without definition (**DO/ND)** ii) dropout without notification to therapist – not attending planned session, not calling back to make another appointment **(DO/WNT)** iii) Treatment duration criteria not met (**DO/CNM**) (e.g., 3 months, 16 sessions) iv) Missing a set number of sessions **(DO/MNS)** (e.g., missed 3 consecutive sessions)
Name of outcome measure and informant (i.e., Therapist, Client)	Indicate which outcome measure has been used as well as the person who assessed the outcome ***NOTE:*** *Author's may have used only a section of an outcome measure (e.g., GSI of the SCL)*	Write full name or abbreviation of the outcome measure used. Indicated the informant by the letter T or C at the end of the name. (e.g., Global Assessment of Functioning Scale – Therapist (GAF-T), Beck Depression Inventory (BDI-C))

Type of correlation	Identify the type of alliance/outcome correlation used. Use the titles in bold provided below **FULL**: full correlation **PARTIAL**: either partial or residualized gain score **CS**: change score only **PreTer**: premature Termination ***NOTE:*** *Use only the partial correlations that control for the initial levels of distress from the same outcome measure. Any other type of partial correlation is not applicable here*	Write the type of correlation using the list provided.
Alliance/Outcome correlation	Indicate the numerical value of the alliance/outcome correlation. Match each numerical value to it corresponding alliance and outcome measure if multiple correlations are reported ***NOTE:*** *Extract only the alliance/outcome data from which the alliance scores where not combined from multiple measures* *(i.e., authors have combined both the therapists' and clients' alliance measure scores to obtain one alliance/outcome correlation)* *If alliance/outcome data has not been reported as a correlation (e.g., reported as a beta value or t-test), convert the data to a correlation using equations provided.*	Code as continuous variable (report to two decimal points)

Appendix C

Outcome Variable Grouping Table

	Outcome measures: categories and subcategories				
General Distress (GD)	**Psychosocial Functioning (GPF)**	**Satisfaction with life and self (SLS)**	**Substance Use (SU)**	**Psychotic symptoms (PS)**	**Dropout (PreTer)**
Anxiety and related measures (ANX) • Penn State Worry Questionnaire (PSWQ) • Clinical Global Impression scale for PTSD (CGI-P) • Clinician Administered PTSD Scale (CAPS) • Horowitz Impact of Event Scale (IES) • Modified PTSD symptoms scale – self report (MPSS-SR)	**Interpersonal Functioning/Problems (IPF)** • Inventory of Interpersonal Problems (IIP) • The Social Adjustment Scale (SAS) • The Wisconsin Personality Inventory (WISPS) • Interpersonal Functioning Composite (IFC/TC) – TC, SCL, IIP, WISPI, GAS – both therapist and client) • UFB resolution scale (UFB) • Structural Analysis of Social Behavior (SASB) • SCID (PD scale)	**Self-Esteem** • Rosenberg's Self-Esteem Scale (SES or RSES) • Satisfaction with Life Scale (SWLS)	• Frequency of use (FSU) • Severity of use (SSU) • Days abstinent while in treatment (DAT)	• PANSS • Brief Psychiatric Rating Scale (BPRS)	• Dropout without definition (**DO/ND**) • Dropout without notification to therapist – not attending planned session, not calling back to make another appointment **(DO/WNT)** • Treatment duration criteria not met (**DO/CNM**) (e.g., 3 months, 16 sessions) • Missing a set number of sessions **(DO/MNS)** (e.g., missed 3 consecutive sessions)

Depression and related measures (DEP)	**General Functioning (GF)**				
• BDI • Hamilton Depression Rating Scale (HAM-D or HRSD)	• Global Assessment Scales (GAS) or Global Assessment of Functioning Scale (GAF) • Brief Psychiatric Rating Scale (BPRS) used to measure overall functioning • Overall Change Rating (OCR) • Target Complaints Method (TC) • Post therapy improvement (PTI) • Composite score of TC & PTI (COM-TC-PTI) • OQ • Schwartz Outcome Scale (SOS) • Anxiety Disorders				

	Interview Schedule for DSM-IV – used the Client Severity Rating (CSR) • Counseling Outcome Measure (COM**)** • The clinical outcomes in routine evaluation – outcome measure (CORE) • West Haven-Yale Multidimentional Pain Inventory (MPI-I) – used to measure how much pain interferes with functioning				
Other Symptoms of Distress (OSD) • Symptom check list 90-R; Symptom check list 10-R; Section of SCL-90R; GSI – section of SCL; Brief					

symptom Index (BSI) • Composite score of SCL and Target Complain symptoms (COM-TC-SCL) • General Distress Composite (GDC/TC) – TC, SCL, IIP, WISPI, GAS – both therapist and client) • Numeric Pain Rating Scale (NRS) • GHQ-28					

Conclusion

With the growing emphasis being placed on the use of evidence-based practice (EBP) within the field of professional psychology, clinicians are being strongly encouraged to use scientific data to inform and direct treatment (American Psychological Association Presidential Task Force on Evidence Based Practice, 2006). This push for EBP has also extended to the relationship between therapist and client, as APA has come to acknowledge the therapeutic alliance being a key component of the effectiveness of psychological treatment (APA, 2013). Having an evidence-based assessment tool is essential to practicing the evidence-based standards set out by APA and CPA. Without the use of a scientifically supported measure, the research, and the treatment practices, resulting from the date will be unsound. However, it is all too frequent that many of the assessment measures being used in psychological research have not been supported by sufficient scientific evidence (Hunsley & Mash, 2008). There seems to be a common misconception among researchers that psychometric properties derived from test development studies, or other studies, are immutable, and therefore are sufficient to establish psychometric applicability of the measure for any population or purpose (Vacha-Haase, Henson, & Caruso, 2002). In order for an assessment measure to be considered as evidence-based it needs to demonstrate solid scientific evidence of its psychometric suitability. This is established through ascertaining empirical evidence that illustrates that the measure yields scores that are reliable and valid for the specific sample and/or conditions for which it is being used.

Although research on the therapeutic alliance has been extensive, many measurement issues related to the reliability and validity of this construct's measures have not been fully addressed or resolved. One of the underlying reasons for the measurement

issues highlighted in the literature is related to the countless ways in which this broadly defined construct is measured. Out of the multitude of alliance measures that exist, no single measure contains representative items for the different components of the alliance concept. Therefore, it is difficult to say whether one alliance measure is more theoretically and psychometrically sound for measuring the alliance than any other measure. The purpose of this dissertation was to add to the scientific evidence for the therapeutic alliance by establishing empirical evidence of the psychometric properties of this construct's most commonly used measures, with the intention of identifying the most psychometrically sound alliance measures. This was first addressed by (a) exploring the typical score reliability of the therapeutic alliance measures most commonly used in the adult psychotherapy research literature and (b) identifying specific study characteristics that may influence these reliability estimates.

Six different alliance measures (ARM, CALPAS, CRF, HAq, TBS, WAI), in various formats and rater versions, were included in the first analyses, resulting in a total of 17 alliance measure variants being included in this reliability generalization (RG) study. In general, most measures were found to be good choices, on the basis of reported reliability estimates, for assessing the alliance. Out of these 17 measures, 10 measures (ARM-28-C, CALPAS-O, CRF-S-C, HAq-O, WAI-C, WAI-O, WAI-S-C, WAI-SR-C, WAI-S-T, WAI-T) showed evidence of having high reliability coefficients. Similar to previous research, the six of the WAI variants, the most frequently used measure, showed excellent reliability. However, the WAI-O, along with the ARM-28-C, CALPAS-O, CRF-S-C, HAq-O, had reliability estimates derived from relatively few studies, and therefore, the results from these measures should be viewed with appropriate caution. Also important to note is that three out of these five

alliance measures (WAI-O, CALPAS-O, HAq-O) are observer-rated alliance measure variants. The limited number of available data for these measures may be the result of researchers' tendency to focus on interrater reliability only when using observer data.

Although their overall reliability estimates were not high as the WAI, the ARM-5-T, CALPAS-24-C, HAq-I-C, and HAq-II-C also appeared to be good choices for assessing the alliance. Whereas, the ARM-5-C and the TBS-C only showed adequate reliability estimates and may not be the optimal choice for measuring the therapeutic alliance. With the exception of the ARM, no substantial differences were found between the therapist and client versions of the same measures.

When looking at the impact of sample characteristics on the reliability estimates, only six of the 17 alliance measure variants (CALAPS-24-C, WAI-C, WAI-S-C, WAI-SR-C, WAI-S-T, WAI-T) could be assessed for moderator analyses. The WAI-S-C, WAI-S-T, WAI-T were the most influenced by different sample characteristics, with moderating effects found for four different variables for each measure's reliability estimates. For the WAI-S-C, its reliability estimates were higher in samples with younger and less experienced therapists, where there was more variability among the client's score on the alliance measure, as well as in samples where treatment provided was either experiential or a combination of theoretical orientations. With regards to the WAI-S-T, this measure's reliability estimates were higher in samples with younger therapists and larger percentage of female therapists as well as with larger sample sizes for both the clients and therapists. As for the WAI-T, samples with younger therapists, more variability among the therapists' age, samples drawn from a community setting, and with smaller therapists' sample sizes all produced higher reliability estimates. As for the three other alliance measures, higher

reliability estimates were found for the WAI-SR-C in samples where clients scored higher on the alliance measure and when those scores were more dispersed, this later pattern also observed for the CALPAS-24-C. For the WAI-C, higher reliability estimates were found in studies that had older publication dates and samples that were being treated for depression or a personality disorder. Although researchers should consider these potential influences on score reliability, it should be noted that their influence will have less of a meaningful impact on a study's reliability estimate when the alliance is assessed from a measure with high reliability and narrow confidence interval levels.

As noted in previous research (e.g., Therrien, 2013), limiting reporting of study-specific reliability continues to be commonplace. In this RG, 56% of the studies included did not report reliability estimates based on their own sample, with 33% of authors relying on reliability induction and 23% failing to mention reliability altogether. Although these numbers are less than ideal, results from this RG may be pointing towards a shift in reporting practices, as the reporting rate is above the average observed in Vacha-Haase and Thompson's (2011) review of the RG literature. Here, they reported that on average, RG studies were based on 18.7% of the primary studies reporting their sample's reliability, whereas, this study's primary study reporting rate was 34.4%. In addition, this RG also observed an increase in reporting rates in recent years, with an average reporting rate of 14.5 studies per publication year from 2010 onward, compared to an average reporting rate of 5.7 studies per publication year for the years between 1982 to 2009. Although this increase is relatively small it does seem to suggest that reporting reliability estimates are moving in the right direction.

In addition to exploring the average reliability estimates of the alliance measures, in this dissertation I also assessed the predictive validity of the most commonly used alliance measures. This was done by exploring (a) the alliance-outcome association reported for the most commonly used therapeutic alliance measure in the adult psychotherapy literature, using only those studies that involved the provision of psychotherapy by professionals trained to provide this service (or those in the process of receiving such training), (b) assessing the alliance-outcome relation for full and partial correlation data separately, and (c) identifying specific study characteristics that might moderate the alliance-outcome relation. Five different alliance measures (CALPAS, CRF, HAq, WAI, VTAS) in various formats and rater versions were included in the analyses, resulting in a total of 15 alliance measure variants being included in this second study. Consistent with previous meta-analyses (Horvath & Symonds, 1991; Martin et al., 2000; Horvath & Bedi, 2002; Horvath et al., 2011), this study found an overall moderate effect for the alliance-outcome relation, with the magnitude of this effect being almost identical for studies using full correlation data (mean effect size of r = .24) and partial correlation data (mean effect size of r = .23). These results suggest that there is relatively no difference in the early alliance's ability to predict client functioning at the end of treatment versus its ability to predict how much the client's functioning has changed over the course of therapy.

Another important finding in this study was the homogeneity of variance evident across the results of studies included in this meta-analysis. This means there was no statistically significant variation in effect sizes across the different alliance measures, the different informants, or in the various ways treatment outcomes were assessed (i.e., client symptoms, level of psychosocial functioning, combination of symptoms and psychosocial

functioning, or premature termination). In other words, the size of the alliance-outcome relation was not influenced by any study, sample, treatment, or measure characteristic.

These results are in contrast to those reported by Horvath et al. (2011) who found a number of measurement effects on the alliance-outcome relation. However, it is likely that the difference in findings is due to the difference in inclusion criteria between the current study and Horvath et al. (2011). In this study I took a more conservative approach, using a more precise definition of what constituted "psychotherapy," an adult only sample, and only studies in which the alliance was assessed at mid-point of treatment or earlier. That being said, these results strongly suggest that when assessing the alliance-outcome relation based on studies of adult psychotherapy in which services were provided by appropriately trained mental health professionals, the association between therapeutic alliance and treatment outcome is consistent. Furthermore, having found no statistical significant difference across all forms of treatment, client presenting problems, measurements of alliance, and outcome variables suggests that the alliance may have a generalized effect on treatment outcome by increasing the likelihood that clients will stay in therapy and actively work on their problems.

What is very interesting to note is when considering the results of both studies, based on reliability and validity analyses, no single measure stood out from the rest as being superior for assessing the alliance. That is, with few exceptions, all measure variants appeared to have comparable strength of evidence supporting their reliability and validity. The results of this dissertation have both practical and conceptual implications. On a practical level, a researcher, or clinician intending to evaluate psychotherapy (as defined in this study) has many solid measures to choose from. From within these choices, a measure

can be selected that best fits the intended use (e.g., facets of the construct included in the measure, length of measure). However, from a theoretical standpoint, these findings present a quandary. If measures based on different conceptual models seem to perform similarly, it is hard to determine which conceptual model best represents the therapeutic construct. Although, differences between models might emerge with specific aspects of the therapeutic process being considered (e.g., predicting client levels of self-disclosure or engagement in between session assignments), there is no firm empirical basis for promoting one model over another. Although, the model endorsed by APA (i.e., goals, tasks, bond) captures the key components of the alliance that have been outlined in the literature, such as emphasis on collaboration and consensus between client and therapist, and has the most supporting evidence, it cannot be stated that it is the "best" measure based on scientific evidence. With that being said, these results appear to support the notion that there is one or more common underlying components to the therapeutic alliance, and that these components are present among all of the different alliance measures. It may well be that it is the components of collaboration and consensus that are being captured in the different alliance measures. That is, they are all measuring, to some degree, how well the client and therapist are working together towards solving the client's problems.

With regards to providing psychotherapy, these findings provide additional evidence of the importance for clinicians to place further emphasis on the therapeutic alliance when treating clients. However, as it does not seem to matter what conceptual alliance model one chooses to work from, as the different measures seems to be capturing a key facet of the alliance, with regards to treatment outcomes, it becomes more imperative

that professional training programs turn their attention to helping clinicians understand what factors, or variables, strengthen the alliance. Although both patient and therapist variables have been outlined in the literature (e.g., Horvath et al., 2011; Sharpless, Muran, & Barber, 2010; Muran, Safran, & Eubanks-Carter, 2011), it makes intuitive sense for training programs to focus on enhancing the therapist variables. For example, a number of these therapist variables that have been associated with stronger alliance, such as professional demeanor, therapeutic skills, confidence, honesty, warmth, empathy, and self-reflection, may be cultivated and refined within the context of supervision. The supervisory relationship between trainee and supervisor may also serve as model for how to explicitly address alliance issues with clients. Receiving specific training designed to develop a greater understand of what fosters a strong alliance and to enhance these alliance building skills will likely in turn increase the likelihood of better treatment outcomes.

Another important finding to highlight is that, although the alliance has been shown to fluctuate over the course of treatment, and that assessing the alliance at multiple times points may be a better predictor of treatment outcomes, early alliance seems to matter. Therefore, it is likely essential for clinicians to ensure they have establish a good alliance from early on and should be at a forefront of every clinician's mind, even during the assessment phase of psychotherapy. However, that is not to say that clinicians should not be continuing to assess the alliance throughout the course of treatment. From the rupture and repair research, it has been shown that development of the alliance is not always linear and that breakdowns in the collaboration, understanding, and communication do occur (Safran et al., 2011). Clients often have negative feelings related to being in therapy as well as to the therapist, and they may not always feel they are able to express them. Therefore, it

is important that clinicians be on the lookout for subtle indications of such ruptures throughout the course of therapy and take the initiative to explore where the breakdown may have occurred and what can be done to repair it (Safran et al., 2011). A way to facilitate this exploration is by having the clinicians make the alliance an explicit component to treatment by directly addressing it at the start of therapy with their clients. This may help to promote an open and honest communication between the therapist and client regarding the work they are doing together and, hopefully, this form of communication follows them throughout the course of therapy. Because, as research has repeatedly shown, having a good alliance likely increases the likelihood the client will experience positive treatment outcomes.

References

American Psychological Association Publications and Communications Board Working Group on Journal Article Reporting Standards (2008). Reporting standards for research in psychology: Why do we need them? What might they be? *American Psychologist, 63,* 839-851. doi: 10.1037/0003-066X.63.9.839

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

American Psychological Association. (2013). Recognition of psychotherapy effectiveness. *Psychotherapy, 50,* 102-109. http://dx.doi.org/10.1037/a0030276

Anderson, R., & Anderson, G. (1962). Development of an instrument for measuring rapport. *Personnel Guidance Journal, 41*, 18–24. doi: 10.1002/j.2164-4918.1962.tb02226.x

Ardito, R. B., & Rabellino, D. (2011). Therapeutic alliance and outcome of psychotherapy: Historical excursus, measurements, and prospects for research. *Frontiers in Psychology*, *2*, 270. http://doi.org/10.3389/fpsyg.2011.00270

Barrett-Lennard, G. T. (1986). The relationship inventory now: issues and advances in theory, method and use. In L. S. Greenberg & W.M. Pinsof (Eds.), *The psychotherapeutic process: a research handbook* (pp. 439–467). New York: Guilford Press.

Beck, A. T., Rush, J. J., Shaw, B. F., & Emery, G. (1979). *Cognitive therapy of depression.* New York: Guilford Press.

Berdondini, R. E., & Shearer, J. (2012). Collaboration in experiential therapy. *Journal*

of Clinical Psychology, 68, 159-167. doi: 10.1002/jclp.21830

Bibring, E. (1937). Symposium on the theory of the therapeutic results of psycho-analysis. *International Journal of Psycho-Analysis, 18*, 170–189.

Bordin, E. S. (1979). The generalizability of the psychoanalytic concept of the working alliance. *Psychotherapy: Theory, Research and Practice, 16*, 252–260. http://dx.doi.org/10.1037/h0085885

Borsboom, D., Mellenbergh, G.J., & van Heerden, J. (2004) The concept of validity. *Psychological Review*, 111, 1061-1071. http://psycnet.apa.org/doi/10.1037/0033-295X.111.4.1061

Castonguay, L. G., Constantino, M. J., McAleavey, A. A., & Goldfried, M. R. (2010). The therapeutic alliance in cognitive-behavioral therapy. In Muran, J. C. & Barber, J. P. (Eds). *The therapeutic alliance: An evidence-based guide to practice* (pp.150-171). New York: Guilford.

Coaley, K. (2009). *Introduction to psychological assessment and psychometrics.* London: SAGE Publications Ltd.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112,* 155–159.

Cook, D. A., & Beckman, T. J. (2005). Current concepts in validity and reliability for psychometric instruments: Theory and application. *The American Journal of Medicine, 119*, 166.e7-166.16. doi:10.1016/j.amjmed.2005.10.036

Crits-Christoph, P., Crits-Christoph, K., & Gibbons, M. B. C. (2010). Training in alliance-fostering techniques. In Muran, J. C. & Barber, J. P. (Eds). *The therapeutic alliance: An evidence-based guide to practice* (pp.304-319). New York, NY: Guilford

Dattilio, F. M., & Hanna, M. A. (2012). Collaboration in cognitive-behavioral therapy.

Journal of Clinical Psychogy, 68, 146-158. doi: 10.1002/jclp.21831

DeVon, H. A., Block, M. E., Moyle-Wright, P., Ernst, D. M., Hayden, S. J., Lazzara, D. J., ... & Kostas-Polston, E. (2007). A psychometric toolbox for testing validity and reliability. *Journal of Nursing Scholarship*, *39*, 155-164.

Diener, M. J., Hilsenroth, M. J., & Weinberger, J. (2009). A primer on meta-analysis of correlation coefficients: The relationship between patient-reported therapeutic alliance and adult attachment style as an illustration. *Psychological Research, 19,* 519-526. doi.org/10.1080/10503300802491410

Dimitrov, D. M. (2002). Reliability: Arguments for multiple perspectives and potential problems with generalization across studies. *Educational and Psychological Measurement, 62,* 783-801. doi: 10.1177/001316402236878

Downing, S. M. (2004). Reliability: on the reproducibility of assessment data. *Medical Education, 38,* 1006-1012.

Elliott, R., Watson, J. C., Goldman, R. N., & Greenberg, L. S. (2004). *Learning emotion-focused therapy. The process-experiential approach to change.* Washington, DC: American Psychological Association.

Elvins, R., and Green, J. (2008). The conceptualization and measurement of therapeutic alliance: an empirical review. *Clinical Psychological Review, 28,* 1167–1187. doi:10.1016/j.cpr.2008.04.002

Field, A. (2001). Meta-analysis of correlation coefficients: A Monte Carlo comparison of fixed-and random-effects methods. *Psychological Methods, 6,* 161-180. http://psycnet.apa.org/doi/10.1037/1082-989X.6.2.161

Ferguson, C. J. (2009). An effect size primer: A guide for clinicians and researchers.

Professional Psychology: Research and Practice, 40, 532-538. doi: 10.1037/a0015808

Flückiger, C., Del Re, A. C., Wampold, B. E., Symonds, D. & Horvath, A. O. (2011). How central is the alliance in psychotherapy? A multilevel longitudinal meta-analysis. *Journal of Counseling Psychology*. *59*, 10-17. doi: 10.1037/a0025749

Freud, S. (1971). "The case of Schreber paper on technique and other works." In James Strachey (Ed. And Trans.), *The standard edition of the complete psychological works of Sigmund Freud* (Vol. 12, pp. 99-108). London: The Hogarth Press. (Original work published in 1911-1913).

Furr, R. M., & Bacharach, V. R. (2008). *Psychometrics: An introduction* (2nd ed.). Los Angeles, CA: SAGE Publications, Inc.

Gaston, L., & Marmar, C. R. (1994). The California Psychotherapy Alliance Scales. In A. O. Horvath & L. S. Greenberg (Eds.), *The working alliance: Theory, research, and practice* (pp. 85-108). New York: John Wiley & Sons.

Geisinger, K. F. (2013). Reliability. In Geisinger, K. F., Bracken, B., Carlson, J. F., Hansen, J. C., Kuncel, N. R., Reise, S. P., & Rodriguez, M. C. (Eds.), *APA Handbook of Testing and Assessment in Psychology, Vol. 1: Test Theory and Testing and Assessment in Industrial and Organizational Psychology* (pp. 21-42). American Psychological Association. Washington, DC.

Graham, J. M., Diebels, K. J, & Barnow, Z.B. (2011). The reliability of relationship satisfaction: A reliability generalization meta-analysis. *Journal of Family Psychology, 25*, 39-48. doi/10.1037/a0022441

Graziano, A. M., & Raulin, M. L. (2004). *Research methods: A process of inquiry (5th ed.).* Boston, MA: Pearson Education Group.

Greenson, R. R. (1965). The working alliance and the transference neurosis. *Psychoanalytic Quarterly*, *34*, 155-181. doi: 10.1002/j.2167-4086.2008.tb00334.x

Greenson, R. R. (1967). *Technique and practice of psychoanalysis*. New York: International University Press.

Hatcher, R. L. (2010). Clinical studies of the therapeutic alliance. In Muran, J. C. & Barber, J. P. (Eds). *The therapeutic alliance: An evidence-based guide to practice* (pp.7-28). New York: Guilford Press.

Hatcher, R. L., & Barends, A. W. (1996). Patients' view of the alliance in psychotherapy: Exploratory factor analysis of three alliance measures. *Journal of Consulting and Clinical Psychology, 64*, 1326–1336. doi: 10.1037/0022-006X.64.6.1326

Haynes, S. N., Smith, G. T., & Hunsley, J. (2011). *Scientific foundations of clinical assessment.* New York: Taylor & Francis.

Hemphill, J. (2003). Interpreting the magnitudes of correlation coefficients. *American Psychologist, 58*(1),78-79. doi.org/10.1037/0003-066X.58.1.78

Henchy, A. M. (2013). *Review and evaluation of reliability generalization research.* (Unpublished dissertation). University of Kentucky, U.S.A.

Henson, R. K. (2001). Understanding internal consistency reliability estimates: A conceptual primer on coefficient alpha. *Measurement and Evaluation in Counseling and Development*, *34*, 177-189.

Horvath, A. O., & Bedi, R. P. (2002). The alliance. In J. C. Norcross.

Psychotherapy relationships that work (pp. 37-69). New York: Oxford University Press.

Horvath, A., Del Re, A. C., Flückiger, C., & Symonds, D. (2011). The alliance. In J. C. Norcross (Ed.). *Relationships that work* (pp. 25-69). New York: Oxford University Press.

Horvath, A. O., & Greenberg, L. S. (1989). Development and validation of the working alliance inventory. *Journal of Counseling Psychology, 36*, 223–233. doi.org/10.1080/10503300500352500

Horvath, A. O., & Symonds, B. D. (1991). Relation between working alliance and outcome in psychotherapy: A meta-analysis. *Journal of Counseling Psychology, 38,* 139-149. doi.org/10.1037/0022-0167.38.2.139

Howard, K. I., & Orlinsky, D. E. (1972). Psychotherapeutic processes. *Annual Review of Psychology, 23*, 615-668. doi: 10.1146/annurev.ps.23.020172.003151

Hunsley, J. & Lee, C. (2010). *Introduction to clinical psychology: An evidence based approach* (2nd Ed.). Toronto, ON: John Wiley & Sons Canada.

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

Hunsley, J. & Mash E. J. (2008). *A guide to assessments that work.* New York: Oxford University Press.

Hunsley, J., & Westmacott, R. (2007). Interpreting the magnitude of the placebo effect: mountain or Molehill? *Journal of Clinical Psychology*, *63*, 391-399. doi: 10.1002/jclp.20352

Kieffer, K. M., & MacDonald, G. (2011). Exploring factors that affect score reliability and variability in the Ways of Coping Questionnaire reliability coefficients: A meta-analytic reliability generalization study. *Journal of Individual Differences, 32*, 26-38. doi: 10.1027/1614-0001/a000031

Krause, M., Altimir, C., & Horvath, A. (2011). Reconstructing the therapeutic alliance: Reflections on the underlying dimensions of the concept. *Clinica y Salud, 22,* 267-283. doi.org/10.5093/cl2011v22n3a7

Luborsky, L. (1976). Helping alliances in psychotherapy. In J. L. Claghorn (Ed.), *Successful psychotherapy* (pp. 92–116). New York: Brunner Mazel.

Luborsky, L. B. (1994). Therapeutic alliances as predictors of psychotherapy outcomes: Factors explaining the predictive success. In A. O. Horvath & L. S. Greenberg (Eds.), *The working alliance: Theory, research, and practice* (pp. 38-50). New York: John Wiley & Sons.

Luborsky, L., Singer, B., & Luborsky, L. (1975). Comparative studies of psychotherapies; "Is it true that everybody has won and all must have prizes"? *Archives of General Psychiatry, 32*, 995-1008. doi:10.1001/archpsyc.1975.01760260059004.

Marmar, C. R., Horowitz, M. J., Weiss, D. S., & Marziali, E. (1986). Development of the therapeutic rating system. In L. S. Greenberg & W. M. Pinsof (Eds.), *The psychotherapeutic process: A research handbook* (pp. 367–390). New York: Guilford Press.

Marmar, C. R., Weiss, D. S., & Gaston, L. (1989). Towards the validation of the California Therapeutic Alliance Rating System. Psychological Assessment. *A Journal of*

Consulting and Clinical Psychology, *1*, 46–52. doi: 10.1037/1040-3590.1.1.46

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

Marziali, E., Marmar, C., & Krupnick, J. (1981). Therapeutic alliance scales: development and relationship to psychotherapy outcome. *Journal of Nervous and Mental Disease, 172,* 417–423. doi:10.1176/ajp.138.3.361

Messer, S. B., & Wolitzky, D. L. (2010). A psychodynamic perspective on the therapeutic alliance: Theory, research, and practice. In Muran, J. C. & Barber, J. P. (Eds). *The therapeutic alliance: An evidence-based guide to practice* (pp.97-122). New York: Guilford Press.

Miller, M. D. (2010). Classical test theory reliability. *International Encyclopedia of Education (3rd ed.),* 27-30. doi 10.1016/B978-0-08-044894-7.00235-9

Muran, J. C., & Barber, J. P. (2010). *The therapeutic alliance: An evidence-based guide to practice.* New York: Guilford Press.

Muran, J. C., Safran, J. D., & Eubanks-Carter, C. (2010). Developing therapist abilities to negotiate alliance ruptures. In Muran, J. C. & Barber, J. P. (Eds). *The therapeutic alliance: An evidence-based guide to practice* (pp.341-354). New York, NY: Guilford

Norcross, J. C. (Ed.) (2011). *Relationships that work.* New York: Oxford University Press.

Nimon K., Zientek L. R., Henson R. (2012). The assumption of a reliable instrument and other pitfalls to avoid when considering the reliability of data. *Frontiers in Psychology, 3,* 1-13. doi:102.10.3389/fpsyg.2012.0010

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill

Orlinsky, D. E., & Howard, K. I. (1975). *Varieties of psychotherapeutic experience: multivariate analysis of patients' and therapists' reports*. New York: Teachers College Press.

Orlinsky, D. E., & Rønnestad, M. H. (2000). Ironies in the history of psychotherapy research: Rogers, Bordin, and the shape of things that came. *Journal of Clinical Psychology, 56,* 841-851. doi: 10.1002/1097-4679(200007)56:7<841::AID-JCLP3>3.0.CO;2-V

Raue, P. J., & Goldfried, M. R. (1994). The therapeutic alliance in cognitive-behavior therapy. In A. O. Horvath & L. S. Greenberg (Eds), *The working alliance: Theory, research and practice* (pp. 131-152). New York: John Wiley & Sons.

Reinhardt, B. (1996). Factors affecting coefficient alpha: A mini Monte Carlo study. In B. Thompson (Ed.), *Advances in social science methodology* (Vol. 4, pp. 3-20). Greenwich,CT: JAI Press.

Rogers, C. R. (1951). *Client-centered therapy.* Boston: Houghton Mifflin.

Rogers, C. R. (1965). Client-centered therapy: its current practice, implications, and theory. Boston: Houghton Mifflin.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*, 638. doi/10.1037/0033-2909.86.3.638

Safran, J. D., & Muran, J. C. (2000). *Negotiating the therapeutic alliance: a relational treatment guide.* New York: Guilford Press

Safran, J. D., Muran, J. C., & Eubanks-Carter, C. (2011). Repairing alliance ruptures. *Psychotherapy, 48,* 80-87. doi: 10.1037/a0022140

Saunders, S. M., Howard, K., & Orlinsky, D. E. (1989). The Therapeutic Bond Scales:

Psychometric characteristics and relationship to treatment effectiveness. *Psychological Assessment: A Journal of Consulting and Clinical Psychology, 1*, 323-330.

Schmidt, F. L., & Hunter, J. E. (1977). Development of a general solution to the problem of validity generalization. *Journal of Applied Psychology, 62*, 529-540. doi: 10.1037/0021-9010.62.5.529

Sharf, J., Primavera, L. H., & Diener, M. J. (2010). Dropout and therapeutic alliance: A meta-analysis of adult individual psychotherapy. *Psychotherapy: Theory, Research, Practice, Training, 47,* 637-645. doi: 10.1037/a0021175

Sharpless, B. A., Muran, J. C., & Barber, J. P. (2010). Coda: Recommendations for practice and training. In Muran, J. C. & Barber, J. P. (Eds). *The therapeutic alliance: An evidence-based guide to practice* (pp.341-354). New York, NY: Guilford

Sterba, R. F. (1934). The fate of the ego in analytic therapy. *International Journal of Psychoanalysis, 115,* 117–126. doi: 10.1177/000306519404200310

Stiles, W. B., Shapiro, D., & Elliot, R. (1986). Are all psychotherapies equivalent? *American Psychologist, 41*, 165-180. doi/10.1037/0003-066X.41.2.165

Stone, L. (1961). The Psychoanalytic Situation: An Examination of Its Development and Essential nature. New York: International Universities Press

Suen, H. K. (1990). *Principles of Test Theories*. Hillsdale: Lawrence Erlbaum.

Swift, J. K., & Greenberg, R. P. (2012). Premature discontinuation in adult psychotherapy: A meta-analysis. *Journal of Consulting and Clinical Psychology*, *80*(4), 547. doi/10.1037/a0028226

Therrien-Poirier, Z. (2013). *Psychometric roperties of instruments used to assess*

anxiety in older adults. Unpublished doctoral dissertation, University of Ottawa, Ottawa, Canada.

Thompson, B. & Vacha-Haase, T. (2000). Psychometrics is datametrics: The test is not reliable. *Educational and Psychological Measurement, 60*, 174-195. doi: 10.1177/0013164400602002

Tyron, W. W., & Bernstein, D. (2003). Understanding measurement. In J. C. Thomas and M. Hersen (Eds.). *Understanding research in clinical and counseling psychology: A textbook* (pp. 27-68). Mahwah: Earlbaum.

Vacha-Haase, T. (1998). Reliability generalization: Exploring variance in measurement error affecting score reliability across studies. *Educational and Psychological Measurement*, *58,* 6-20. doi: 10.1177/0013164498058001002

Vacha-Haase, T., Henson, R., & Caruso, J. (2002). Reliability Generalization: Moving toward improved understanding and use of score reliability. *Educational and Psychological Measurement*, *62*, 562-569. doi: 10.1177/0013164402062004002

Vacha-Haase, T., Kogan, L. R. & Thompson, B. (2000). Sample composition and variabilities in published studies versus those in test manuals: Validity of score reliability inductions. *Educational and Psychological Measurement*, *60*, 509–552. doi: 10.1177/00131640021970682

Vacha-Haase, T., & Thompson, B. (2011). Score reliability: A retrospective look back at 12 years of reliability generalization studies*, 44,* 159-168. doi: 10.1177/0748175611409845

Wallerstein, R. S. (1986). *Forty-two lives in treatment: A study of psychoanalysis and psychotherapy.* New York: Guilford Press.

Watson, J. C., & Greenberg, L. S. (1994). The alliance in experiential therapy: Enacting the relationship conditions. In A. O. Horvath & L. S. Greenberg (Eds.), *The working alliance: Theory, research, and practice* (pp. 153-172). New York: John Wiley & Sons.

Watson, J. C., & Kalogerakos, F. (2010). The therapeutic alliance in humanistic psychotherapy. In Muran, J. C. & Barber, J. P. (Eds). *The therapeutic alliance: An evidence-based guide to practice* (pp.191-209). New York: Guilford Press.

Wilkinson, L., & Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*, 594-604. doi.org/10.1037/0003-066X.54.8.594

Wright, J. H., & Denise, D. (1994). The therapeutic relationship in cognitive-behavioral therapy: Patient perceptions and therapist responses. *Cognitive and Behavioral Practice, 1*, 25-45. doi/10.1016/S1077-7229(05)80085-9

Yin, P., & Fan, X. (2000). Assessing the reliability of Beck Depression Inventory scores: Reliability generalization across studies. *Educational and Psychological Measurement, 60*, 201-223. doi: 10.1177/00131640021970466

Zetzel, E. R. (1956). Current concepts of transference. International Journal of Psychoanalysis, 37, 369-37.

Zuroff, D. C., & Blatt, S. J. (2006). The therapeutic relationship in the brief treatment of depression: contributions to clinical improvement and enhanced adaptive capacities. *Journal of Consulting and Clinical Psychology, 74*, 130–140. doi: 10.1037/0022-006X.74.1.130