Development and Evaluation of a Measure of Collaborative Empiricism in Cognitive Therapy

Submitted by

John M. Tee BSc GDipPsych BPsySci (hons)

A thesis submitted in partial fulfilment

of the requirements for the degree of

Doctor of Clinical Psychology

School of Psychological Science

Faculty of Science, Technology and Engineering

La Trobe University
Bundoora, Victoria, 3086
Australia

July 2013

Acknowledgements

I would like to thank my supervisor, Dr. Nikolaos Kazantzis, for his invaluable assistance and guidance with the conduct and reporting of this research. His wealth of knowledge, resourcefulness, and unwavering optimism have been invaluable. Grateful acknowledgement is also given to Dr. Keith Dobson, who kindly provided access to archival recordings of therapy sessions, and gave important advice on methodology. I would also like to thank Dr. Art Stukas for his valuable input early in the project and his help with the development of the scale, Dr. David Farchione, from the La Trobe University Department of Statistics, for his assistance with the statistical models, and the academic and administrative staff of the School of Psychological Science at La Trobe University for their support throughout the years of the project. Thank you also to my fellow research students in the Cognitive Behaviour Therapy Research Unit at La Trobe University for their good cheer and wise words as we struggled forward together. Finally, I would like to thank my friends, for their continued support through the inevitable tough times, and for sharing the milestones along the way.

Table of Contents

List of Tables

List of Figures

Abstract

Collaborative empiricism (CE) is a key CBT process, in which the client and therapist collaborate to empirically explore the client's cognitions and cognitive processes. The aims of this study were to develop and psychometrically evaluate the first measure of CE in the literature, the Collaborative Empiricism Scale (CES), and use the measure to examine whether CE predicts reduction in depression severity in CBT for depression. The CES employs a novel design and rating methodology, allowing the rating of collaboration and empiricism at multiple levels, and at multiple times, within each therapy session. A theoretical model of CE is proposed that seeks to explain the motivational aspects of the construct, and integrate the collaborative and empirical elements of CE.

Three studies are reported. In Study 1, 30 expert CBT researchers and practitioners evaluated the draft CES via an online feedback questionnaire, in a process consistent with the Delphi method. Experts rated the CES very highly for content validity, ecological validity, and ability to capture variance in CE, and highly for clarity. Study 2 reports on pilot testing of the scale. In Study 3, psychometrics of the CES are reported, and CE is examined as a predictor of depression severity in a sample of 44 adults, at three time points, with a total of 3548 ratings of empiricism and collaboration across 132 therapy sessions. Reliability of the CES was high, and concurrent validity demonstrated with the Cognitive Therapy Rating Scale. Collaborative empiricism was found to predict reduction in depression severity at termination, after controlling for pre-therapy depression, the working alliance, and therapist competence. A medium effect size of CE on depression was obtained (R^2 = .088), larger than that reported in the literature for the working alliance or therapist competence. This effect was maintained at 6, 12, 18, and 24 months post-therapy.

Statement of Authorship

Except where reference is made in the text of the thesis, this thesis contains no material published elsewhere or extracted in whole or in part from a thesis submitted for the award of any other degree or diploma.

No other person's work has been used without due acknowledgment in the main text of the thesis.

All research procedures reported in this thesis were approved by the La Trobe University Faculty of Science, Technology and Engineering Human Ethics Committee, approval number FHEC09.R59.

This thesis has not been submitted for the award of any degree or diploma in any other tertiary institution.

John M. Tee

Date: 01/07/2013

Chapter 1: Collaborative Empiricism in the Literature[1]

Aims and outline

This chapter focuses on key aspects of the psychotherapy literature relevant to collaborative empiricism (CE) in cognitive-behaviour therapy (CBT). First, a definition of CE is derived from the seminal CBT literature. Inconsistencies and contradictions in the literature regarding the role and scope of CE are reviewed. The relationship between CE and the therapeutic alliance as operationalized in CBT research is then discussed, followed by an analysis of the inability of the most common form of the therapeutic alliance in psychotherapy research, Bordin's working alliance, to capture the CE construct. The possibility is considered that past alliance research has failed to adequately account for the role of CE in the alliance. This leads to consideration of CE within the frameworks of evidence-based relationship attributes and core competencies in CBT, with a focus on the work of the APA Task Force on Evidence-Based Therapy Relationships. The role of CE as a change process is then discussed, including its potential role as a mediator of therapeutic outcomes. Existing theory relevant to CE is reviewed, and a new theoretical model is proposed to explain the motivational role of CE in therapeutic change.

Collaborative Empiricism in the Literature

A contradiction awaits the reader who explores the literature on collaborative empiricism. On the one hand, there is 40 years of widespread agreement regarding the central role of CE in Beckian CBT (A. T. Beck, Emery, & Greenberg, 1985; A. T. Beck, Rush, Shaw, & Emery, 1979; S. D. Hollon & Beck, 1979; Kuyken, Padesky, & Dudley, 2009; Madsen, 2009; Persons, 2008; W. W. Tryon & Misurell, 2008). On the other, there has been little theoretical analysis and no specific empirical examination

[1] This chapter and the theoretical model of CE described within were published as Tee, J., & Kazantzis, N. (2011). Collaborative empiricism in cognitive therapy: A definition and theory for the relationship construct. *Clinical Psychology: Science and Practice*, *18*, 47-61.

of the construct. Collaborative empiricism is frequently referred to as a core component of the therapeutic alliance in CBT, and as a distinctive and defining characteristic of CBT process (e.g., A. T. Beck, 1989; Bishop & Fish, 1999; Cross, 2007; Dattilio, 2000; Dudley & Kuyken, 2006; Friedberg & Clark, 2006; Friedberg & Gorman, 2007; Kuyken, Padesky, & Dudley, 2008; Merali & Lynch, 1997; Overholser, 1995; Rutter & Friedberg, 1999; Segal & Shaw, 1996; Will, 1995). Despite this, it remains unclear whether CE is related to treatment outcomes, or moderates or mediates other processes. Importantly, there is no measure of CE in the literature, which may partly explain the lack of empirical work. A measure of CE would also aid in addressing recent calls for empirical data to guide the content and evaluation of clinical training programs for CBT (Simons, Rozek, & Serrano, 2013). This chapter aims to address these incongruities, to build a case for the development of the first measure of CE, and to propose a first theoretical model of CE, based on Self-Determination Theory, that is capable of integrating the collaborative and empirical aspects of CE, and which addresses the motivational aspect of the construct.

A Definition of Collaborative Empiricism

Collaborative empiricism is not a new concept in CBT. The fundamental ideas underpinning it are present in Beck's early work on the cognitive theory of depression (A. T. Beck, 1967). The term 'collaborative empiricism' appears in Beck and colleagues' work a few years later (A. T. Beck et al., 1979; S. D. Hollon & Beck, 1979). These seminal texts characterize CE as a CBT-specific feature of the therapeutic relationship that distinguishes CBT from other therapy approaches (A. T. Beck et al., 1985; A. T. Beck et al., 1979). Collaborative empiricism is described as an active collaboration in which the client and therapist work together to identify problem situations, and to design, implement, and evaluate empirical explorations of

the client's beliefs (S. D. Hollon & Beck, 1979). Sharing the work of therapy in CE is not simply agreement or co-operation of the client with the plans of the therapist (Alford & Beck, 1994). A key objective is for CE to be applied by the client outside therapy, and to bring the results of between sessions work back to subsequent sessions as data for further investigation (A. T. Beck, 1970; J. S. Beck, 1995). There is also a pervasive role for the collection and use of data within sessions. This empirical focus is not just on the use of data to evaluate beliefs, but flavours the whole interaction between client and therapist, including, for example, framing between session work as the collection and evaluation of data, and the use of data gathered in session as the basis of treatment planning and review of the therapeutic interventions (A. T. Beck et al., 1979).

These early descriptions of CE have had a profound and enduring influence on later discussions of the construct, and will be adopted here as the definition of CE. A more detailed operational definition of the construct will be outlined following an analysis of the components of CE in Chapter 2 (on collaboration) and Chapter 3 (on empiricism).

Inconsistencies in the Literature

Recent discussions of the role of CE in CBT have reinforced the fundamental importance ascribed to it in early work (A. T. Beck & Dozois, 2011; Deborah J. G. Dobson & Dobson, in press; Kazantzis, Freeman, Fruzzetti, Persons, & Smucker, 2013; Kuyken et al., 2008, 2009; Overholser, 2011; Persons, 2008; W. W. Tryon & Misurell, 2008). However, despite long standing consensus in the literature regarding the importance of CE, contradictions and incompatibilities have emerged in the roles and characteristics ascribed to the construct. Importantly, the discussion has remained largely based on practical, common sense grounds, rather than empirical evaluation.

The fundamental role attributed to CE is reflected in definitions in which it is described as the "distinguishing feature" (van Oppen, 2004, p. 339) or "hallmark of cognitive therapy" (Waddington, 2002, p. 186). Sometimes, however, CE is not just a distinguishing feature but a necessary and ubiquitous one: the "key to effective psychotherapy" (Westefeld et al., 2000, p. 479), "crucial for effective therapy" (Hazlett-Stevens, 2008, p. 68), "embedded in every clinical action and decision" (Friedberg & McClure, 2002, p. 44), which comprises the "cornerstone of CBT", and is "used throughout the entire course of treatment" (Dattilio, 2000, p. 39). Despite agreement on the fundamental importance of the construct, it is unclear whether CE is to be understood as one unique feature of CBT among many, or whether it is so primary and integral that it is necessary for effective CBT to proceed.

There are also inconsistencies regarding whether CE is a specific technique or a 'common factor' that pervades the entire therapeutic relationship. In some cases CE is referred to as the "major therapeutic method used by Beck" (Abramson, Alloy, & Dykman, 1990, p. 198), a "method of reality testing" (Bell, Grech, Maiden, Halligan, & Ellis, 2005, p. 148), or a "cognitive strategy" (Dattilio, 2000, p. 39). In others, rather than a method or specific technique, it is a "philosophy which forms the foundation of cognitive therapy" (Wilkes, 1994, p. 309). Sometimes it is both a "process" and an "approach" (Gleeson & McGorry, 2004, pp. 209-211), both a "specific technique" and an "atmosphere" that pervades the client-therapist relationship (Turkington & Siddle, 1998, p. 237). To date, there has been neither recognition nor discussion of these discrepancies in the literature.

Collaborative Empiricism and the Therapeutic Alliance

Inconsistency is also evident regarding the relationship between CE and the therapeutic alliance. The therapeutic alliance can be broadly defined as the alliance of

a client's reasonable side with a therapist's working side (C. J. Gelso & Hayes, 1998; Horvath & Greenberg, 1994). In some cases CE is described in the literature as "the ideal cognitive therapy alliance" (Bishop & Fish, 1999, p. 117), or as the "basis of the working alliance" (A. T. Beck et al., 1985, p. 37). In other cases it is a "cardinal feature of the therapeutic relationship in cognitive therapy" (K. S. Dobson & Shaw, 1988, p. 674), and in others it is neither the relationship, nor a component of it, but a "stylistic fulcrum that permits the helping alliance to thrive" (Stein, Kupfer, & Schatzberg, 2006, p. 359).

Clarifying the relationship between CE and the therapeutic alliance is important because there is reliable meta-analytic evidence of a robust relationship between therapeutic alliance and treatment outcomes in CBT (Horvath, Del Re, Fluckiger, & Symonds, 2011; Horvath & Symonds, 1991; Martin, Garske, & Davis, 2000; Shirk & Karver, 2003). The relationship between alliance and outcomes holds across different therapies and disorders, despite whether outcome or alliance are rated by therapist, client, or observer (Martin et al., 2000; Shirk & Karver, 2003).

The working alliance.

To examine the relationship between CE and the therapeutic alliance requires an understanding of the construct of the therapeutic alliance and the measures used to study it in CBT. The conception of the alliance most often used in research studies of CBT is the *working alliance* developed by Bordin (Bordin, 1979, 1994). Bordin (1979) proposed a trans-theoretical model of the working alliance as a collaborative, negotiated, and interactive relationship between client and therapist. In Bordin's view, the working alliance consists of three core components: an interpersonal, affective *bond* between therapist and client, agreement on the *goals* of therapy, and

collaboration on *tasks* designed to address the client's therapeutic goals (Bordin, 1979).

Bordin described collaboration as a 'working relationship' between client and therapist (Bordin, 1979, 1994). At face value, this working relationship seems similar to the meaning of collaboration as 'sharing the work' in CE. However, Bordin conceived of collaboration in the working alliance as an "agreed-upon contract" between client and therapist regarding the tasks and goals of therapy (Bordin, 1994, p. 254). For Bordin, agreement is "the key process in building an initial, viable alliance" (Horvath & Greenberg, 1994, p. 1). In contrast, collaboration in CE is not simply willing participation by the client, nor agreement on tasks or goals. In CE, the therapist aims to engage the client in a genuine sharing of the work of goal setting and creative authorship of therapeutic tasks, progressively encouraging the client to take the lead role in these activities as far as is practicable (A. T. Beck et al., 1985; A. T. Beck et al., 1979; J. S. Beck, 1995, 2011). The aim of this partnership is to actively share the work of empirically exploring and evaluating the client's cognitions and cognitive processes, an aim that goes beyond simple agreement on a program of therapeutic tasks and goals.

The Working Alliance Inventory.

If the concept of collaboration in Bordin's (1979) working alliance fails to capture the meaning of collaboration in CE, it is possible that measures of the alliance based on Bordin's conception may similarly fail to capture the meaning of collaboration in CE. If so, past research on the working alliance in CBT that has employed such measures may not have adequately accounted for the role of CE. This possibility is demonstrated here by an analysis of the most commonly used measure

of the working alliance in CBT, the Working Alliance Inventory (WAI; Horvath & Greenberg, 1989, 1994).

The WAI is a 36 item questionnaire designed specifically to measure Bordin's working alliance (Horvath & Greenberg, 1989, 1994). It has been widely used in alliance research, across many types of therapy (Horvath, 2001; Horvath & Greenberg, 1989, 1994; Horvath & Luborsky, 1993; Horvath & Symonds, 1991; Martin et al., 2000). A commonly used version, the 12-item WAI short form revised (WAI-SR; Hatcher & Gillaspy, 2006), has improved psychometric properties and better differentiates Bordin's three alliance dimensions of bond, tasks, and goals. The WAI-SR items are presented in Appendix H. Analysis of the WAI-SR reveals that none of the items measure collaboration (or indeed empiricism) as defined in CE. No items in either the Task or Bond subscales refer to collaboration. Three items in the 4-item Goal subscale do refer to 'agreement' and 'good understanding', but no other aspects of collaboration are discussed. The third item refers to 'collaboration in setting goals' but this phrase also does not capture the meaning of collaboration as an active sharing of the work of therapy that is central to CE. Collaboration as sharing the work of empirically validating the client's beliefs is largely missing from the WAI-SR. Both the most frequently used construct of the therapeutic alliance in CBT (Bordin's working alliance), and a frequently used instrument developed to measure it (the WAI-SR), fail to capture the key aspects of collaboration or empiricism as described in CE.

Evidence-Based Therapy Relationships.

The two APA Task Forces on Evidence-Based Therapy Relationships (Norcross, 2001; Norcross & Lambert, 2011) conducted a landmark series of meta-analyses of psychotherapy outcome research studies, with the aim of identifying

evidence-based attributes of the therapeutic relationship. The Task Force identified three relationship attributes related to collaboration as either 'demonstrably effective' or 'probably effective'; however, as will be discussed in detail in Chapter 2, the processes identified do not capture collaboration in CE. The Task Force recommended that future research investigate further relationship attributes (Norcross & Wampold, 2011). Chapter 8 of this thesis takes up the Task Force's recommendation in an examination of the relationship between CE and therapeutic outcome.

Collaborative Empiricism as a Change Process in Cognitive Behaviour Therapy

Collaborative empiricism has been hypothesized to act as a change process in CBT (D. J. G. Dobson & Dobson, 2009; Kuyken et al., 2008, 2009). There is some support for this hypothesis in the literature. An example of this is the use of CE by clients to evaluate the accuracy of their own delusional thinking (e.g., Alford & Beck, 1994; Bell et al., 2005; Chadwick, Birchwood, & Trower, 1996). Traditionally, change in CBT has been understood via the cognitive mediation hypothesis (A. T. Beck, 1970; DeRubeis, Tang, & Beck, 2001). In this view, improvements in therapeutic outcome result from direct change in clients' cognitions. Underlying the use of CE in the treatment of delusions is the idea that belief change is more likely if the rationale for change comes from clients' own experience-driven insight, rather than from the therapist (Dattilio & Padesky, 1990; Hutton & Morrison, in press). Self-evaluation of clients' beliefs, using experiments they collaboratively design, avoids the confrontation implicit in directly challenging delusional beliefs and hallucinations (Alford & Beck, 1994). Self-efficacy may be lower if the therapist is seen as the change agent, increasing the probability of relapse (Horvath & Greenberg, 1994). Collaborative empiricism may also function in the change process by reframing CBT

as a shared, challenging, empirical experiment (Klosko & Sanderson, 1999). This may increase psychological activation and a sense of hope, which may as important as specific cognitive or behavioural tasks (Cross, 2007). The atmosphere of curiosity modelled in CE may also foster flexible thinking, encourage a broader examination of the client's problems, and allow the client and therapist to get at core subjective experiences that clients may not report (Friedberg & McClure, 2002). Better access to the client's problems and subjective experiences may in turn allow a more focused and productive examination of the client's cognitions, which has been shown to improve outcomes in CBT of anxiety disorders (Butler, Fennell, & Hackmann, 2008).

Existing Theory and Collaborative Empiricism

There has been little theoretical work specifically investigating the mechanisms underpinning CE, or attempting to explain how or why CE may influence therapeutic outcomes. This lack of analysis is not uncommon in process research (Orlinsky, Ronnestad, & Willutzki, 2004). Nevertheless, it is a shortcoming of a literature in which CE is widely described as a fundamental process. This section will evaluate the degree to which it is possible for well-validated aspects of psychological theory to underpin a theory of CE, and the extent to which theory can support the roles ascribed or hypothesized for CE in the literature.

Kuyken, Padesky, and Dudley (2009) conducted the only theoretical analysis to date specifically incorporating CE. They considered CE to be one of three fundamental principles informing the case conceptualization process in CBT. Although they focused on the role of CE in case conceptualization, their approach is relevant to an analysis of CE itself. They argued that the collaboration and empiricism components of CE function in tandem as the driving force underlying the process of case conceptualization. In case conceptualization, the client's unique experience

determines how both the theory and empirically validated principles of CBT are applied to that individual in their particular circumstances (J. S. Beck, 1995, 2011; Charlesworth & Greenfield, 2004; Persons, 2008). In Kuyken et al.'s account, the genuine collaboration inherent in CE aids the development of an accurate shared understanding of the client's problem situation. Both the client and therapist have important information to contribute to understanding the client's presenting situation in its unique historical, social, cultural, developmental, and psychological contexts. Increasing the accuracy of this shared understanding increases the amount of useful and relevant information available to develop the conceptualization. The empirical aspect of CE then functions as a check on the accuracy of the information generated in collaboration. Aspects of the conceptualization that do not fit the client's situation can be revised or discarded after empirical testing. The process of CE firstly increases the range of information available to the case conceptualization via collaboration, and then filters out inaccurate information and confirms accurate information via empirical testing. In this view, CE mediates the relationship between the case conceptualization and therapeutic outcomes by increasing the relevance and accuracy of the conceptualization.

Collaborative empiricism may also increase the accuracy of the case conceptualization by overcoming heuristic biases inherent in therapists' interpretations of their clients' problem situations (Bieling & Kuyken, 2003; Kuyken et al., 2009). In developing case conceptualizations, therapists have working hypotheses about their client's difficulties, and interpret information arising in session in terms of these hypotheses. Therapists, like all people, are prone to overreliance on or misapplication of heuristics in decision making (Kahneman, 2003). The therapist who decides, for example, that a client has Panic Disorder may selectively discount

information contrary to this hypothesis. It is practically impossible to entirely avoid the effects of these biases (Baron, 2007). However, the structured collaboration of CE, combined with explicit empirical testing of the client's cognitions, may reduce the impact of these biases or distortions in cognitive processing (Bieling & Kuyken, 2003).

In a related point, CE may also reduce erroneous or biased recall on the part of either client or therapist, by subjecting recall to ongoing empirical validation (S. D. Hollon, 1999). Reducing biased recall by the therapist may increase the accuracy of the case conceptualization, which may then improve outcomes as discussed above. It also seems reasonable that reducing the client's biased recall may increase the accuracy of the client's input into both the collaborative and empirical components of CE process. For example, an important aspect of CBT involves the client learning to objectively evaluate their own cognitions. Collaborative empiricism may increase the objectivity of this evaluation by reducing reliance on potentially biased recall, and by providing the client with a systematic method of objectively evaluating their beliefs (S. D. Hollon & Kriss, 1984). The merit of these approaches is that they articulate mechanisms by which case conceptualization may be related to CE. However, further work is required to develop a plausible theoretical account of CE per se.

Behavioural experiments.

A theoretical account of CE is informed by the literature on behavioural experiments. Behavioural experiments are empirical tests of clients' beliefs that clients undertake themselves, usually between therapy sessions (Bennett-Levy, 2003; Bennett-Levy et al., 2004). It is widely held in the literature that behavioural experiments affect change in CBT by providing evidence against maladaptive beliefs, and for new, more adaptive ones (A. T. Beck et al., 1979; J. S. Beck, 1995; Bennett-

Levy et al., 2004; D. J. G. Dobson & Dobson, 2009). If collaboratively designed and evaluated, behavioural experiments are an important and common example of CE in action. As such, existing theoretical accounts of behavioural experiments may inform a theoretical account of CE.

It has been argued that behavioural experiments may be more effective than techniques such as thought records because behavioural experiments manifest greater synchrony across cognitive, affective, and behavioural systems (Bennett-Levy, 2003; Bennett-Levy et al., 2004). This assertion can be supported by theories in cognitive science and learning theory (Bennett-Levy et al., 2004). According to the interacting cognitive subsystems model (ICS; Bennett-Levy et al., 2004; Teasdale, 1997a, 1997b), behavioural experiments provide powerful evidence for or against beliefs because they activate the non-linguistic, emotion-linked, automatic information processing system, rather than only the verbal, propositional, and rational information processing system. The model proposes that evidence from behavioural experiments thereby has a deeper 'felt sense', and greater subjective weight.

Teasdale (1997b) argued that change in beliefs requires the development of alternative schematic models at the deeper emotional level of processing. In this view, behavioural experiments provide experiences in which these new models may be created directly. The ICS model is supported by evidence from memory research (Bennett-Levy et al., 2004). Memory is usually better for self-performed actions – the so-called 'enactment effect' (Engelkamp, 1988; Helstrup, 2004; Masumoto et al., 2006), and is facilitated by stronger emotion (Bradley, Greenwald, Petry, & Lang, 1992; Christianson, 1992; S. B. Hamann, 2001). This view also accords with the common clinical practice of selecting the 'hot thought' as the focus of cognitive

restructuring with thought records. The hot thought is precisely the thought that elicits the strongest (most distressing) emotion in the client.

The value of behavioural experiments can be further supported by the experiential learning model of Lewin, Kolb, and colleagues (Kolb, 1984; Lewin, 1946). In this model, learning occurs in a cycle through four stages: plan, experience, observe, and reflect. A typical CBT session might begin with a shared case conceptualization (reflect), used to develop a behavioural experiment (plan), which the client completes (experience), and then (observes) the results. The point here is that behavioural experiments involve more than just carrying out an experiment. The cycle of learning incorporates experiential learning in the context of procedural planning.

This theoretical work provides a plausible mechanism for the efficacy of behavioural experiments, and by extension the empirical component of CE. Although this work goes some way towards an explanation of 'how' empirical investigation may affect therapeutic outcomes, it does not explain 'why' a client may be motivated to engage in the usually arduous process of changing their beliefs. Many clients, perhaps most, come to therapy driven by a pain they wish to relieve. However, as the motivational interviewing literature points out, clients often have conflicting motivations for engaging in therapy (W. R. Miller & Rollnick, 2002). Change may offer greater freedom or less pain, but comes with its own costs. A theory that explains this motivational dimension of CE is lacking. In addition, while the theoretical work discussed above is relevant to elements of CE, a theory that integrates the collaboration and empiricism aspects of CE with its motivational dimension is needed.

Towards a Theory of Collaborative Empiricism

Self-Determination Theory.

A theoretical foundation for CE that addresses the motivational aspect of the construct is proposed here using Self-Determination Theory (SDT; Deci & Ryan, 2000). Self-Determination Theory states that people have a fundamental propensity towards growth and the resolution of psychological discrepancy (Deci & Ryan, 1980, 1985; R. M. Ryan & Deci, 2000, 2002). The theory seeks to explain how individuals self-regulate behaviour and internalize extrinsic motivation in order to engage in autonomous action (Markland, Ryan, Tobin, & Rollnick, 2005; R. M. Ryan & Deci, 2000; Vansteenkiste & Sheldon, 2006). It is also concerned with the effect of environmental factors, such as directives or rewards, on intrinsic motivation to engage in behaviours (Deci & Ryan, 2000).

Self-Determination Theory proposes that behaviours are regulated on a continuum of autonomy, ranging from intrinsic (autonomous) regulation, to introjected regulation, and then external regulation. Autonomy refers to the locus of causality. External regulation denotes a behaviour that is regulated by external factors such as rewards or punishments. A direct attempt to persuade a client to change a belief is an example of an externally regulated motivation. Introjected regulation refers to motivation that is internalized. In this case, the pressure or drive to engage in a behaviour comes from within the person in the form of self-approval when engaging successfully in the behaviour and shame or self-criticism when failing to engage successfully in the behaviour (R. M. Ryan & Deci, 2000). Introjected regulation is more likely to lead to sustained engagement in behaviour than external regulation (Deci & Ryan, 2000); however, it can be accompanied by significant ambivalence, as well as negative affect and inner conflict (Markland et al., 2005; R. M. Ryan, Rigby,

& King, 1993). At the other end of the spectrum is intrinsic motivation. Intrinsically motivated behaviours are engaged in willingly, and are experienced as rewarding of themselves (Deci & Ryan, 1980). Self-Determination Theory proposes that three key needs underlie the innate tendency towards growth: competence, autonomy, and relatedness (Deci & Ryan, 2000). That is, humans have an innate need to experience themselves as competent, autonomous actors, who are related to others in meaningful and satisfying relationships. These factors work to increase the level of autonomous motivation associated with behaviour.

Empirical support for Self-Determination Theory.

There is considerable support for SDT in the social psychology literature. Autonomous regulation of behaviour is performed with higher quality and attention (Deci & Ryan, 1980), is more stable over time (Deci & Ryan, 1985), and is associated with higher positive affect (Deci & Ryan, 1980; R. M. Ryan & Deci, 2002). Autonomous motivation has also been shown to be related to better outcomes in psychotherapy and medical care (for review, see Markland et al., 2005).

According to SDT, supporting the client's sense of autonomy is required to augment competence and relatedness and promote optimal motivation for change (Deci & Ryan, 2000; Markland et al., 2005). Self-efficacy, the belief that one is competent to engage in a behaviour, is not sufficient to maximize motivation. Clients may feel able to engage in behaviour but not feel motivated to do so. According to SDT, increasing the level of autonomous regulation of a behaviour is required to maximize motivation to engage in the behaviour (Deci & Ryan, 1985; Reeve, 1998, 2002). There is evidence for the ability of at least four environmental conditions to support client's autonomy: a meaningful rationale for the behaviour, minimal external contingent reinforcers (rewards and punishments), active participation and exercising

of choice, and acceptance and acknowledgement of negative feelings (Markland et al., 2005; Reeve, 1998, 2002).

Collaborative empiricism and self-determination.

Of the four environmental conditions described above that are hypothesized in SDT to support client's autonomy, CE can be seen to directly influence the first three, and the fourth as part of general good CBT practice. A meaningful rationale for changing beliefs arises from the client's experience of the results of the empirical test of their beliefs. By definition, a good empirical test of the client's beliefs is one in which alternative explanations for data resulting from the experiment have been accounted for in the design of the experiment (e.g., J. S. Beck, 1995). The elimination of alternative explanations of the results of experiments increases the validity of the data. It is suggested here that it also increases the degree of internally autonomous relevance of the data for the client. The client experiences this data as intrinsically his or her own. The meaning of the data is judged using criteria that the client decides upon before the experiment is conducted. The source of these criteria is the client, and consequently the source of the regulatory capacity of the criteria is internal to the client. Reasons for changing beliefs are ideally experienced as the client's own reasons, rather than the therapist's reasons. As a result, the data is more likely to be experienced as an internal and autonomous source of regulation. According to SDT, this is likely to increase the motivation of the client to change their belief. In addition, the process of collaboration minimizes external contingent reinforcers, further increasing internal motivation. Although CE is directive in the sense that the client is directed to engage in the process of CE, the CE process itself is genuinely collaborative. The client is encouraged to participate in all aspects of CE and to

actively exercise choice and creativity in the design and evaluation of empirical tests of their beliefs.

It is suggested here that SDT can provide a theoretical basis for the efficacy of CE in belief change. In this view, CE encourages autonomously motivated belief change by promoting the acceptance and internalization of new beliefs, and by increasing the meaningfulness of new beliefs for the client.

Conclusion

It is likely that existing measures of the alliance in CBT fail to capture the core meaning of collaboration in CE, casting doubt on whether past research using these measures has adequately accounted for the role of CE in CBT. A measure specific to CE is required to address these issues. Existing theory provides a plausible explanation for elements of the collaborative and empiricism components of CE, but remains piecemeal and ignores the motivational role of the construct. A new theoretical model of CE has been proposed, drawing on well-validated social-cognitive theory, which integrates the collaborative and empirical components of the construct and its motivational aspect. There is some support in the literature for the efficacy of CE as a change process in CBT. However, no theoretical or empirical work to date incorporates the core elements of the construct.

The next two chapters will examine collaboration and empiricism in CE in detail, including aspects of CE for which there is empirical support in the literature, followed by a detailed analyses of therapist and client behaviours related to collaboration and empiricism in CE.

Chapter 2: Collaboration in Psychotherapy

Aims and outline

This chapter reviews the theoretical and research literature relevant to collaboration in CE. First, collaboration in psychotherapy is discussed in relation to the therapeutic alliance, and then as an empirically supported relationship attribute. Conceptions of collaboration in the existing literature are then compared with collaboration in CE and important differences examined. Finally, the key elements of collaboration in CE, as an active sharing of the work of therapy, are described.

Conceptual Work

Collaboration and the therapeutic alliance.

Collaboration has been considered in the research literature of the last 30 years primarily as an attribute of the therapeutic alliance (Castonguay, Constantino, & Holtforth, 2006; Clemence, Hilsenroth, Ackerman, Strassle, & Handler, 2005; Del Re, Flückiger, Horvath, Symonds, & Wampold, 2012; Elvins & Green, 2008; Hatcher, 2010; Horvath, 2006; Horvath et al., 2011; Hougaard, 1994; Leahy, 2008; Lejuez, Hopko, Levine, Gholkar, & Collins, 2005; Norcross & Lambert, 2011). The therapeutic alliance is an aspect of the relationship between client and therapist that supports the capacity of the client to engage with the therapist in the productive tasks of therapy (Dryden, 1989; Gaston, Thompson, Gallagher, Cournoyer, & Gagnon, 1998; Hatcher, 2010; Horvath, 2000, 2001, 2005, 2006; Lambert & Barley, 2001; Leahy, 2008). Research has demonstrated a reliable association between the alliance and therapeutic outcomes, with an effect size of approximately $r = .26$ (Horvath & Bedi, 2002; Horvath & Symonds, 1991; Martin et al., 2000; Shirk & Karver, 2003). By convention, an r of .26 is a medium size effect in psychological research (J. Cohen, 1988). The nature of the association between alliance and outcomes, however,

and the causal role of the alliance in improving outcomes, remains less clear (Crits-Christoph, Connolly Gibbons, & Hearon, 2006; Crits-Christoph, Gibbons, & Hearon, 2006; Del Re et al., 2012; DeRubeis, Brotman, & Gibbons, 2005; Hatcher, 2010; Horvath et al., 2011; Safran & Muran, 2006). Previous research has found that therapist characteristics and therapist variability in the alliance are related to alliance formation and therapeutic outcome (e.g., Ackerman & Hilsenroth, 2001, 2003; Baldwin, Wampold, & Imel, 2007; Ulvenes et al., 2012). This suggests that therapists' behaviours related to alliance formation, such as collaboration, are important for outcomes in therapy.

The idea of an alliance between client and therapist developed from Freud's 'analytic pact' between a client's 'reasonable' or 'working' side and the therapist (Freud, 1912; Horvath et al., 2011; Kanzer, 1981; Wiseman, Tishby, & Barber, 2012). The structure of Freud's therapy, in particular working with the transference in the relationship, highlighted the need for a 'working relationship' outside the transference, within which therapeutic dialogue could take place. That is, the idea of the working relationship arose for Freud as a necessary consequence of the mechanism of change of the therapy. The concept of the therapeutic alliance was subsequently developed further by Greenson, who distinguished the affective bond between client and therapist from the alliance (Greenson, 1965). Greenson argued that an affective bond between the client and therapist was necessary to promote the therapeutic alliance and thereby foster the client's engagement in the tasks of therapy.

Bordin subsequently sought to generalize the concept of the alliance from its psychoanalytic roots, by grounding the alliance in a pan-theoretical conception of therapy as 'purposeful work' (Bordin, 1979, 1994; Castonguay et al., 2006; Constantino, Castonguay, & Schut, 2002; Horvath, 2000, 2005; Horvath et al., 2011;

Horvath & Greenberg, 1994; Norcross & Lambert, 2011; Samstag, 2006). Bordin's *working alliance* integrated the affective bond described by Greenson with a focus on the goals of therapy and the tasks needed to achieve these goals (Bordin, 1979). These concepts are encapsulated in the three operational components of Bordin's working alliance: bond, goal, and task (Bordin, 1979, 1994).

Collaboration (on goals and tasks) is the key factor that differentiates Bordin's conceptualization of the pan-theoretical working alliance from previous alliance constructs (Hatcher, Barends, Hansell, & Gutfreund, 1995; Horvath et al., 2011; Luborsky, 1976). Bordin's view of therapy as purposeful work affords a ready framework for situating the concept of client and therapist collaboration. By definition, 'working together' necessitates some form of collaboration. In Bordin's conception, collaboration functions in a reciprocal relationship with the working alliance: the quality of the alliance is a direct consequence of collaboration between client and therapist (Bordin, 1979, 1994; Horvath et al., 2011), and the central purpose of the alliance is to foster collaboration and consensus on the goals and tasks needed to facilitate fruitful work in therapy (Bordin, 1979; Hatcher et al., 1995). A key point here is that collaboration in the working alliance takes the specific form of consensus on the goals of therapy, and agreement on the tasks required to achieve these goals (Bordin, 1979; Horvath et al., 2011).

The working alliance has been the prevailing model of the therapeutic alliance in the past 30 years (Bordin, 1979, 1994; Horvath, et al., 2011; Norcross & Lambert, 2011), and the alliance itself is currently the most researched aspect of the therapeutic relationship (Norcross, 2011; Ribeiro, Ribeiro, Gonçalves, Horvath, & Stiles, 2012). The predominance of the working alliance has had two consequences of importance here: (1) the idea of collaboration as consensus and agreement, embedded in Bordin's

conception of the working alliance, has become the predominant conception of collaboration in therapy, and (2) the focus on the working alliance as a pan-theoretical construct has obscured forms of collaboration specific to different therapies.

The centrality of collaboration in the working alliance is reflected in research employing the alliance construct. The collaborative relationship remains the primary common component across conceptions of the alliance and alliance measures (Horvath & Bedi, 2002; Horvath et al., 2011; Karver, Handelsman, Fields, & Bickman, 2006; Lambert & Cattani, 2012; Leahy, 2008; Ribeiro et al., 2012; G. S. Tryon & Winograd, 2001, 2011). This point is illustrated in a meta-analysis of the relationship between the alliance and outcomes conducted by Horvath and colleagues (Horvath et al., 2011). This research reviewed 201 studies, of cognitive behaviour therapy, interpersonal psychotherapy, psychodynamic therapy, and counselling for substance issues, that employed 30 different alliance measures. Two thirds of the included studies used one of four main measures of the alliance. Despite considerable variation between measures, the central theme across the shared factor structure of the main four measures was 'the collaborative relationship' (Horvath et al., 2011).

There are several difficulties with aspects of alliance research that are relevant to the current study. Firstly, there is imprecision in terminology and a lack of consensus regarding definitions of the alliance construct, and its role in therapy (Fitzpatrick, Iwakabe, & Stalikas, 2005; Horvath, 2005, 2006, 2011; Horvath et al., 2011; Horvath & Luborsky, 1993; Meissner, 2006; Saketopoulou, 1999). Bordin himself did not clearly distinguish his conception of the working alliance from previous alliance constructs (Bordin, 1979; Horvath et al., 2011). The lack of consensus encompasses the relationship between the alliance and collaboration (Horvath, 2006; Meissner, 2006). In some cases, collaboration has been equated with

the alliance (Colson et al., 1988; Frieswyk et al., 1986), or considered the primary core of it (Hatcher & Gillaspy, 2006; Ribeiro et al., 2012; Summers & Barber, 2010), in others, the role of the alliance is to facilitate client engagement in therapy, thereby enabling a collaborative working together (Castonguay et al., 2006; Hatcher, 2010).

In addition to a lack of consensus regarding the definition of the alliance and its relationship with collaboration, there has been little theoretical work on collaboration apart from considering it as an enabling attribute of the working alliance (Hatcher & Barends, 2006; Horvath et al., 2011; G. S. Tryon & Winograd, 2011). This lack of theory has been managed in practice by defining collaboration in terms of the instruments devised to assess it (Bachelor, Laverdière, Gamache, & Bordeleau, 2007; Horvath, 2011; Horvath et al., 2011). This has contributed to a proliferation of measures in the literature (G. S. Tryon & Winograd, 2001, 2011). Although many of these measures have adequate individual reliability (Horvath et al., 2011), there has been little integration, possibly due to the lack of a clear basis of shared definitions and theory (Bachelor et al., 2007; Horvath et al., 2011).

Researchers have begun to critique the lack of theoretical work on collaboration and the alliance in the literature, and to call for greater theoretical articulation of these constructs (Bachelor et al., 2007; Horvath et al., 2011; Muran & Barber, 2010; Ribeiro et al., 2012; Wiseman et al., 2012). Recent examples of this include the grounding of a theory of collaboration in the assimilation model of therapeutic change, in which collaboration is conceived of as acting directly as a therapeutic change agent by enabling the internalization of adaptive self-narratives (Ribeiro et al., 2012); and the theory of collaborative empiricism in terms of Self-Determination Theory presented in the first chapter of this thesis (Tee & Kazantzis, 2011).

A third issue relates to a prevailing tendency in alliance research to equate the working alliance with its operational components of bond, goal, and task (Castonguay et al., 2006; Constantino et al., 2002; Horvath et al., 2011). Bordin viewed the bond between client and therapist as important for therapy because it directly supports the working alliance. The working alliance in turn allows client and therapist to collaborate on the tasks and goals of therapy. Thus, in Bordin's theory, bond has a direct impact on collaboration (Bordin, 1979). Bordin observed that different therapies required different levels or types of bond between client and therapist (Bordin, 1980, 1994). He contrasted CBT with psychoanalysis in this regard. The consequence of this is that different therapies require different forms of the working alliance (Bordin, 1979; Hatcher, 2010), and therefore different forms of collaboration. This insight is obscured in research in which the broader sense of the working alliance as a theory of therapy as purposeful work is subordinate to a focus on the operational components of the theory (Hatcher, 2010).

The reduction of the alliance to its operational components has consequences for research on collaboration because the alliance components of goal and task embed a particular conception of collaboration, as agreement on goals and consensus on tasks, into the alliance construct. This results in a conception of collaboration that is overly focused on consensus and agreement, rather than incorporating the broader sense of collaboration as sharing the purposeful work of therapy. The understanding of collaboration as focused on agreement on goals and tasks continues to reverberate in the literature. For example, a recent review focused on guiding practitioner collaboration in psychotherapy introduced collaboration as "essential, as the process of reaching goals requires the therapist and client to agree on the necessary steps to reach such goals." (Lambert & Cattani, 2012, p. 209). The review went on to mention

other aspects of collaboration, but remained focused primarily on collaboration as outcome monitoring and 'goal achievement', demonstrating the pervasive nature of the conception of collaboration as consensus and agreement on goals and tasks.

A further consequence of the narrowed focus on operational components is the loss of distinction between different forms of collaboration in different therapies. Bordin anticipated that 'working together' would take different forms in different therapies, and that consequently the alliance needed to facilitate working together would also take different forms in different therapies (Bordin, 1979, 1994; Horvath et al., 2011). In this sense, the pan-theoretical working alliance provides a general framework for comparing the alliance and collaboration across different therapies. If, however, the focus of the alliance is narrowed to its operational components, the different forms of alliance intrinsic to different forms of therapy may be reduced to consensus on goals and agreement on tasks (supported by an affective bond). Obscuring the difference between specific forms of the alliance in different therapies in turn obscures the different forms of collaboration embedded in the alliances of different therapies. Obscuring the form of collaboration specific to different therapies is likely to obscure important aspects both of collaboration and of the therapies themselves (Kazantzis & Kellis, 2012). Collaboration means something different in the context of Freudian psychoanalysis or Rogerian client centred therapy, for example, than it means in CBT. In the case of CBT, conceptualizing collaboration as consensus and agreement deemphasizes the active, creative role of the client in developing and assessing activities carried out in session in real-time. This latter form of collaboration is a key aspect of collaboration in collaborative empiricism.

Static vs. emergent alliance.

Another issue relates to the distinction between the alliance as a static property versus a fluid, emergent property changing in real time. The alliance has typically been measured as an attribute of a whole session, with the implicit assumption that it remains constant across a session, or in some cases across multiple sessions (Bachelor, 1991; Horvath et al., 2011; Safran & Muran, 2000). The most commonly used measures of the working alliance give an overall rating of the alliance for a whole session (Hatcher & Gillaspy, 2006; Horvath & Greenberg, 1989; Tichenor & Hill, 1989). It is clear, however, that the alliance is affected by client and therapist in-session emergent behaviour (Charles J. Gelso & Carter, 1994; A. Ryan, Safran, Doran, & Moran, 2012; Stiles et al., 2004; Stiles, Honos-Webb, & Surko, 1998), and is subject to ruptures and repair (Aspland, Llewelyn, Hardy, Barkham, & Stiles, 2008; Safran, Muran, & Eubanks-Carter, 2011; Safran, Muran, Samstag, & Stevens, 2001).

Collaboration has also been frequently measured as a property of a whole session (Lepper & Mergenthaler, 2007; J. E. Young & Beck, 1990); yet it seems likely that collaboration will be affected by emergent in-session behaviour, if for no other reason than it is embedded within the therapeutic alliance. There is also evidence that particular elements of collaboration may be more or less effective with different clients (Brotman, 2004; Hatcher, 2010). This suggests that collaboration may be sensitive to the emergent input of the client, and that research may usefully examine collaboration as it changes within a therapy session, rather than taking a single measure for a whole session.

Specific versus Common Factors

Although there is near universal agreement in the literature that attributes of the therapeutic relationship play a crucial role in successful therapy, there is debate

regarding the relative importance of ‘specific’ factors, understood as technical factors that are treatment or therapy modality specific, and ‘common’ factors, such as the therapeutic relationship, which are assumed to be shared across all forms of therapy (Castonguay, Goldfried, Wiser, Raue, & Hayes, 1996; DeRubeis et al., 2005; Imel & Wampold, 2008; Karver, Handelsman, Fields, & Bickman, 2005; Norcross, 2011). In their seminal work describing collaborative empiricism, Beck and colleagues designated attributes of the therapeutic relationship such as warmth, accurate empathy, and genuineness as ‘necessary but not sufficient’ to produce optimum therapeutic effect (A. T. Beck et al., 1979). Other views hold that the therapeutic relationship is itself the key agent of therapeutic change (Lambert & Barley, 2001). The specific and common factors debate is beyond the scope of this study; however, it is noted here that, using the definitions of ‘specific’ and ‘common’ factors above, collaborative empiricism functions as both a treatment specific intervention (e.g., empirical exploration of the client’s cognition and cognitive processes) and as a relational intervention (active collaboration on shared empirical work). Beck and colleagues statement that the relationship is not sufficient for change in CBT does not mean that the relationship is not an key aspect of CE (Leahy, 2008; Safran et al., 2011; Safran & Muran, 2000; Thwaites & Bennett-Levy, 2007; J. E. Young & Beck, 1980b; J. E. Young, Klosko, & Weishaar, 2006).

Responsiveness and mutuality.

Two further aspects of collaboration in the psychotherapy literature that are relevant to CE are *therapist responsiveness* and *mutuality* between client and therapist (Stiles et al., 1998). These concepts describe central elements of the actively shared work characteristic of collaboration in CE. Responsiveness refers to the extent to which one participant’s actions address the other’s previous communications, actions,

needs, or wishes (L. C. Miller & Berg, 1984). A responsive action addresses the previous communicated meaning, and elaborates appropriately (D. Davis & Perkowitz, 1979), uses appropriate non-verbal communication to signal listener cues (Derlega, Hendrick, Winstead, & Berg, 1992), and is appropriately timed (i.e., not delayed) so that it is perceived as meaningfully connected to the behaviour of the other person (Derlega et al., 1992; Piner, Berg, & Miller, 1986).

Responsiveness has also been discussed in the literature in terms of the therapist's responsiveness to the needs and capacities of the client and the therapy. Stiles and colleagues have discussed responsiveness as the degree to which the content and process of therapy emerge in real time throughout the therapy, rather than being planned from the outset, and such that the therapist does what is required to skilfully avail the desired outcome (Stiles et al., 1998). This means that the therapy is tailored for the individual, rather than inflexibly following a manualised treatment model. This sense of responsiveness captures an important aspect of CE as involving an emergent and fluid interaction between client and therapist, although in the case of CE the flow is ideally two way, and a product of the dyad, rather than solely the therapist.

Beck et al. (1979), in their seminal treatment manual for cognitive therapy of depression, explicitly discuss responsiveness as a core element of good cognitive therapy practice. In their guide for the 'selection of targets and techniques', Beck et al. recommend that therapists respond to client factors such as the degree, extent, and severity of the client's problems, as well as the client's 'style', coping strategies, and level of interpersonal sophistication. They note that finding a fit between the client's problems and appropriate interventions in therapy will require 'trial and error' and that several potential interventions may be trailed before finding one that best fits the

client (A. T. Beck et al., 1979, p. 169; cited by Stiles et al. 1998, p. 446). This process, of trialling interventions in order to determine how well they fit an individual client's circumstances and clinical needs, is an example of responsiveness at the treatment level, manifest as an empirical process. Client and therapist share the work of planning and developing interventions and then trialling them to determine the utility of an intervention relative to the client's goals for therapy.

Mutuality between client and therapist is a key factor that distinguishes Bordin's concept of the working alliance (Bordin, 1979) from other conceptions of the therapeutic alliance, such as Roger's (1957, 1961) concept of therapist provided conditions for therapeutic growth, and Strong's (1968) interpersonal influence conceptualization (Horvath & Greenberg, 1989). Mutuality can also distinguish the more didactic approach to the alliance in Ellis' (1980) application of rational disputation as a means for correcting 'irrational' thinking. For Bordin, mutuality incorporates the degree of agreement, concordance, shared purpose, and co-operation underlying each of the relationship dimensions of bond, goal, and task (Bordin, 1979, 1994), rather than the degree to which the actual work of designing and carrying out these tasks is shared. This concept of *sharing the work* is a distinguishing feature of collaboration in CE and will be elaborated below.

Empirical Work

Collaboration as an empirically supported relationship attribute.

Previous research has demonstrated a small but reliable association between collaboration and improved outcomes, irrespective of therapeutic modality (Colson et al., 1988; Frieswyk et al., 1986; Hatcher & Barends, 1996; Horvath & Bedi, 2002; McGuire-Snieckus, McCabe, Catty, Hansson, & Priebe, 2007; G. S. Tryon & Winograd, 2001, 2011). These results are supported by meta-analyses showing a

relationship between outcomes and the therapeutic alliance, of which collaboration is a key part (Horvath et al., 2011; Karver et al., 2006; Martin et al., 2000; Norcross, 2002).

The APA Task Force on Empirically Supported Therapy Relationships.

As with the theoretical work discussed above, empirical research on collaboration in psychotherapy has generally approached the construct as an attribute of the therapeutic relationship between client and therapist (Kazantzis & Kellis, 2012; Norcross & Wampold, 2011; Ribeiro et al., 2012; G. S. Tryon & Winograd, 2002, 2011). Relationship attributes are aspects of the therapeutic relationship that support and promote the client's engagement in the therapy process, thereby providing a foundation for productive therapeutic work (Lambert, 2004; Norcross, 2011).

In a landmark review, the APA Task Force on Empirically Supported Therapy Relationships (Norcross, 2001, 2002) commissioned a series of meta-analyses, guided by an expert panel, to identify effective attributes of the therapeutic relationship, and to specify the level of empirical support for these attributes. A second Task Force, convened in 2011, revised and extended the work of the first (Norcross, 2011; Norcross & Lambert, 2011). The second Task Force commissioned over 20 meta-analyses and examined 12 elements of the therapeutic relationship. They categorized the level of empirical support for the relationship attributes they investigated into three levels: 'demonstrably effective', 'probably effective', and 'promising but insufficient research to judge'. Demonstrably effective were the alliance, cohesion, and collecting client feedback; probably effective were goal consensus, collaboration, and positive regard; and promising but insufficient research to judge were congruence/genuineness, repairing alliance ruptures, and managing countertransference (Norcross & Wampold, 2011). Three of these attributes share

important features with collaboration in CE: collecting client feedback (demonstrably effective), goal consensus (probably effective), and collaboration (probably effective). Empirical work on these attributes will be discussed in turn.

Collecting client feedback.

In examining 'collecting client feedback', the Task Force conducted a meta-analysis of studies that had employed either of two systems for assessing feedback: the Partners for Change Outcome Management System (PCMOS) and the Outcome Questionnaire (OQ) of the Psychotherapy Quality Management System (Lambert & Shimokawa, 2011). The Task Force reported a combined weighted random effect size for the PCMOS of $r = .23$, 95% CI [.15, .31], $p < .001$, $k = 3$, $n = 558$, and for the OQ system of $r = .25$, 95% CI [.15, .34], $p < .001$, $k = 4$, $n = 454$. They concluded that assessing clients' progress via feedback was an important and empirically validated method for overcoming therapists' inability to detect client worsening in therapy. The authors made the point, however, that the association between actions and outcomes in therapy may not be neatly temporally related, making evaluation of the feedback collected difficult.

The results of this meta-analysis demonstrate empirical support for 'collecting client feedback'. It is important to note, however, that the studies reviewed by the Task Force operationalized 'collecting client feedback' as the therapist collecting and using feedback to revise their own input into the therapeutic process, rather than requesting or sharing feedback on progress with the client. Also, the measures used to operationalize 'collecting client feedback' were symptom measures (the PCMOS and the OQ). Operationalizing feedback as gathering data on outcomes is a restriction of feedback to a single aspect of that described in the conceptualization as 'collecting client feedback'. It is also a far narrower conceptualization of feedback compared

with feedback as understood in CE. Feedback in CE is a two-way, dyadic utilization of the client's cognition and cognitive processes as data for exploration and employment in interventions in real time. Feedback in CE may involve, for example, ascertaining whether the client and therapist share an understanding, whether they are on the same track, or whether the client feels the current work is helpful. Another key aim of feedback in CE is to fine-tune the collaboration between client and therapist (Kuyken et al., 2009). Feedback may also demonstrate empathy and understanding from the therapist, facilitating collaboration and trust (J. S. Beck, 2011). Seeking feedback in these broader senses may be important for outcomes. Recent research has found that it is the client's rating of the relationship in therapy that relates to outcomes, rather than the therapist's rating of either the relationship or the bond in the working alliance (Lo Coco, Gullo, Prestano, & Gelso, 2011). This suggests that it may be important for therapy outcomes to elicit feedback from the client not only regarding their symptoms but also their judgments and attributions.

Goal consensus.

Goal consensus is the agreement about treatment goals and the steps by which the client and therapist will reach these goals (G. S. Tryon & Winograd, 2011). As such, the construct of goal consensus closely reflects the central role of agreement on tasks and goals in Bordin's conception of the working alliance (Bordin, 1979, 1994; G. S. Tryon & Winograd, 2011).

Research has demonstrated a relationship between agreement on goals and therapeutic outcomes. Client-rated agreement with the therapist and client-rated experience of goal consensus at session two has been shown to predict reduction in symptoms on the Symptom Check List (SCL; Derogatis, 1974), including up to six months later (Dormaar, Dijkman, & de Vries, 1989). Therapist-client discussion of

goals and expression of thoughts and feelings has been associated with observer ratings of 'goodness' of therapy sessions (Hoyt, Xenakis, Marmar, & Horowitz, 1983). Simply specifying goals is enough to increase clients' ratings of therapy helpfulness (Goldstein, Cohen, Lewis, & Struening, 1988).

The second APA Task Force (Norcross & Lambert, 2011) conducted a meta-analysis of 15 studies published from 2000 to 2009 relating goal consensus to therapy outcome, with a total sample size of 1,302 (G. S. Tryon & Winograd, 2011). The results showed a mean correlation between goal consensus and therapy outcomes of $r = .34$, 95% CI [.23, .45], by consensus a medium sized effect in psychotherapy research (J. Cohen, 1992). The second Task Force used an expanded definition of goal consensus, compared with that adopted by the original Task Force, to include how well goals are discussed and specified, and the client's understanding and commitment to goals. Despite this, the construct remained largely centred on consensus on goals, agreement on the origin of the 'patient's problem', agreement on responsibility for the problem and its solution, and commitment on the part of the client to working towards the agreed goals (G. S. Tryon & Winograd, 2002, 2011). This is appropriate for a construct focused on 'consensus'; however, as discussed above, collaboration in CE goes beyond agreement or consensus to incorporate a deeper sharing of the design, employment, evaluation, and review of therapeutic tasks. In addition, in the work discussed above, 'goals' is focused on therapy goals, rather than the more detailed empirical review, analysis, and discussion of goals on multiple levels in CE. Nevertheless, agreement on goals and agreement on the target and rationale for goals are important aspects of collaboration in CE, and the research discussed above provides support for the efficacy of these aspects.

Collaboration.

The APA Task Force identified 'collaboration' as a specific relationship attribute for investigation. Tryon and Winograd (2011) conducted a meta-analysis of 19 studies published between 2000 – 2009. These studies employed a group design of individual adult psychotherapy and each included at least one measure of collaboration and one outcome measure. The authors viewed collaboration from the perspective of Bordin's working alliance (Bordin, 1979, 1994). Collaboration was defined in the review as "the active process of working together to fulfil treatment goals" (G. S. Tryon & Winograd, 2011, p. 50).

In the studies reviewed, collaboration was operationalized chiefly in terms of the agreement, co-operation, or compliance of the client with the agreed upon goals of therapy, and the tasks needed to reach these goals. Measures of collaboration included ratings of involvement in the client role, mutual involvement in the helping relationship, and client co-operation (G. S. Tryon & Winograd, 2011). Homework was also used as an 'indicator' of collaboration. This was operationalized as homework completion or compliance, the quality of homework, and the client's rating of homework. The results of the meta-analysis showed a mean correlation between collaboration and therapy outcomes of $r = .33$, 95% CI [.25, .42], a medium sized effect (J. Cohen, 1992). The authors concluded that collaboration appears to considerably enhance psychotherapy outcomes.

This work provides important meta-analytic confirmation of a relationship between collaboration and outcomes in therapy. Again, however, collaboration was conceived in terms of Bordin's working alliance, and operationalized largely as the agreement, co-operation, or compliance of the client with agreed upon tasks and goals of therapy. As discussed above, although this captures important aspects of

collaboration, it does not explicitly capture collaboration as sharing the work of designing, implementing, and evaluating the tasks and goals of therapy. Functioning 'as a team', when operationalized in terms of agreement, co-operation, or compliance, does not discriminate how much of the actual work is done by client or therapist.

This problem is illustrated in a clinical vignette provided by Tryon and Winograd (2011) to demonstrate goal consensus and collaboration at work. In this vignette, the client reports worry and social avoidance related to fears of being judged over a recent, first episode of psychosis. After inquiring about the problem, the therapist gives an interpretation in terms of a cognitive formulation and the client agrees. The authors label this single act of agreement "goal consensus: congruence on patient problem" (G. S. Tryon & Winograd, 2011, p. 53). This is a good example of high agreement but low shared work. (It would rate no higher than 3 out of 5 for collaboration on the Collaborative Empiricism Scale). The vignette goes on to describe excellent collaboration involving genuinely shared therapeutic work, but the authors conclude that the client would be likely to endorse items on the Working Alliance Inventory related to consensus, such as, 'we agree about what is important for me to work on'. The point here is that 'agreeing about what is important' does not discriminate between the client simply agreeing with the therapist's interpretation (early in the vignette) or the client herself developing hypotheses about her own thinking and behaviour, and ways to explore these (later in the vignette). Collaboration operationalized as consensus, such as when measured using the WAI, does not discriminate the shared work characteristic of collaboration in CE.

Recommendations of the Task Force on Empirically Supported Therapy Relationships

As an element of therapy that has been consistently described over 40 years as a core aspect of the therapeutic relationship in CBT, collaborative empiricism is conspicuous in its absence from the Task Force's review. This is understandable because the Task Force based their recommendations on the research data available at the time. Despite the absence of specific mention of CE, the Task Force made several concluding remarks regarding their review of the empirical literature that speak directly to the potential importance of collaborative empiricism for treatment and research (Norcross & Wampold, 2011). The Task Force noted that therapists 'who assume or intuit' their clients' perceptions of the success of the relationship or the therapy are not infrequently wrong, and they called for ongoing collection and use of empirical data, in the form of feedback, to enhance the alliance and prevent ruptures in therapy (Lambert & Shimokawa, 2011). The Task Force also highlighted the central importance of the client's perspective and noted that "privileging the client's experiences is central" (Norcross & Wampold, 2011, p. 428). Finally, they encouraged practitioners to seek more demonstratively effective ways of tailoring therapy for the needs of the client (Norcross & Wampold, 2011). As discussed in Chapter 1, the twin processes of collaboration and empiricism in CE can play an important role in tailoring the case conceptualization to the idiographic circumstances of the client.

Other Aspects of Collaboration in the Literature

In addition to feedback and goal consensus, previous research has found support for the efficacy of other aspects of collaboration in psychotherapy. These include topic determination, shared decision making, and collaborative involvement.

Elements of these constructs relate directly to the active shared work inherent in collaboration in CE.

Topic determination.

There is evidence that *topic determination*, the proportion of topic initiations subsequently followed by the other participant in therapy, predicts the continuation of therapy past session three (Tracey, 1986). Similarly, higher *response congruence*, in which the therapist's verbal response directly addresses the subject of the client's immediately preceding statement, is associated with higher likelihood of the client returning for therapy after intake (Duehn & Proctor, 1977).

Notably, not all studies have found a relationship between aspects of topic determination and outcome. Beyebach and Carranza (1997) failed to replicate the relationship between topic determination and engagement reported by Tracey (1986). However, non-engaged clients in the Beyebach and Carranza study used more domineering language with therapists, interrupted more often, and engaged in more conflict. It is possible that these behaviours disrupted collaboration, resulting in lower continuation of therapy (G. S. Tryon & Winograd, 2002).

Shared decision making.

Another construct related to collaboration in CE is *shared decision making*. Shared decision making is an interactive process in which both client and therapist take steps to share decisions about the form and process of therapy, share information about options, and come to some form of consensus regarding treatment processes (Schauer, Everett, del Vecchio, & Anderson, 2007). Shared decision making is explicit in discussions of CE in the seminal CBT literature (A. T. Beck, 1967; A. T. Beck et al., 1979; J. S. Beck, 1995, 2011). There is support in the literature for the wide-ranging utility of sharing decisions in therapy. Clients are more likely to return

for therapy after an intake interview in which the therapist shares the problem formulation and negotiates mutual treatment goals, compared to an interview where neither is done (Tracy, 1977). Sharing decisions with the client regarding the treatment plan, and providing a clear rationale and explanation for the plan, is associated with clients' satisfaction with that session (Eisenthal, Koopman, & Lazare, 1983). Many clients like sharing decisions in therapy (J. Hamann, Cohen, Leucht, Busch, & Kissling, 2005), which is itself likely to increase engagement in the therapy process. The involvement of clients who are by disposition less actively engaged in sharing decisions can be increased by providing communication training decision aids (Adams & Drake, 2006), and sharing decisions can reduce conflicts regarding medication adherence (Deegan & Drake, 2006). From the point of view of CE, sharing decisions is also likely to be helpful for therapy because sharing decisions is a direct method of gathering empirical data on how the client thinks and feels, and how the client thinks and feels about therapy. This in turn increases the data derived from client feedback, and fosters evidence-based decision making.

Collaborative involvement.

Collaboration has also been conceptualized as c*ollaborative involvement*. Collaborative involvement is the mutual involvement of patient and therapist in a helping relationship (G. S. Tryon & Winograd, 2002). The term 'mutual involvement' suggests that this construct has a meaning closer to 'sharing the work' in CE; however, studies using the construct have operationalized collaborative involvement as the compliance or cooperation of the client to the strategies of the therapist. For example, Schmidt and Woolaway-Bickel (2000) operationalized collaborative involvement as homework completion. They found that therapists' ratings of compliance with homework predicted outcomes, but clients' ratings of compliance

did not. This suggests a relationship between clients' compliance and outcomes, but 'mutual involvement' measured as homework completion does not encompass the genuine sharing of the work described in the literature on CE. It is not clear, for example, whether clients in the above studies co-developed homework tasks or were simply compliant with those set by the therapist.

In an interesting study, O'Malley, Suh, and Stupp (1983) found that client involvement at session 3 predicted outcome, whereas involvement at sessions 1 and 2 did not. They concluded that client involvement may not be simply a quality that the client brings to therapy, but may be facilitated by qualities of the therapist. Windholz and Silberschatz (1988) found that therapist ratings of client involvement were related to outcomes, but client and observer ratings were not. Further research is needed to determine whether collaborative involvement is a function of the interaction between client and therapist, rather than a quality of one or the other in isolation. If true, the active sharing of the work of therapy inherent in CE could be expected to increase client involvement.

Conclusion

Collaboration has been predominantly considered in the literature as an attribute of the therapeutic alliance. In general, there has been little theoretical underpinning of this work. Research has managed this lack of theory by defining collaboration and the alliance in terms of the instruments used to measure them. This has led to a proliferation of measures with little theoretical integration.

Despite the volume of research employing the alliance construct, important difficulties remain. There is imprecision in language and a lack of consensus on definitions of collaboration and the alliance. Frequently, the alliance has been identified with its operational components of bond, goal, and task. This narrowing of

focus has led to a parallel narrowing of collaboration to consensus and agreement, as well as obscuring the differences between types of collaboration in different therapies. Research is needed that conceptualizes collaboration more broadly than consensus and agreement; operationalizes the construct in more detail than bond, goal, and task; and measures collaboration as an emergent property of the therapeutic dialogue, rather than an attribute of the therapy sessions as a whole.

The empirical work reviewed above demonstrates support for a relationship between aspects of collaboration and therapeutic outcomes. This suggests that collaboration in CE, which incorporates and extends these aspects of collaboration, may also relate to outcomes. Nevertheless, collaboration in CE goes beyond the meaning of collaboration in these studies. In addition, previous work has largely focused on collaboration as a client or therapist factor, rather than a property of the dyad, and as a property of the whole session, rather than a fluid construct. Although existing research provides qualified support for a relationship between collaboration in CE and therapeutic outcomes, research focusing specifically on CE is needed. For this, a measure is needed that incorporates the aspects of collaboration discussed above, and widens the operationalization of collaboration to include the meaning of collaboration in CE as ‘sharing the work’. These aims are taken up in the development of the Collaborative Empiricism Scale in the present research.

Chapter 3: Empiricism

Aims and Outline

This chapter reviews empiricism in CBT. First, a definition of empiricism in CBT is derived from the seminal literature of A. T. Beck and colleagues. Theoretical work relevant to empiricism is then discussed, followed by a review of empirical research. The process of empiricism in a typical CBT session is then discussed.

Levels of Empiricism in Cognitive Behaviour Therapy

Empiricism is a fundamental guiding principle of CBT (A. T. Beck et al., 1979; J. S. Beck, 2011; Keith S. Dobson, 2012; Friedberg & Brelsford, 2011; Kazantzis, Cronin, Dattilio, & Dobson, in press), which operates at multiple levels within the therapy (Kazantzis et al., 2013; Kazantzis, Tee, Dattilio, & Dobson, in press; Kuyken et al., 2008). At the first level is empirical support for CBT, through the scientific process (e.g., K. S. Dobson et al., 2008; Haaga, Dyck, & Ernst, 1991; Haubert & Dobson, 2007), the scientist-practitioner model (Davison, 1998), and evidence-based practice (D. J. G. Dobson & Dobson, 2009; Spring, 2007). A second level is empirical support for specific interventions (Kazantzis et al., 2013; O'Donohue & Fisher, 2008). And a third level of empiricism is the responsive adaptation of interventions and therapy process to the unique historical, cultural, and personal circumstances of clients (Kazantzis, Deane, Ronan, & L'Abate, 2005; Kuyken et al., 2009; O'Brien, 2010; Persons, 2008).

A Definition of Empiricism in Cognitive Behaviour Therapy

In their seminal work, Beck and colleagues described collaborative empiricism as a process in which the client and therapist explore the client's cognitions by weighing empirical evidence rather than proceeding by rhetorical argument or debate (A. T. Beck et al., 1979). The weighing of empirical evidence is based on, concerned

with, or verifiable by observation or experience rather than on logical or a priori theoretical grounds (A. T. Beck et al., 1985; A. T. Beck et al., 1979). A key aim of empiricism in CBT is, therefore, the identification, observation, and exploration of the client's cognitions and cognitive processes in an objective manner (A. T. Beck et al., 1985).

The objective exploration of cognitions and cognitive processes is necessary because, in general, the therapist does not know in advance whether a specific cognition or cognitive process is helpful, functional, or valid (J. S. Beck, 2011). The client also frequently does not know in advance whether their cognitions or cognitive processes are helpful or correct. The inability of the client to intuit the helpfulness or validity of their own cognitions may be in part why they are in therapy. And conversely, a key part of their therapy may be learning to gauge the helpfulness and validity of cognitions based on more carefully scrutinized data, drawn from their experience related to these cognitions (J. S. Beck, 2005).

The term 'objective' also distinguishes evidence from interpretation. For example, a client who is anxious because 'their boss is angry with them' may support this belief by reporting that their boss often glares or scowls during conversations, but this interpretation of their bosses' behaviour may be incorrect. The boss may be busy or stressed, or concerned about another matter. The distinction between evidence and interpretation may be difficult for some clients to grasp, or may be culturally more or less salient or valued (Wong, in press).

Discussions of empiricism in the CBT literature also note two other key aspects of the construct. First, empiricism should ideally be employed at multiple levels throughout the therapy, including case conceptualization, treatment planning, and carrying out interventions, such that the client and therapist work together to

"examine the resulting pool of data, develop a conceptualization and plan intervention strategies" (A. T. Beck et al., 1979, p. 55). Secondly, the therapist's role is fundamentally to "guide the client to collect information in which the unhelpful cognitions are embedded" (A. T. Beck et al., 1985, p. 182). These components of empiricism have been embraced in later discussions of empiricism in CBT (J. S. Beck, 1995, 2011; Dattilio, 2010; Kuyken et al., 2008, 2009).

Drawing on these discussions in the literature, empiricism in CBT is defined here as a core feature employed across multiple levels of the therapy, focused on the client exploring their world (including cognitions, cognitive processes, behaviour, emotions, and problem situations). The client's cognitions are treated as hypotheses, to be explored by gathering and evaluating data from the client's experience, and the therapist's primary role is to guide and foster the process of the client's empirical exploration.

Theoretical Work

Empiricism and cognitive change.

Although there has been little theoretical work focused specifically on empiricism in CBT, theoretical work on change processes is relevant for an understanding of the construct. A central tenet of CBT that relates to the exploration and evaluation of cognitions is the cognitive mediation hypothesis. The cognitive mediation hypothesis states that change in clients' cognitions mediates therapeutic outcomes (A. T. Beck, 1970; DeRubeis et al., 2001). In the context of CBT for depression, the focus of the present study, the cognitive theory of depression developed by Beck and colleagues states that depressed people have latent depressogenic schemas, which when activated by stressful events lead to patterns of negative automatic thoughts and negative cognitive processes (such as negative

perceptual biases and discounting of positive evidence) (A. T. Beck, 1967, 1976; A. T. Beck et al., 1979). These negative biases result in a reinforcing spiral of increasingly negative affect, behaviour, and cognition, leading eventually to a depressive disorder (A. T. Beck et al., 1979; J. S. Beck, 2011). Stated broadly, the cognitive mediation hypothesis asserts that change in key negative cognitions leads to improvement in depressive symptoms and associated behaviours. There is considerable support for the cognitive mediation hypothesis (Garratt, Ingram, Rand, & Sawalani, 2007; Haaga, 2007; Haaga et al., 1991; Haubert & Dobson, 2007; Ingram, Miranda, & Segal, 1998; Reilly, Ciesla, Felton, Weitlauf, & Anderson, 2012; Romero, Agnew, & Insko, 1996).

Several models have been proposed to explain the mechanism of cognitive mediation (Garratt et al., 2007). Depressive schemas may be modified profoundly (the accommodation model) (Steven D. Hollon, Evans, & DeRubeis, 1990), or may remain unchanged but be deactivated progressively during treatment (the activation – deactivation model) (Ingram & Hollon, 1986), or remain unchanged but compensatory schemas developed that nullify their negative effects (the compensatory model) (Steven D. Hollon et al., 1990).

In these models, depressogenic schemas are either changed at their roots, deactivated, or neutralized by a compensatory schema. The question of relevance for an understanding of empiricism is how these changes in schemas take place. In each case, theory proposes that changes in the client's thinking result from the consideration of empirically derived data. That is, it is the thoughtful interaction of the client with their own experience that brings about change in schemas. Existing beliefs are used to generate hypotheses, based on the client's experience, combined with input from the therapist, who is guided by clinical theory and research (A. T.

Beck et al., 1979; Garratt et al., 2007). The client and therapist then test these hypotheses in an empirical process of comparing the predicted consequences of hypotheses with the observed consequences from the client's experience. The client may revise or amend the belief underpinning the hypothesis, depending on the outcome of hypothesis testing. Thus, the change in cognition central to the cognitive mediation hypothesis is itself mediated by a fundamentally empirical process of deriving the necessary consequences of current thinking as predictions and comparing this with empirically derived data from the client's experience. Although not explicitly articulated in the model, empiricism underpins the cognitive mediation hypothesis as its engine, placing empiricism at the heart of change in CBT.

Empiricism in Case Conceptualization.

Empiricism has been hypothesized to play a key role in case conceptualization (Kuyken et al., 2009). Case conceptualization is important because it is a primary process for guiding the selection and development of interventions to achieve therapeutic goals, for integrating theory and clinical practice, and for assessing the progress of therapy (Charlesworth & Greenfield, 2004; Hutton & Morrison, in press; Needleman, 1999; Persons, 2008). Case conceptualization typically begins with a description of the client's problems and resources. Over time, this descriptive level is integrated with explanatory theory and consideration of relevant developmental and cultural history (Butler et al., 2008; Persons, 2008). Empiricism has been theorized to function throughout this process, as a method of checking the utility and accuracy of the conceptualization, for selecting between alternative explanations or interventions (Kuyken et al., 2009), and to improve the case conceptualization by counterbalancing mistakes in decision-making (Kuyken et al., 2009). These mistakes stem from characteristic errors of reasoning, cognitive biases, or the over-application of

heuristics that distort our judgments (Kahneman, 2003). Although the hypothesized role of empiricism in case conceptualization has been articulated clearly in the literature, there has been no empirical evaluation to date. An aim of the present study is to facilitate such evaluation by developing a measure of collaborative empiricism in CBT.

Empiricism in collaborative empiricism.

Empiricism is also conceptualized to play a key role in the theoretical model of CE proposed in Chapter 1 of this thesis (Tee & Kazantzis, 2011). In that theory, based on Self-Determination Theory (Deci & Ryan, 2008; R. M. Ryan & Deci, 2008), the client's intrinsic motivation to change is enhanced by the employment of the client's *intrinsic data* as the basis of exploring the client's cognitions and cognitive processes. Intrinsic data are data derived by the client from the client's experience, and evaluated by the client using their experience. It is the high personal relevance and experiential immediacy of intrinsic data that is theorized to leverage empiricism and thereby enhance intrinsic motivation for change.

Empirical Work on Empiricism

There is no research focused specifically on empiricism in psychotherapy known at this writing. However, an area of empirical research of relevance to empiricism is the literature on skill acquisition and use in CBT. The use of concrete skills related to specific, problem-focused aspects of CBT has been shown to predict reduction in depression severity at termination, whereas abstract techniques such as psychoeducation regarding the cognitive model, did not (DeRubeis & Feeley, 1990; Feeley, DeRubeis, & Gelfand, 1999). Concrete skills in these studies included 'asking for specific examples of beliefs', 'reporting cognitions verbatim', 'examining evidence concerning beliefs', and 'assigning or reviewing self-monitoring'. Abstract

skills included ‘relation between thoughts and feelings’, ‘the cognitive therapy rationale’, ‘exploring the personal meaning of thoughts’, and ‘encouraging independence’. This research is of interest in the present context because the concrete skills described are likely to be high in empirical content or focus, and the abstract skills low in empiricism. The association reported between concrete skills, high in empirical content, and improved outcomes suggests a role for empiricism in mediating the relationship between skills and outcome. Some support for this suggestion comes from research showing that clients’ ability to learn and implement compensatory CBT skills in session (such as ‘generating an alternative explanation of a belief’ and ‘planning to test a belief or idea’) is related to depression severity at termination (Jacques P. Barber & DeRubeis, 1992, 2001), and at 12 months post-therapy (Strunk, DeRubeis, Chiu, & Alvarez, 2007). Skills such as planning a test of an idea or generating an alternative explanation are again likely to be highly empirical because they are explicitly structured as empirical tests of an idea (or alternative idea) taken as a hypothesis.

Previous Measures of Empiricism

There is no specific measure of empiricism known in the literature at this writing. As discussed above, it seems reasonable to suppose that concrete CBT skills would have a high loading on empiricism, and therefore that measures of these skills may tap into empiricism as a construct (e.g., measures in Jacques P. Barber & DeRubeis, 1992; DeRubeis & Feeley, 1990). It is difficult, however, to see how to reliably distinguish empiricism as a separate construct on these measures. One measure that, for a short time, had a specific item measuring empiricism is the Cognitive Therapy Rating Scale (CTRS; J. E. Young & Beck, 1980a). In reporting the psychometric properties of the unpublished first version of the CTRS, Vallis and

colleagues listed Item 7 as ‘Empiricism’ (Vallis, Shaw, & Dobson, 1986). The ‘Empiricism’ item had the highest inter-rater reliability of the scale items, measured using intraclass coefficients (ICCs; Shrout & Fleiss, 1979), of .59, from a range of .27 – .59. The item was replaced on subsequent versions of the CTRS by an item titled ‘Guided Discovery’. The Guided Discovery item measures guided discovery using a 4-point Likert scale, ranging from ‘primarily debate or persuasion’ at the low end to ‘examining evidence, considering alternatives, weighing advantages and disadvantages, rather than through debate’ at the high end. A disadvantage with the structure of this item as a model for an item measuring empiricism is that it provides a single value for the whole session. Rating empiricism once per session is contrary to the understanding of empiricism described in the literature reviewed above, in which empiricism functions at multiple levels and occurs in multiple behaviours throughout a session. To capture empiricism in multiple contexts, a measure of empiricism would need to be able to freely rate empiricism as it occurs in the current focus of the session.

The Process of Empiricism Within a CBT Session

In the theory, research, and seminal literature reviewed above, empiricism in CBT is understood broadly as an empirical research process that is employed at multiple conceptual levels within the therapy. Empiricism in a session is focused on and expressed in key therapy processes. Key processes identified in the literature above that are a focus of empiricism include (a) identifying and selecting targets for discussion and intervention, or selecting interventions to employ; (b) rating emotion and the degree of belief associated with cognitions and cognitive processes; (c) general exploration of the client’s phenomenal world via ‘weighing empirical evidence’; (d) developing alternative explanations for beliefs; (e) designing and

conducting experimental tests of cognitions and cognitive processes; and (f) evaluating the results of these experimental tests.

These empirical processes can occur in any of the structural components of a CBT session, such as agenda setting, homework review, and homework planning. Figure 1 presents an example of the typical occurrence of empirical processes in a CBT session and their relation to the structural components of the session.

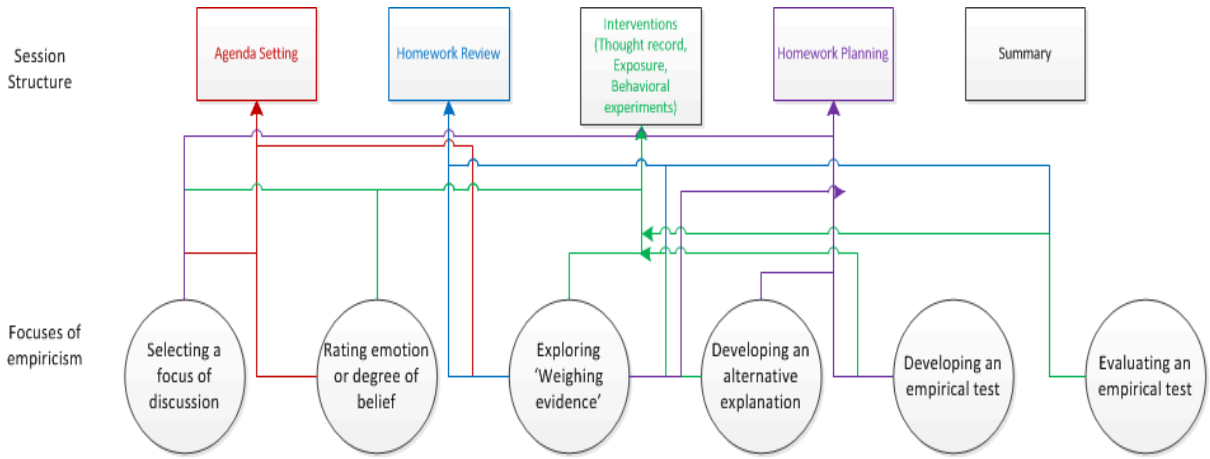


Figure 1. Empirical processes typically occurring in a CBT session, and their relation to session structure.

The processes described in Figure 1 that are a focus of empiricism occur throughout a CBT session. In a typical session, the client and therapist may begin by selecting a problem situation to discuss (from the agenda) because it is distressing for the client (client's data used to select a focus for discussion). In discussing this problem, there may be some rating of emotion associated with various thoughts that are characteristic of or central to the problem situation. The rating of emotion produces new data, which may result in thought record work (client's data used to select an intervention) on a particularly distressing (client's data) and frequent (client's data) thought. The thought record work might lead to the client and therapist developing a test of the thought (an empirical test, taking the thought as a hypothesis), which itself has the purpose of producing data, and then evaluating that data produces

more data, and so on… These processes need not occur in any particular order. Empiricism moves from 'a little to a lot' through a process of questioning, clarifying, summarizing, exploring, and testing, and sits on a foundation of collaboration. The six characteristic processes described in Figure 1, that are a focus of empiricism in a session, will be operationalized as *CE Events* in the Collaborative Empiricism Scale, described in the next chapter.

Conclusion

This chapter defined empiricism in CBT from the seminal literature, as a core feature employed across multiple levels of the therapy. Three levels of empiricism were discussed: the level of the therapy, the specific intervention, and the individual client. Theoretical work related to empiricism was then reviewed, including the role of empiricism in case conceptualization and as a mediator of cognitive change. Previous research relevant to empiricism was discussed, including the role of empiricism in the acquisition and use of therapy skills, and the measurement of empiricism on the CTRS. Lastly, the process of empiricism within a CBT session was described from the literature reviewed. Empiricism in a CBT session is expressed in key therapy processes, such as selecting a topic to discuss, rating emotion, and exploring an issue by weighing evidence. These processes typically occur multiple times in a session and in multiple contexts, for example agenda setting and homework review. Six characteristic empirical processes were identified, which will underpin the operationalization of empiricism in the Collaborative Empiricism Scale.

Chapter 4: Development of the Collaborative Empiricism Scale

Aims and Outline

This chapter describes the development of the Collaborative Empiricism Scale (CES; Tee, Kazantzis, & Stukas, 2012). First, the work of previous chapters is brought together in an overview of the process of CE in a typical CBT session. Overviews are then given of the scale itself, and the scale development process. Global aspects of the scale are then described. This is followed by a discussion of the development of the collaboration and empiricism items and the common rating scheme underpinning the empirical items.

Overview of the Process of Collaborative Empiricism in a CBT Session

Collaborative empiricism is a basic strategy or form of interaction that underlies the relationship between client and therapist in CBT (A. T. Beck et al., 1979; J. S. Beck, 1995, 2011). As such, CE is pervasive throughout a session and throughout the interaction between client and therapist. Collaborative empiricism is perhaps unique as a construct in that everything that happens in a session can be more or less empirical and more or less collaborative, and this extends to between sessions homework and the client's responses to their problems and challenges outside of therapy in everyday life. Accordingly, CE is not localized to specific structural components of a session, such as agenda setting or reviewing homework, nor is it localized to specific interventions. The fundamental elements of CE, the actively shared work of identifying and exploring problems, and the design, implementation, and evaluation of interventions to improve problems, are centred instead on the particular action, activity, or discussion that the client and therapist are engaged in at any particular time (A. T. Beck, 1970; A. T. Beck et al., 1985; J. S. Beck, 2005,

2011). When done well, CE flows throughout the session, flavouring and framing each interaction from start to end.

Collaborative empiricism also functions on multiple levels in a session. Data is gathered from the client's experience and evaluated using the client's experience. The gathered data is used as the basis of therapeutic interventions, but also in case conceptualization, treatment planning, and progress review (A. T. Beck et al., 1979; Dattilio & Hanna, 2012; Kazantzis, Beck, Dattilio, Dobson, & Rapee, in press; Kuyken et al., 2009; Overholser, 2011; Persons, 2008). Although centred on the momentary flow of the session, the process of CE takes time. The careful work of gathering data and sifting through evidence to distinguish it from interpretation may slowly evolve over one or many sessions, as the client develops familiarity and skills with the process.

In a typical CBT session, the client and therapist may start by setting a session agenda. This might be directed by the therapist, 'Let's start with the homework'; or might be more collaborative, 'What would you like to start with?'; and might be more empirical, 'Which of these things has caused you the greatest distress this week?', or, 'How should we decide which to discuss first?'; or less empirical, 'Ok, let's start with that' (the client's first suggestion, without further reference to the client's experience). Having chosen a first topic, the client and therapist might then explore this, again with more or less collaboration and empiricism, until the therapist suggests (or more collaboratively, asks the client whether) they should do a thought record. The thought record might involve rating the degree of emotion associated with several thoughts, in order to choose a clinically important 'hot' thought to examine. Again, rating the degree of emotion can be more or less empirical and more or less collaborative. For example, more empirical might involve discussing specific anchor points on a

subjective scale of distress, grounded in the client's experiences of similar past situations. Less empirical might be, 'How anxious? Give me a number from one to ten.' The client and therapist might then move on to explore the thought and develop an alternative interpretation of it – again, each with a higher or lower degree of empiricism and collaboration.

The process of CE in a session is centred on the specific focus of activity at any particular time. This focus of activity changes with the flow of the session. The current activity may be focused on a specific intervention, or at the level of the treatment plan, or the broader case conceptualization. Accordingly, the approach taken in the CES is to focus the measure on the current focus of the session, and to mirror the flexibility of application of CE, so that the measure can be used to rate CE as it occurs in each separate focus of activity across a session, whether at the level of a specific intervention, or at the level of treatment planning, or case conceptualization.

Overview of the Collaborative Empiricism Scale

The Collaborative Empiricism Scale centres on the current *focus of discussion* in a therapy session. The focus of discussion is the current object of activity or conversation, for example, talking about a problem at work, setting an agenda for the session or a goal for therapy, or carrying out in vivo exposure or a behavioural experiment. Each focus of discussion is rated as an *empirical event*. An empirical event is a specific exploration of a cognition, cognitive process, feeling, behaviour, or situation, centred on the current focus of discussion in the session. Examples of empirical events are 'rating the degree of emotion associated with a thought' (rated using Item B: 'Rating the degree of belief or emotion') and 'selecting a thought or feeling to explore' (rated using Item A – 'Selecting a focus of discussion').

Empirical events range in degree of empiricism and collaboration. Empiricism may range from no mention of the client's experience to a very high level of use of the client's experience. Similarly, collaboration may range from unilateral decision-making and one-sided, directive actions, to actively shared decisions and creatively shared work. An empirical event may also lead to other empirical events. For example, in exploring a distressing thought (rated using Item C – 'Exploring the focus of discussion') the client and therapist may go on to discuss an alternative to the thought (rated using Item D – 'Developing an alternative interpretation'). In this case, both these events would be rated separately.

To use the scale, a rater first identifies the current focus of discussion in the therapy session, and then decides whether this focus of discussion is the target of one of six empirical events matching the empirical items A – F. If so, the identified event is rated for empiricism using the matching item. The same event is then rated for collaboration using Item G. That is, each empirical event is rated twice, once using one of the empirical items A – F and once using the collaboration item G. The rater then rates the next focus of discussion and so on across the session. The CES is presented in Appendix A.

Overview of the Scale Development Process

The development of the Collaborative Empiricism Scale was guided by recommendations in the psychometrics literature (Baer, 2010; Christ & Boice, 2009; L. A. Clark & Watson, 1995; L. L. Cohen et al., 2008; Dawis, 1987; DeVellis, 2003; Hinkin, 1995; Worthington & Whittaker, 2006), and a review of the development processes used by previous authors in developing therapy process scales (Allen, Newsom, Gabbard, & Coyne, 1984; J. P. Barber, Liese, & Abrams, 2003; Blackburn et al., 2001; Denton, Johnson, & Burleson, 2009; Kim, Boren, & Solem, 2001;

McCormack, McCarthy, Wright, & Coffey, 2009; O'Malley et al., 1983; Shelef & Diamond, 2008; Tichenor & Hill, 1989; Vallis et al., 1986). Scale development proceeded through nine stages. A summary of the stages of development is presented in Table 1.

Table 1

Summary of Stages of Development of the Collaborative Empiricism Scale

Stage	Description	Summary of the development process
1	Definition of CE	A definition of CE was developed from a conceptual review of seminal CBT literature describing CE, and analysis of existing measures of the therapeutic alliance in CBT (see Chapter 1)
2	Review of collaboration and empiricism	Expanded definitions of collaboration and empiricism in CE were established by reviewing theoretical and empirical work on these constructs (see Chapters 2 and 3)
3	Core elements of CE identified	Core elements of the construct were identified from a conceptual review of the literature, including analyses of other therapy process measures, as well as close analyses of discussion of CE and CBT process in seminal texts in CBT
4	Item construction	Scale items were then developed to capture each of these construct elements, in an iterative process of creation and evaluation of prospective scale items
5	Textual revision of items	Textual revision of items and overall scale structure was undertaken to increase clarity and usability

6	Expert feedback on scale psychometrics and clarity	Thirty expert CBT researchers and practitioners provided quantitative and qualitative feedback on the construct validity, ecological validity, ability to capture variance in CE, clarity, and usability of the scale (Study 1: Chapter 5)
7	Scale revision	The scale was then revised to take account of expert feedback
8	Pilot testing to assess reliability and usability	The reliability and usability of the revised scale was assessed through two pilot studies (Study 2: Chapter 7)
9	Revision of rating methodology	Issues identified in the pilot studies were addressed, and the examination of the relationship between CE and therapy outcome commenced (Study 3: Chapter 8)

Global Aspects of the Collaborative Empiricism Scale

Once the definition of CE had been established from the literature but before item generation began, several decisions were made regarding the structure and focus of the CES.

Initial design choices.

Two design choices regarding the measurement of empiricism were made at the outset of scale development. First, it was thought that it would be easier to identify CE, and facilitate good reliability, if the scale focused on concrete therapy behaviours that could be conducted empirically, rather than, for example, conceptual dimensions of CE measured broadly across the session, as in the Working Alliance Inventory

(Hatcher & Gillaspy, 2006) and the Cognitive Therapy Rating Scale (J. E. Young & Beck, 1980a).

Second, CE has been consistently discussed in both the clinical and research literature as composed of two components, empiricism and collaboration, that merge together to form the CE construct (A. T. Beck et al., 1985; A. T. Beck et al., 1979; J. S. Beck, 1995, 2011). This two-component structure was retained in the CES in the form of two sub-scales of collaboration and empiricism. This structure has the additional advantages that it is consistent with the seminal literature, the theoretical analysis of CE outlined in Chapter 1 (see also, Tee & Kazantzis, 2011), and facilitates comparison with past research.

The focus of the scale: Process vs. session structure.

Based on the definition developed from the literature of CE as a fluid and emergent construct, it was decided to focus the scale on measuring empiricism and collaboration in several key aspects of therapy process in which empiricism characteristically occurs, rather than focusing on structural components of a therapy session, such as agenda setting or homework review. This approach has the advantage of allowing the scale to measure empiricism and collaboration across multiple levels of CBT process, from developing the overall case conceptualization, to specific interventions such as a thought record or behavioural experiment, to more fine grained analysis of individual segments of CBT process, such as rating emotion, or interpreting the results of a homework task.

Accounting for the dyad.

CE involves working together to empirically explore the client's cognition, cognitive processes, emotions, behaviour, and problem situations (A. T. Beck, 1967; A. T. Beck et al., 1979). As such, CE is an inherently dyadic construct. In working

together, the client's or therapist's attempts to collaboratively share empirical work may be closely related to the other's responses, or they may be distinct from them. Situations may arise, for example, in which the client or therapist is attempting to work in a collaborative or empirical manner, but for some reason the other party is not engaging in the process. In these situations, the client's or therapist's efforts to work collaboratively or empirically may be of a very high standard, but little collaborative or empirical work is actually accomplished. To reflect this, the CES was designed to measure the quality of CE actually achieved by the client and therapist dyad, rather than measuring attempts at CE or fostering conditions for CE.

Item Construction

Collaboration in the CES.

A list of elements of collaboration in CE were derived from the review of collaboration in Chapter 2, combined with a review of collaboration in CE in two seminal CBT texts (A. T. Beck et al., 1979; J. S. Beck, 2005). At the first level of analysis, behaviours of relevance to collaboration in CBT were identified and grouped into broad categories. Client and therapist behaviours related to collaboration in CE are presented in Table 2.

Table 2

Categories of Client and Therapist Behaviours Related to Collaboration in CE

Therapist	Client
Offers choices	Makes choices
Asks for suggestions	Makes suggestions
Gives other time to reflect, think, formulate a contribution	Gives other time to reflect, think, formulate a contribution
Incorporates or subsequently makes use of	Incorporates or subsequently makes use of

other's input	other's input
Yield's to other's judgment when they feel it is reasonable (responsiveness / engagement)	Yield's to other's judgment when they feel it is reasonable (responsiveness / engagement)
Asks open questions. Seeks opinions, priorities, information, and cognitions.	Contributes information: Reports introspective data such as thoughts, feelings, and wishes. Responds constructively to therapist's input
Uses techniques to foster participation. E.g., brainstorming, perspective taking.	Participates actively. Identifies useful beliefs or data to think about, Reflects on / recognizes automatic negative interpretations of experience. Participates in techniques to generate ideas
Seeks regular feedback: Checks understanding	Gives feedback
Supports client's autonomy: Follows client's lead / direction / priorities	
Fails to foster (encourage, support) client contributions	
Fails to give sufficient time for the other to think and contribute	Fails to give sufficient time for the other to think and contribute
Ignores the other's contributions	Ignores the other's contributions
Overrides the other, returns to own agenda	Overrides the other, returns to own agenda

The behaviours identified in Table 2 were then grouped into key aspects of collaboration in CE. Categories of behaviours related to collaboration in CE, and examples of each category, are presented in Table 3.

Table 3

Categories of Behaviours Related to Collaboration in CE

Category	Behaviour related to collaboration
Engagement	
Involvement in the process of therapy	Exerting effort
	Participating in tasks and interventions
Eliciting (or stifling) contributions	
	Asking for suggestions, opinions, priorities / goals / aims / agenda items
	Creating opportunities for input (e.g., waits, gives time for client to think or process input)
	Providing information
	Generating ideas / Brainstorming
	Perspective taking / alternative viewpoints / others' points of view on client's belief
Responsiveness	
	Makes responses that are congruent with the others' immediately preceding statement.
	Summarizes, or paraphrases.
	Topic determination. Follows other's topic

	initiations. Addresses previous communicative content and offers appropriate elaboration
	Ignores, or overrides other's input (i.e., returns to own agenda).
	Response is well-timed
Feedback	
	Seeks regular feedback
	Offers feedback
	Integrates feedback into further work
Shared decision making	
	Asks for input on decisions
	Offers choices regarding decisions
	Makes choices in decisions
	Identifies viable options
	Evaluates possible implications of choices / options

Key aspects of collaboration.

The categories of behaviour described in Table 3 above were further combined to form three core aspects of collaboration in the CES.

Mutually responsive interaction.

In the CES, the concepts of mutuality and responsiveness are combined to form *mutually responsive interaction*, as one of three core components of collaboration. Mutually responsive interaction represents a process that is mutual,

involving both parties in the client – therapist dyad; responsive, such that the interaction is flowing, emergent, and appropriately meets the required aims of the current therapy process; and interactive, integrating and building on contributions from each member of the dyad.

Feedback.

Feedback in the CES involves seeking and providing data, such as information or suggestions, in order to check the level of shared understanding and to adjust and improve interventions, goals, or the case conceptualization. Feedback can function at multiple levels within a session, including, for example, within an specific technique, at the level of a whole intervention, or goals for the entire course of therapy. Ideally, seeking and making use of feedback is done by both the client and therapist, although early in therapy it is more likely to be led by the therapist.

Feedback is a core component of CE because engagement in a collaborative sharing of the work necessarily requires judgments about the other party's thoughts, beliefs, feelings, and motivations, etc. Such judgments may shape and underpin both individual interventions and the therapy as a whole but the accuracy of these judgments may not be determinable by introspection, particularly between client and therapist (Kuyken et al., 2009; Messer, 1991; Persons, 2008). As a result, it is important to monitor how accurate and helpful these judgments are, as well as how they are being employed in the work of therapy. Feedback also directly improves interventions by enabling empirical checking of attributions and judgments about the other partner in therapy. Accordingly, in the CES, feedback appears both as a key element of collaboration, and as a specific element in the response choices of the empirical items.

Active contribution.

Colson and colleagues (1988) argued that the therapeutic alliance is best measured as client collaboration, which they defined as the degree to which the client cooperates with the intentions of the therapist, and is able to make use of the treatment offered (Colson et al., 1988). They operationalized cooperation as the extent of client engagement in the requisite treatment tasks, measured as the amount of client work done in session, the client's contribution of material (thoughts, feelings, issues, etc.), application of the work done in therapy between sessions, and adoption of processes, such as self-monitoring, with a view to using these autonomously in future. This conception of engagement captures several important aspects of sharing the work from the client's side, for example, the contribution of problems, phenomenal material (cognitions, feelings, etc.), and active participation. However, focusing solely on the client's engagement will not capture the work shared by the client and therapist dyad. The degree to which the client and therapist genuinely share the selection, development, deployment, and evaluation of therapeutic interventions is a quintessentially dyadic concept. In order to capture this, the CES extends the concept of engagement to the dyad, by focusing on the degree to which client and therapist both *actively contribute* to the session.

Sharing the work.

A key aspect of CE in CBT is that the client *shares the work* of therapy at all levels (A. T. Beck et al., 1979). This sharing the work is described as a 'collaborative effort' shared between client and therapist (A. T. Beck et al., 1985). The role of the client in CBT is also frequently described as a 'researcher' (e.g., A. T. Beck et al., 1985; A. T. Beck et al., 1979; J. S. Beck, 1995; D. J. G. Dobson & Dobson, 2009). This description refers to the client's active participation in planning interventions, collecting data by carrying out experiments (e.g., in the form of thought records,

behavioural experiments, or skills practice), and analysing the collected data. This shared work also explicitly informs the case conceptualization process (A. T. Beck et al., 1985; Kuyken et al., 2009). The collaboratively developed case conceptualization is then used to guide the selection of therapeutic goals, and development of further interventions in a cyclical process (D. J. G. Dobson & Dobson, 2009). A goal of CBT is for the client to learn to carry out this process independently, but the therapist actively guides and teaches, particularly early in treatment (D. J. G. Dobson & Dobson, 2009). In this sense, sharing the work is both the ideal process of CE in CBT, and an end goal of the therapy itself.

From the conceptual review above, and the literature review reported in earlier chapters, sharing the work in the CES was explicitly defined as: "Sharing the work involves shared decisions, mutually responsive interaction, and contributions from both parties (engagement, participation, effort). This is more than simply agreement or cooperation" (The Collaborative Empiricism Scale; Tee et al., 2012, p. 11).

The definition of sharing the work in the CES also incorporates the concept of *balance* as a further element of the degree of sharing the work in the session. Balance reflects the fact that in cognitive therapy the therapist typically takes a lead early in therapy in guiding interventions and formulation, but that ideally this becomes more balanced as the client develops skills in these processes. A consequence of this is that sharing the work may involve different activities for client and therapist at different stages of the therapy. As sessions progress over the course of therapy, the client will ideally be encouraged and fostered to act 'as their own therapist', and will take a more equal and active role in the development and evaluation of therapeutic interventions. The degree of shared work is therefore measured relative to the quality of the three core components described above, not whether the client is carrying out the same

behaviours as the therapist. The key idea here is that sharing the work need not involve client and therapist doing the same things per se, but each contributing to the process. In this context, *ideal balance* is defined in the CES as:

> The client and therapist share the work to the extent that they are able given the stage of therapy. For example, early in therapy, the client may contribute less to the design of an empirical test, as this is a new activity for the client, but the client can still fully share the work by making contributions, involvement in shared decisions, and mutually responsive interactions (The Collaborative Empiricism Scale; Tee et al., 2012, p. 11).

Empiricism in the CES.

Definition of empiricism in the Collaborative Empiricism Scale.

Beck and colleagues defined empiricism in cognitive therapy as a "weighing of empirical evidence", which is "based on, concerned with, or verifiable by observation or experience rather than theory or pure logic" (A. T. Beck et al., 1985, p. 182). This involves the therapist "encouraging the patient to identify, observe, and evaluate his thoughts in an objective manner" (A. T. Beck et al., 1979, p. 55). Based on this definition, empiricism in the CES was defined as a process of identifying and exploring the usefulness and functionality of the client's cognitions, cognitive processes, emotions, and behaviour, by exploring these in terms of the client's experience, rather than solely from general principles or on logical grounds. This empirical exploration is typically focused in one of several key aspects of empiricism, each of which corresponds to an empirical item on the scale.

Key aspects of empiricism.

The elements of empiricism focused on in the CES were derived from an analysis of discussions of CBT process in two CBT manuals that are regarded as seminal and authoritative, and are widely employed in training and research (A. T. Beck et al., 1979; J. S. Beck, 1995). First, a close analysis was made of discussions of the process of CBT in Beck et al. (1979), in order to identify processes relevant to empiricism. This analysis was then cross-referenced and augmented by repeating the process with J. S. Beck (1995). Unlike the structural aspects of a typical cognitive therapy session (e.g., agenda setting, homework review), which characteristically occur as discrete stages in a session, empiricism in cognitive therapy is typically focused on several identifiable processes that can occur at any time, often multiple times in a session, and may function at different levels within a session.

This analysis identified broad categories of empirical behaviours, such as, 'choosing a focus of discussion', 'eliciting raw data from the client's experience', and 'evaluating the results of an empirical test'. The elements of empiricism identified were then cross referenced with the CBT workflow illustrated in both texts in order to derive a final list of focuses for empiricism in CBT. As an illustration of this procedure, J. S. Beck (1995) breaks the process of 'examining thoughts' down into several component elements. Table 4 presents these component elements and the focuses of empirical behaviours corresponding to them.

Table 4

Components of the Process of 'Examining Thoughts' and Corresponding Focuses of Empiricism

Component of 'examining thoughts'	Focus of empiricism
Identifying a thought to examine	Selecting something to examine or explore
Uncovering evidence	Exploring (e.g. within the 5 part model of CBT)
Questioning automatic thoughts	Exploring the usefulness or validity of a cognition or cognitive process
Devising a reasonable alternative belief	Developing an alternative explanation

Focuses of empiricism measured in the CES.

As a result of the analyses described above, a final list of six categories of empirical process were identified as focuses of the empirical items in the CES. These six processes are presented in Table 5.

Table 5

Focuses of the Empirical Items in the Collaborative Empiricism Scale

	Focuses of the Empirical Items in the Collaborative Empiricism Scale
1	Selecting a focus of discussion (cognition, cognitive process, behaviour, intervention, therapy goal, etc.)
2	Rating the degree of belief or emotion associated with the current focus of the session
3	General exploration of the focus of session, with respect to the client's experience, taken as raw data for objective examination
4	Discussing / developing alternative explanations for cognitions, cognitive processes, emotions, etc.
5	Developing an empirical (experimental) test of a cognition or cognitive process
6	Evaluating the results of an empirical test of a cognition or cognitive process

The empiricism items (A – F).

Once the core empirical processes to be measured had been identified, items were generated to measure each process. Any general discussion in a cognitive therapy session can be undertaken with greater or lessor reference to the client's experience. Accordingly, Item C 'Exploring the focus of discussion' was designed to measure the quality of empiricism in any general exploration of the current focus of discussion in the session. Five other items each focus on the quality of empiricism in one of the empirical processes outlined in Table 5 above. Table 6 lists each empirical process and its corresponding scale item.

Table 6

Empirical Processes and Corresponding Scale Items

Empirical process	Scale item measuring the process
Selecting a focus of discussion	Item A. Selecting a focus of discussion
Rating the degree of belief or emotion	Item B: Rating the degree of belief or emotion.
Exploring of the focus of discussion	Item C. Exploring the focus of discussion
Developing an alternative explanation for cognitions or cognitive processes	Item D. Developing an alternative explanation
Developing an empirical test	Item E. Designing an empirical test
Evaluating an empirical test	Item F. Evaluating the results of an empirical test

Item A. Selecting a focus of discussion.

This item focuses on the act of selecting something to do or discuss. Selecting a therapeutically efficacious cognition or cognitive process to work on is an important aspect of CBT that is routinely discussed in CBT training manuals (e.g., A. T. Beck et al., 1979; J. S. Beck, 2011; Padesky & Greenberger, 1995). Selecting a focus of therapeutic discussion or activity may also involve, for example, choosing a behaviour to evaluate, or a situation or decision to problem solve. Selecting also frequently operates at multiple levels within a session. At the level of the case conceptualization, specific goals may be selected from alternatives, and interventions selected as part of working towards these goals. At the level of discussion, problematic cognitions and issues may be selected to discuss, or priorities set about

which issue to discuss first (e.g., setting an agenda). Each act of selection may involve a high or low degree of empirical reasoning and exploration.

Item B. Rating the degree of belief or emotion.

Rating the degree of belief or emotion associated with a cognition or cognitive process has several established functions in CBT. It is typically employed in measuring levels of belief or emotion before and after an intervention. It is therefore a key process in gauging change in cognition or emotion, and thereby also a measure the effectiveness of therapy interventions. Common examples include monitoring changes in emotion or degree of belief during work with a thought record or evaluating the effect of specific behaviours or therapeutic interventions on beliefs or mood (J. S. Beck, 1995, pp. 94-104; Leahy, 2003; Persons, 2008), or before and after in vivo or imaginal exposure, or behavioural experiments (O'Donohue & Fisher, 2008). Rating the degree of emotion also occurs in contexts other than measuring change. For example, if clients have difficulty identifying the strength of an emotion, this may be overcome by associating the client's various experiences of the emotion with key points on a subjective self-rating scale (e.g., 0, 25%, 50%, 75%, 100%) (J. S. Beck, 1995).

Item C. Exploring the focus of discussion.

As discussed above, Item C aims to capture the use of the client's experience in the general process of discussion in a therapy session, including, for example, during guided discovery. Exploring involves the examination, surveying, review, or evaluation of the client's cognition, cognitive processes, feelings, or behaviours, with respect to the client's experience (of self, world, and other) and the key situations in which these occur.

Item D. Developing an alternative explanation.

Developing an alternative explanation involves "the active investigation of other interpretations or solutions to the patient's problems" (A. T. Beck et al., 1979, p. 158). As such, it involves using collaborative empiricism to explore new ways of thinking and behaving, and is therefore also closely linked to the process of change in CBT. Developing an alternative explanation typically appears in the process of devising a reasonable alternative to a belief in a thought record, as well as the reattribution and cognitive restructuring techniques described by Beck and colleagues (A. T. Beck et al., 1979, pp. 157-158; J. S. Beck, 2005, 2011).

Items E & F: Designing and evaluating an empirical test.

Items E and F are designed to rate the development and subsequent evaluation of an empirical test. Empirical tests are a distinguishing feature of CBT. They typically involve the experimental exploration of the accuracy or helpfulness of beliefs and other cognitive content, cognitive processes, emotions, or behaviours. The paradigm example of an empirical test is the behavioural experiment, carried out either during the therapy session or between sessions (Bennett-Levy et al., 2004). Item E rates the quality of empiricism and use of the client's experience in the development of an empirical test. It focuses on the degree to which the experiment is grounded in the client's experience. Also important is the quality of the test design. This includes, for example, distinguishing subjective interpretations of a situation from more objective evidence, anticipating difficulties in carrying out the experiment, and articulating clear criteria for the outcome of the test (Bennett-Levy et al., 2004; Nelson, 1997).

Global Rating Scheme

In keeping with the aim of flexibly measuring empiricism across multiple contexts in CBT, and to facilitate reliable rating, the scale was designed with a common underlying rating scheme across all empirical items. Each empirical item on the scale is rated using a shared scheme that is independent of the particular empirical process being measured, apart from specific elements required to account for unique aspects of the process itself.

An early draft of the rating scheme was organized around detecting 'key features' of empiricism in the focus of discussion being rated. Higher ratings reflected more key features, and the absence of problems with the key features. An early version of the global rating scheme for the empirical items is presented in Table 7.

Table 7

Early Version of the Global Rating Scheme for Empirical Items

Rating	Use of the client's experience in the rated focus of discussion
1	MINIMAL. Minimal or no incorporation of key features
2	INADEQUATE. Some key features but others missing
3	ADEQUATE. All key features, but some problems that reduce the quality of key features
4	EFFECTIVE. All key features. Minor problems that do not substantially reduce key features
5	EXCELLENT. All key features. No problems or problems resolved

As development of the scale proceeded, it became clear that some general psychotherapy competences could also impact on the quality of CE. These competencies are aspects of good CBT practice but not core aspects of CE per se. An

example of this issue is choosing an appropriate belief to work with in session. Choosing an appropriate belief is part of good CBT practice but is not a specific element of CE per se. It is important for CE, however, because CE may be of little use clinically if a therapeutically unimportant belief is chosen to explore. Examples of such a belief is one that is marginal to the client's concerns, not held strongly, occurs in few or limited situations, or is not central to the case formulation (J. S. Beck, 1995).

To address this issue, the rating scheme was revised to focus specifically on the use of the client's experience in carrying out the processes being observed. The revised rating scheme describes a broad progression of increasing quantity and quality of the use of the client's experience in CE. This ranges from *mention* of the client's experience to active *use* of the client's experience, within which there may be either *problems with the use of experience*, or *minor problems* that do not directly reduce the quality of the use of experience per se; and finally to excellent use of the client's experience, where problems are absent or resolved. The global rating scheme underpinning the empirical items is presented in Table 8.

Table 8

Global Rating Scheme Underpinning the Empirical Items (A – F)

Rating	Use of the client's experience in the rated focus of discussion
1	No *mention* of the client's experience
2	*Mention* or discussion of the client's experience, but no *use* or application of the client's experience to explore or evaluate cognitions or cognitive processes
3	*Use* of the client's experience to explore the focus of discussion, *but with problems with empiricism* that reduce the quality of the use of experience. (Examples of problems with empiricism include mistaking interpretations of experience as facts; failing to specify clear criteria for the outcomes of an empirical test; ratings of belief or emotion that are vague or unclear; and empirical evidence sourced from the therapist, rather than from the client)
4	Application of the client's experience to the focus of discussion, but with minor problems. (Minor problems do not reduce the adequacy of empiricism as the basis for exploring. Examples of minor problems can be problems in counselling skills, or missed opportunities to do adequate CE better.)
5	Application of the client's experience, in which there were no problems, or any difficulties were resolved, and feedback was offered or sought.

The global rating scheme is common to all empirical items; however, minor differences exist in the anchor point descriptions of some items, in order to capture the unique features of the process being measured in the item. For example, the global

rating scheme rates an item ‘1’ if no mention is made of the client’s experience in carrying out the rated process. In the case of in Item E (Designing an empirical test), it is not possible that no use is made of the client’s experience because the test itself is either carried out in session and therefore involves the client’s experience in real time, or is set for homework and involves the client’s life experience between sessions. To account for this, Item E is rated ‘1’ if ‘only general or vague reference is made to the client’s experience’. A small number of similar minor adjustments apply to the global rating scale where needed to account for the unique characteristics of specific items.

Concept check: Rating Empiricism in CBT Interventions

As a check on the structure of the empirical items during development, the CES was used to rate standard interventions in a CBT, such as the thought record and the behavioural experiment, to ensure that the scale was able to capture empiricism as it occurred in these interventions. To do this, the student researcher rated a number of CBT training video recordings produced by the American Psychological Association. These recordings featured expert cognitive therapists, including Dr. Judith Beck, Dr. Arthur Freeman, & Dr. Christine Padesky, conducting therapy sessions to demonstrate core competencies of CBT. Test ratings were used to improve the ease of rating the scale and to clarify technical definitions for the rater.

Conclusion

This chapter reported on the development of the Collaborative Empiricism Scale. First, overviews were presented of the process of CE in a CBT session, the scale, and the scale development process. Global aspects of the scale were then discussed, including initial design choices to centre the scale on concrete behaviours focused on key empirical processes, to focus on the dyad, and to retain the two subscale structure of collaboration and empiricism from the literature. The process of

item development was then discussed. Building on the literature review of evidence-based relationship attributes, the process of collaborative empiricism was closely analysed in two seminal CBT texts, spanning 35 years of the literature. Discussions of collaboration and empiricism in these texts were categorized into key aspects of CE, and individual scale items designed to measure each key aspect. Lastly, the global rating scheme common to the empirical items was discussed.

Chapter 5: Expert Review of the Collaborative Empiricism Scale

Outline and Aims

This chapter reports on a review of the CES by expert CBT researchers and practitioners. The aims of the study were to get expert feedback, identify problems, and revise the scale in preparation for pilot testing. Subject matter experts were invited to review the validity and clarity of the CES, as well comment on the structure of the scale, the response options available for each item, and any other issues they identified. The study was successful in attracting the participation of 30 experts, who rated each scale item on a number of Likert scales and made 104 feedback comments. The gathered expert feedback was then evaluated in detail and used to revise the scale prior to pilot testing.

Content Validity

A primary aim of the expert feedback study was to assess the content validity of the scale. Content validity is the agreement of the content of an instrument with the domain it purports to measure (DeVellis, 2003). Ensuring that a measure adequately captures its target construct is a crucial aspect of scale design (C. T. Beck & Gable, 2001; Grant & Davis, 1997; Sireci & Geisinger, 1995). A recommended and commonly used method of assessing content validity is to measure the agreement between subject matter experts regarding how accurately a scale's items describe aspects of the target construct (C. T. Beck & Gable, 2001; L. L. Davis, 1992; Gajewski et al., 2012; Lawshe, 1975).

Expert Review

In the context of scale development, expert review has been frequently recommended as a method for evaluating the content validity of items generated during construction of a measure (for review see, Delgado-Rico, Carretero-Dios, &

Ruch, 2012), and as an important stage of pretesting a scale, particularly when the measure is the first to capture a construct (DeVellis, 2003; Nunnally & Bernstein, 1994; Sireci, 1998). Expert review has been found to be a highly productive method of identifying problems in a measure (Willis, Schechter, & Whitaker, 2000). It has been employed in developing behaviourally anchored rating scales and behavioural summary scales (Mastaglia, Toye, & Kristjanson, 2003; Stark, Chernyshenko, & Guenole, 2011), measures of patient attributes (En-Hong & Yong, 2012), and assessing a measure's content domain representation (Sireci & Geisinger, 1995).

The Delphi method.

The Delphi method is a structured communication process that allows a group of experts to deal effectively with a complex problem (Linstone & Turoff, 1975; Yousuf, 2007) and enable more accurate assessments than those obtainable by individuals alone (Rowe, Wright, & Bolger, 1991). A key focus of the Delphi method is the attainment of consensus between experts in a group (Fish & Busby, 1996). The details of the Delphi method vary between specific implementations (Powell, 2003; S. J. Young & Jamieson, 2001); however, several core features of the technique are typically evident, including anonymity of experts, controlled feedback to the group, and statistical aggregation of the group's responses (Dalkey, 1972; Keeney, Hasson, & McKenna, 2001).

To conduct a Delphi study, a group of subject-matter experts is identified. The size of the group varies in the literature but between 10 to 15 is regarded as sufficient if the group is homogenous and not evaluating multiple domains (Delbecq, Van de Ven, & Gustafson, 1975). A majority of studies have included between 15 and 20 experts (Boulkedid, Abdoul, Loustau, Sibony, & Alberti, 2011; Ludwig, 1997). Once the group is identified, an initial questionnaire is sent, asking experts to assess a target

issue on one or more criteria. The collected data, which is kept strictly anonymous, is evaluated to determine expert consensus. Consensus is defined beforehand as agreement between experts to a specified level (J. Jones & Hunter, 1995). The anonymity of the Delphi process overcomes several drawbacks of evaluation by groups, such as the potential for influence of the group by prestigious, senior, or dominant individuals (Dalkey, 1972; Murphy et al., 1998), and the fear of potential social or professional repercussions for divergent views (Linstone & Turoff, 1975). The de-identified, statistically aggregated results of experts' analyses may then be sent back to experts for re-evaluation and revision of their analyses if they choose.

The Delphi method has been widely used in psychotherapy research outside of scale development, including to develop guidelines for clinical trial protocol content (Tetzlaff, Moher, & Chan, 2012), predict future trends in psychotherapy practice (Couch & Childers, 1991; Norcross, Hedges, & Prochaska, 2002) and clinical child psychology research and practice (Kaufman, Holden, & Walker, 1989), identify factors that effect engagement in psychosocial treatment for personality disorders (Jinks, McMurran, & Huband, 2012), determine discredited aspects of treatment (Norcross, Koocher, Fala, & Wexler, 2010; Norcross, Koocher, & Garofalo, 2006), develop a model of competency for implementing CBT (Sburlati, Lyneham, Mufson, & Schniering, 2012), evaluate the competency in psychotherapy of psychiatric residents (Giordano & Briones, 2003), identify issues in the training of family therapists (Sori & Sprenkle, 2004), determine appropriate and inappropriate uses of humour in psychotherapy (Thomson, 1990), identify critical competencies for psychotherapeutic practice with eating disordered clients (Williams & Haverkamp, 2010), and in the development of a scale to assess the quality of reports of

randomized controlled trials for psychological treatments (Yates, Morley, Eccleston, de, & Williams, 2005).

Despite the widespread implementation of Delphi in psychological research, and recommendations in the psychometrics literature to employ expert review in the scale development process (L. L. Davis, 1992; Delgado-Rico et al., 2012; Haynes, Richard, & Kubany, 1995), to date few process measures in psychology have utilized expert review, apart from the experts who designed the measures. As an illustration of this, a search was conducted on 2 May 2013 via the PsycINFO (1987 – 2013) database of the Journal of Consulting and Clinical Psychology, Journal of Abnormal Psychology, Psychological Assessment, Behaviour Research and Therapy, British Journal of Clinical Psychology, and Clinical Psychology Science and Practice. Search terms used were 'expert AND review' OR 'expert AND feedback' OR 'expert AND survey'. The search returned only 7 articles, none of which involved the use of expert review in scale development or psychotherapy process research. Similarly, a recent review of measures of homework adherence reported no mention of expert feedback as part of the development process (Kazantzis, Deane, & Ronan, 2004).

Expert review of the CES.

The CES benefitted from two levels of expert review. First, the working definition of CE and the theoretical underpinnings of the scale were reviewed as part of the process of publishing a review/theory paper in a peer reviewed journal (Tee & Kazantzis, 2011) (see Appendix B for reprint). This paper outlined a definition of CE, grounded closely in the seminal work of A. T. Beck et al. (1979) and J. S. Beck (1995), as well as a new theoretical model of the CE construct. This review process was continued in subsequent discussion in the literature of the paper and the theory of CE underpinning the scale (D. A. Clark, in press; Deborah J. G. Dobson & Dobson, in

press; Hutton & Morrison, in press; N. Kazantzis et al., in press; Nikolaos Kazantzis, Timothy J. Cronin, et al., in press; Kazantzis et al., 2013; Wong, in press). As a result of this peer review and discussion, the CES attained a first level of consensus regarding the working definition and theory underpinning the scale.

Post-development evaluation.

Once initial scale development was complete, a formal process of expert review was undertaken. In the context of scale development, the Delphi method has typically been employed early in the development process to seek expert consensus on whether a provided list of prospective attributes of a construct are essential or redundant (Fish & Busby, 1996; Linstone & Turoff, 1975; Yousuf, 2007). This approach was not taken with the CES because it was considered that the utility of the scale and its relevance for past and future research would be increased by grounding the scale closely in the seminal CBT literature. This was deemed particularly important because the CES is the first measure of CE in the literature. In accordance with this aim, the definition of CE used in the scale, and the specification of the individual empirical items, were drawn from close analyses of seminal CBT texts by A. T. Beck et al. (1979) and J. S. Beck (1995). These two texts have been widely cited as authoritative and have previously served as the basis for the definition of cognitive therapy in a large-scale study of CBT for depression (Dimidjian et al., 2006). This process was consistent with the Delphi method in that it was highly anonymous and sought consensus from a group of subject matter experts on ratings of a number of core attributes of the CES. It is also consistent with long standing variations of Delphi to use a more structured questionnaire (e.g., Kerlinger, 1973).

The approach taken with the CES also differed from the most common Delphi approach in that experts rated several attributes of the scale items on 5-point Likert

scales. This process provides a more quantified evaluation of aspects of the scale than that obtained using a pre-development expert review of which aspects of a construct to include or exclude from a measure. Post-development evaluation of a draft scale allows experts to make the same comments they might have made in a pre-development Delphi questionnaire, but additionally allows comparison of experts' views with a scale that is grounded in a previously peer reviewed definition and theoretical conception of the construct in the literature.

Definition of expert agreement.

Gathering expert feedback via 5-point Likert scales raises the issue of how to define agreement for the purposes of determining expert consensus. In the Delphi method, consensus is typically defined as a percentage of expert agreement. There is no clear standard in the literature for defining the cut-off point on a Likert scale in a Delphi review. Possibilities that have been used include 80% of ratings falling between two points on a 7-point scale (Ulschak, 1983) and 70% of ratings above 3 on a 4-point scale, with a median score of at least 3.25 (Hsu & Sandford, 2007).

An existing approach in the literature for assessing content validity on a Likert scale is the Content Validity Index (CVI; C. W. Waltz & Bausell, 1981). The CVI has been frequently used and evaluated in health research (C. T. Beck & Gable, 2001; Gajewski et al., 2012; Lynn, 1986; Polit & Beck, 2006; Polit & Hungler, 1991; Wynd, Schmidt, & Schaefer, 2003). The CVI measures the degree to which a scale item accurately describes an aspect of the construct it purports to measure (DeVon, Block, Moyle-Wright, Ernst, & et al., 2007; C. W. Waltz & Bausell, 1981). Typically, the CVI is calculated by rating each item on a four-point ordinal scale with anchor points 1 = *not relevant*; 2 = *unable to assess relevance without item revision or item is in need of such revision*; 3 = *relevant but needs minor alteration*; and 4 = *very relevant*

and succinct. The CVI is then calculated as the proportion of experts that rate the item at least 3 out of 4 (L. L. Davis, 1992).

During development of the expert feedback questionnaire, a design decision was taken to use a 5-point Likert scale to rate scale items. It was considered that would be informative to allow experts to express a middle point 'somewhat' rating, rather than force a choice either for or against each item. Adding a middle anchor point has the additional benefit that positive ratings for an item can be regarded as more deliberate, and therefore more valid, because the opportunity was provided for the expert to make an ambivalent middle rating. As a result of this shift to a 5-point Likert scale, the CVI for each item was calculated as the proportion of experts who rated the item either a 4 (*large extent*) or 5 (*very large extent*). This strengthens the CVI as an index of validity by raising the standard of expert agreement.

Psychometric analysis of the CVI has demonstrated that with 10 experts, 78% of experts rating 3 or higher on a 4-point Likert scale is sufficient to establish agreement between experts at the .05 level of statistical significance (Lynn, 1986). With more experts, a lower figure is required. Combining the Delphi and CVI standards in the post-development evaluation of the CES, consensus was defined to be agreement between at least 80% of experts (the Delphi standard), with 78% of experts rating 4 or higher on each 5-point scale (following the CVI). This definition of agreement provides a very high standard of expert consensus, particularly given that the expert review of the CES attracted the participation of 30 experts, three times the 10 required for statistically significant expert consensus on the CVI.

Method

Expert participants

Experts were sought with experience in scale construction and data analysis, as well as expertise in the target areas of CBT process research and therapy practice. Expert selection followed established practices in the literature (e.g., Dillman, 2007; Polit & Hungler, 1991), and detailed criteria for expert selection in the Delphi method (Adler & Ziglio, 1996; L. L. Davis, 1992). Experts were invited to complete a feedback questionnaire if they met one or more of the following selection criteria: (a) they were authors of scales used in the present study, including the Beck Depression Inventory (used as an outcome measure) and the Cognitive Therapy Rating Scale (used as a measure of therapist competence); (b) were keynote speakers in the past 10 years at the World Congress of Behavioural and Cognitive Therapies; (c) had published commentary or clinical discussion on collaborative empiricism; or (d) were corresponding authors in the last decade on publications in CBT process research, in particular on the therapeutic relationship, collaboration, empiricism, or CE in CBT. To identify authors in criterion (d), a PsycINFO database search was conducted on 28 Feb 2011, for the years 2000 – 2010, with the search terms: “collaborative empiricism”, “collaboration AND empiricism”, “collaboration AND cognitive therapy”, “empiricism AND cognitive therapy”, and “empiricism AND process AND psychotherapy”. This search returned 149 articles, 96 in peer reviewed journals. Additional sorting was undertaken to remove articles not relevant (e.g., articles focused on psychodynamic therapy, articles in disciplines other than psychology), resulting in a total of 57 articles.

Using the above criteria, two groups of experts were identified. One group was comprised of expert CBT researchers, many of whom also had a past or present

professional practice in psychology. The other group were expert CBT practitioners. Expert practitioners were defined to be individuals that met criteria (b) and (c) above (key note speakers, authors of clinical commentary or discussion on CE) but who did not have an identified program of research, and who had not published an empirical study in the past year. In total, 106 experts were identified (86 researchers, 20 practitioners). These individuals were then invited by e-mail to complete the online feedback questionnaire. E-mail addresses were identified from those listed as corresponding authors of journal articles, by searching the websites of academic institutions, and via Google search. The final number of subject matter experts who undertook the feedback questionnaire was 33.

Measures

Subject-matter expert feedback questionnaire.

The questionnaire was constructed using Qualtrics online questionnaire development tools (Version 37.892, http://www.qualtrics.com). Qualtrics allows researchers to develop questionnaires that are completed online using a web browser. Two versions of the questionnaire were created that involved evaluating CES items on a series of 5-point Likert scales, ranging from *to a very small extent*, *to a small extent*, *somewhat*, *to a large extent*, and *to a very large extent*. The first version of the questionnaire was used to seek practitioner feedback. Expert feedback on each item of the CES was sought with two questions for the practitioner version. The researcher version included additional psychometric content and consisted of five questions repeated for each of the seven scale items. A summary of the content of both versions of the questionnaire is presented in Table 9.

Table 9

Summary of Question Content for Researcher and Practitioner Questionnaires

Feedback question	Questionnaire content: Researcher version	Questionnaire content: Practitioner version
Q1	Accurate description of CE[a]	Accurate description of CE
Q2	Captures variability in CE	Captures variability in CE
Q3	Item clear / comprehensible	-
Q4	Anchor points clear	-
Q5	Routinely used in your practice	-

Note. Feedback questions were scored on a 5-point Likert scale ranging from *To a very small extent*, *To a small extent*, *Somewhat*, *To a large extent*, and *To a very large extent*.
[a] For example, Question 1 read: To what extent does the item description accurately reflect an element of collaborative empiricism?

Several suggestions from the literature were incorporated into the questionnaire design (see Dillman, 2007; Perepletchikova, Treat, & Kazdin, 2007; Prochaska & Norcross, 1983; Sheehan, 2001). For ease of use, the questionnaire had a common structure across all questions, and instructions for use were simplified wherever possible. Additionally, instructions were presented in blue font, so that participants could easily distinguish instructions from the scale items in black font. Figure 2 presents a sample page from the questionnaire. The full questionnaire is presented in Appendix C.

Please read "Item B – Rating Emotion or Belief in Thoughts and Thought Processes" and answer the underlined Questions for Expert Feedback below.

Item B – Rating Emotion or Belief in Thoughts and Thought Processes

Strongly held thoughts that elicit strong emotion (sometimes referred to as 'hot thoughts'), as well as thoughts processes (e.g., catastrophizing or minimizing positive experience) are often important for CBT. Rating the strength of belief or associated emotion can serve as the basis for an empirical exploration of a thought or thought process, such as in the context of a 'thought record'. The emotion again serves as the basis for evaluating the utility of the intervention.

B	HOW WELL did the client and therapist rate the strength of belief or emotion associated with a thought?	
1	A thought was rated, but NO REFERENCE was made to the client's experience, and NO USE OF THE RATING was made to determine how helpful evaluating a thought had been, or whether to continue with that thought or move on (e.g., Therapist: "Which of those thoughts should we look at first?" Client: "Maybe, 'If my report is late, I'll get fired'?" Therapist: "Okay.")	
2	PASSING REFERENCE was made to the client's experience of a thought, but this was NOT THE BASIS for rating the strength of belief or emotion associated with that thought (e.g., Therapist: "Which of those thoughts should we look at first?" Client: "Well, I'm really worried about my report being late." Therapist: "Okay, let's start there.")	
3	The client's EXPERIENCE WAS THE BASIS for rating the strength of belief or emotion associated with a thought (e.g., Therapist: "Which of those thoughts do you feel most stressed about?" Client: "Getting fired." Therapist: "More than being judged by colleagues?" Client: "Yes"). However, there were SOME PROBLEMS which reduced the use of experience to rate the strength of belief or emotion associated with a thought (e.g., mistaking interpretations for evidence, ratings of strength of belief or emotion were vague, ambiguous, or were inadvertently suggested by the therapist, rather than sourced from the client).	
4	The client's EXPERIENCE WAS THE BASIS for rating the strength of belief or emotion associated with a thought. However, there were MINOR PROBLEMS, which did not reduce the use of experience as the basis for rating the strength of belief or emotion associated with a thought (e.g., some variability in focus on the task, some doubt or concern about rating remained unaddressed).	
5	The client's EXPERIENCE WAS THE BASIS for rating the strength of belief or emotion associated with a thought. There were NO PROBLEMS, or any difficulties WERE RESOLVED.	

Questions for Expert Feedback:

To what extent does the item description accurately reflect an element of collaborative empiricism?

To a very small extent ○ To a small extent ○ Somewhat ○ To a large extent ○ To a very large extent ○

To what extent does this item capture variability in collaborative empiricism?

To a very small extent ○ To a small extent ○ Somewhat ○ To a large extent ○ To a very large extent ○

To what extent is this item clear and easy to understand?

To a very small extent ○ To a small extent ○ Somewhat ○ To a large extent ○ To a very large extent ○

To what extent are the anchor points (1-5) in this item clear and easy to understand?

To a very small extent ○ To a small extent ○ Somewhat ○ To a large extent ○ To a very large extent ○

Figure 2. Sample page from the Expert Feedback Questionnaire.

Procedure

Prospective participants were sent a recruitment e-mail informing them that the researchers had developed a new measure of CE and that expert feedback was sought regarding the measure. Recruitment emails were addressed to participants personally and it was highlighted that participants' input was highly valued and would aid the development of a new measure of an important therapy process in CBT. Up to three emails were sent to each prospective participant. The first e-mail introduced the research project, provided a Participant Information Sheet, and included a web link to a PDF reprint of a paper conveying the background and conceptual underpinnings of the scale (Tee & Kazantzis, 2011). Up to two reminder emails were sent at intervals of six weeks. The reminder email frequency was set at six weeks as the experts contacted were known to receive large quantities of email, and 2 – 4 weeks is generally considered a reasonable time for responses to unexpected requests for consultation. The Participant Information Sheet and recruitment emails are presented in Appendix D and E respectively.

Ethics approval.

The Participant Information Sheet also described the ethical conditions of the questionnaire and gave information relating to the confidentiality of participants' responses. The project received La Trobe University Faculty of Science, Technology and Engineering Human Ethics Committee approval (number FHEC09.R59). In line with Ethics Committee approval, questionnaire responses were automatically de-identified by the questionnaire software and responses were not linked to individuals.

Results

SPSS Version 21 was used to conduct all statistical analyses. Inferential statistics were evaluated against an alpha level of .05.

Preliminary Data Analyses

Demographics were not collected, as a high degree of anonymity was considered imperative for the methodology. This was both to protect the privacy of experts and to encourage experts to feel free to be critical in reviewing the scale, particularly as many of those contacted also had existing relationships with La Trobe University Cognitive-Behaviour Therapy Research Unit researchers. As a result, it was not possible to further distinguish group characteristics of participants who completed the questionnaire from those who did not.

Missing data.

The SPSS FREQUENCIES procedure was used to screen for missing data. Data from three researcher participants were excluded due to partial responding. One partial responder scored the first four (of 35) questions as '5' (out of a possible 1 – 5) and then did not complete any further questions (88.6% missing); one provided no further responses after question 20 (of 35; 42.8% missing); and the third had 10 (of 35; 28.6%) responses randomly missing. Data from these participants were excluded listwise from subsequent analyses. Excluding the third partial responder with randomly missing data had the negligible effect of increasing the overall mean score from 4.03 to 4.04.

Out of the remaining data, 11 responses out of 882 (1.25%) were missing (researchers = 9 (1.01%), practitioners = 2 (0.23%)). No pattern could be detected in these data and they were considered to be randomly scattered. Less than 5% of missing data in a random pattern is considered in the literature to be less serious and to respond similarly to methods for treating it (Tabachnick & Fidell, 2007). Accordingly, for these 11 missing data, the respondents' data were retained but excluded pairwise in further analyses.

Outliers.

A respondent was deemed to be an outlier if the mean of their scores was more than three times the interquartile range above or below the mean of all scores ($M \pm 3*(IQR)$) (Tabachnick & Fidell, 2007). One respondent met this criteria. This respondent had the lowest mean feedback score in the group. As this was expert feedback, these data were retained unmodified but the low mean score for this respondent was noted when interpreting qualitative feedback comments for individual questions.

Quantitative Results

Experts' ratings of the CES.

The mean score of experts' ratings (researchers and practitioners combined) across all feedback questions was high ($M = 4.04$, out of 5), and variation between experts' ratings was low ($SD = 0.54$). A Kolmogov-Smirnov test indicated that mean scores on feedback questions were normally distributed, $D\ (31) = 0.132$, $p = .178$. The means and standard deviations of scores for each feedback question (Q1 – Q5) for all experts combined are presented in Table 10.

Table 10

Means and Standard Deviations of Scores for Each Feedback Question (Q1 – Q5) for all Experts Combined

Feedback Question	Question Content	*M (SD)*
Q1	To what extent does the item description accurately reflect an element of collaborative empiricism?	4.27 (0.57)
Q2	To what extent does this item capture variability in collaborative empiricism?	4.03 (0.63)
Q3	To what extent is this item clear and easy to understand?	3.82 (0.62)
Q4	To what extent are the anchor points (1 – 5) in this item clear and easy to understand?	3.79 (0.64)
Q5	To what extent does this item describe an aspect of collaborative empiricism that you routinely use in your professional practice?	4.07 (0.73)

As seen in Table 10, experts rated the scale items very highly (above 80%) for content validity (Q1), ecological validity (Q5), and capturing variability in CE (Q2), and highly (above 75%) for the clarity of the items (Q3) and clarity of the item anchor points (Q4). This pattern was consistent across all experts and all feedback questions. The uniformly high scores indicate that experts strongly endorsed the validity, clarity, and ability to capture variance of the CES. The slightly lower (but still high) scores for the two questions relating to clarity indicate that experts had some concerns regarding the clarity of the items and anchor points. This pattern of results is highly

consistent with the feedback comments experts made, discussed in the qualitative results section below.

The pattern of experts' responses for each feedback question across each scale item can be seen in Table 11, which presents the means and standard deviations of scores for each feedback question across each scale item for researchers and practitioners combined.

Table 11

Means (and Standard Deviations) of Scores for Each Feedback Question for Each Scale Item for Researchers and Practitioners Combined

	Feedback question *M* (*SD*)				
Scale Item	Q1	Q2	Q3	Q4	Q5
A: Selecting	4.23	4.00	3.59	3.59	3.95
	(0.73)	(0.79)	(0.67)	(0.73)	(0.90)
B: Rating emotion or belief	3.93	3.93	3.62	3.59	3.73
	(0.98)	(0.83)	(0.81)	(0.73)	(1.12)
C: Exploring the focus of discussion	4.23	4.03	3.77	3.68	4.09
	(0.90)	(0.96)	(0.87)	(0.78)	(0.97)
D: Developing an alternative interpretation	4.32	4.00	3.86	3.64	4.18
	(0.61)	(0.72)	(0.83)	(0.85)	(0.85)
E: Designing an empirical test	4.33	4.03	3.90	4.05	4.27
	(0.88)	(0.93)	(0.97)	(0.84)	(0.94)
F: Evaluating an empirical test	4.43	4.07	4.05	3.86	4.09
	(0.63)	(1.03)	(0.72)	(0.89)	(1.02)
G: Collaboration	4.41	4.10	3.95	3.75	4.14
	(0.73)	(0.92)	(0.79)	(0.85)	(0.89)

As seen in Table 11, scores for Q1, Q2, and Q5 were uniformly high. Scores for Q3 (clarity of the items) and Q4 (clarity of anchor points) were 5 – 7 % lower on average.

Comparison of researcher and practitioner feedback.

An independent samples *t* test was conducted to compare the overall mean scores of researchers (n = 22) and practitioners (n = 8). Neither the Shapiro-Wilk statistic nor Levene's test were significant, indicating that the assumptions of normality and homogeneity of variance were supported. The *t* test was not statistically significant, indicating no difference between groups, $t(28) = -1.061$, $p = .289$. This result indicates that the overall ratings across all feedback questions were consistent for both researcher and practitioner experts.

Figure 3 presents a comparison of the mean scores of researchers and practitioners for individual feedback questions (Q1 – Q2), Figure 4 presents a comparison of the mean scores of researchers for individual feedback questions (Q1 – Q5), and Figure 5 presents a comparison of the mean scores of researchers and practitioners for all scale items (A – G). Error bars represent 95% confidence intervals.

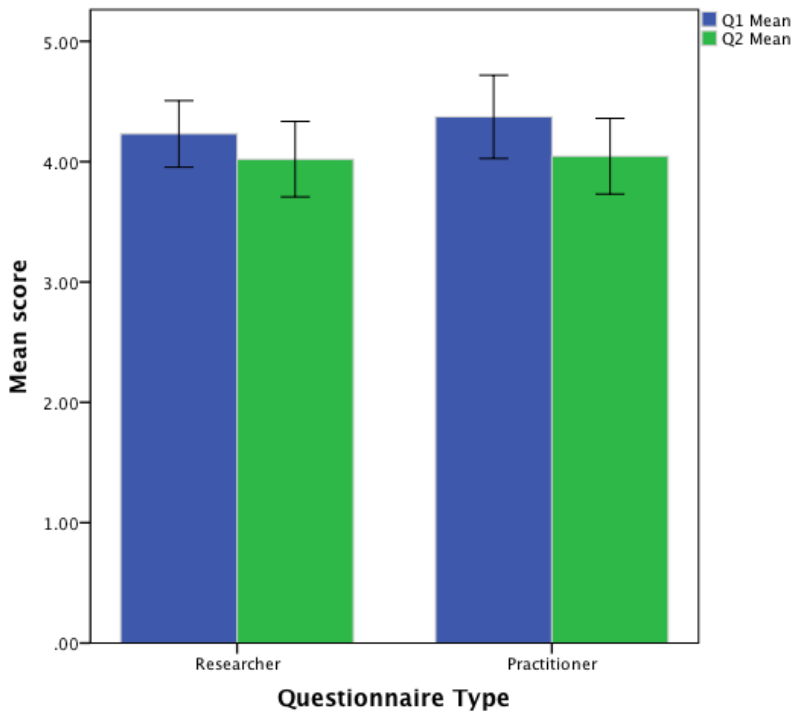


Figure 3. Comparison of the mean scores of researchers and practitioners for feedback questions (Q1 – Q2). Error bars represent 95% confidence intervals.

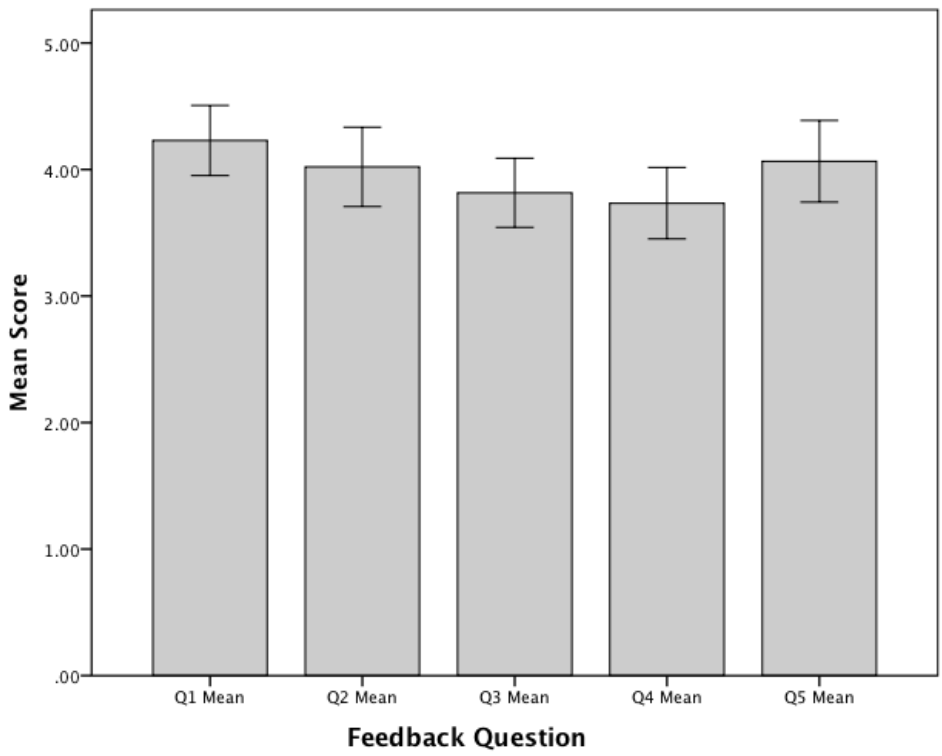


Figure 4. Comparison of the mean scores of researchers for feedback questions (Q1 – Q5). Error bars represent 95% confidence intervals.

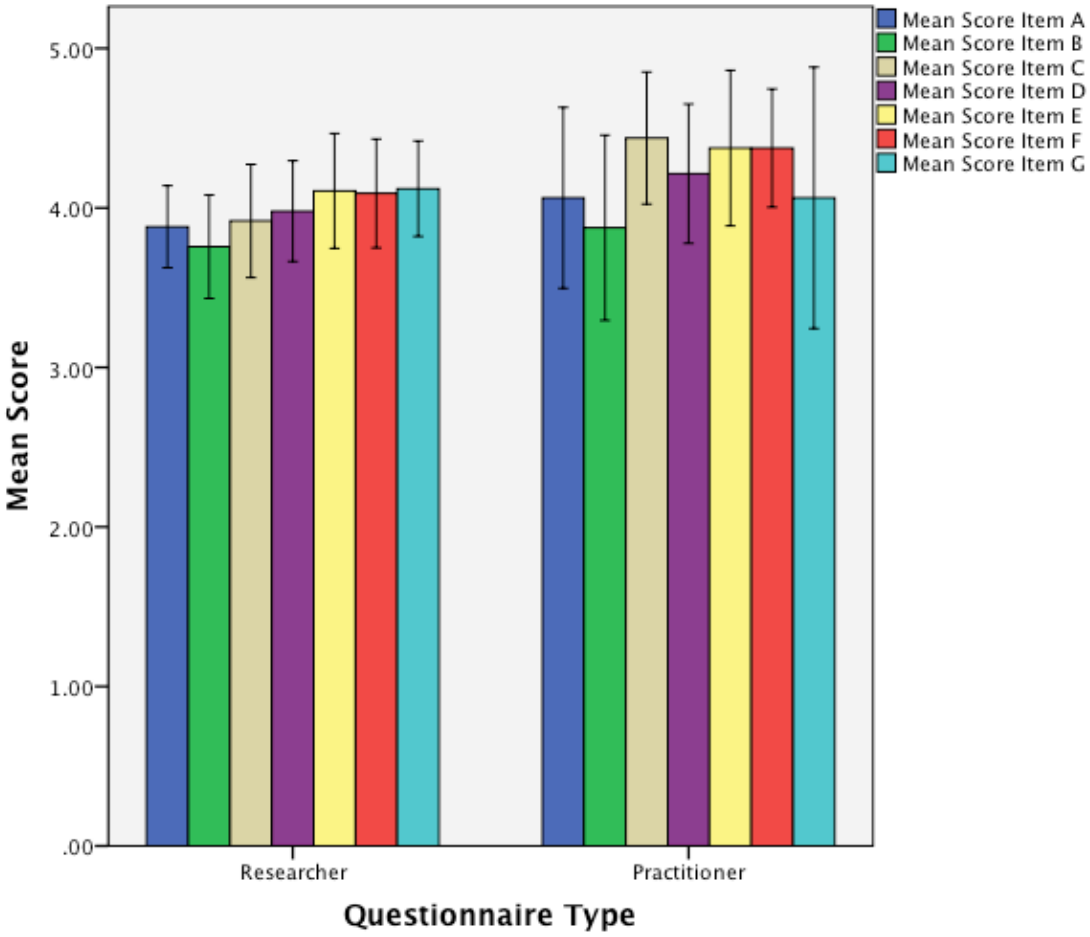


Figure 5. Comparison of the mean scores of researchers and practitioners for scale Items A – G. Error bars represent 95% confidence intervals.

As seen in Figures 3, 4, and 5, all 95% confidence intervals overlap in all figures, indicating no statistically significant differences between researchers' and practitioners' scores on any feedback question or for any scale item. The high degree of consistency between researchers' and practitioners' ratings indicates that both groups rated the scale equally. In particular, both groups rated content validity equally and both found that the scale reflected the use of CE in their professional practice (ecological validity).

Results for individual feedback questions.

Content validity.

This question sought feedback on the degree to which each item accurately described an element of CE. Table 12 presents the number of experts (researchers and practitioners combined) who endorsed each level of content validity for each scale item.

Table 12

Number of Experts Endorsing Each Level of Content Validity for Each Scale Item

		Expert ratings of content validity				
Item	Aspect of CE	Very Small	Small	Somewhat	Large	Very large
A	Selecting	-	1	2	16	11
B	Rating belief or emotion	1	1	6	13	9
C	Exploring	1	-	3	13	13
D	Alternative Explanation	-	-	2	15	11
E	Developing an experiment	1	-	2	12	15
F	Evaluating an experiment	-	-	2	13	15
G	Collaboration	-	-	4	9	16

Note. N = 30.

Data from Table 12 were used to calculate the Content Validity Index. As described above, the CVI was calculated as the proportion of experts who rated the scale item as an accurate description of an aspect of CE to a large or very large extent. The CVI for each scale item and the scale overall is presented in Table 13.

Table 13

Content Validity Index (CVI) for Each Scale Item (A – G) and for the Scale Overall

Scale Item	Aspect of CE	CVI
A	Selecting	.90
B	Rating belief or emotion	.73
C	Exploring	.87
D	Alternative Explanation	.93
E	Developing an experiment	.90
F	Evaluating an experiment	.93
G	Collaboration	.86
Scale Overall	N/A	.87

Note. CVI required for content validity = .78

Table 13 shows that the CVI cut-off for content validity was met by a strong margin for the scale overall and all scale items except Item B. This result demonstrates a very high level of endorsement of content validity by these 30 experts. This is particularly strong given that the CVI is calculated here to a far higher standard using the top two anchor points from a 5-point Likert scale, rather than the usual 4-point Likert scale, and using agreement between 30 experts, rather than the required 10 for statistical significance. Item B, rated at .73, narrowly missed the

required .78 cut-off. From Table 12, it can be seen that the main difference between Item B and the other items is that experts made six ratings of *Somewhat* for Item B compared to a mean of 2.2 *Somewhat* ratings for the other items. This indicates that four experts were more ambivalent about Item B, rather than rating it low or very low on content validity. This result is consistent with comments made regarding Item B, in which a small number of experts were unsure about the role of rating emotion or strength of belief in CBT. This issue will be discussed further in the qualitative results section below.

Ability to capture variability in CE.

Experts strongly endorsed the ability of the scale to capture variability in CE (M = 4.03, or 80.6%, SD = 0.63). Table 14 shows the number of experts (researchers and practitioners combined) who endorsed each level of 'ability to capture variability' for each scale item.

Table 14

Number of Experts Endorsing Each Level of Ability to Capture Variability for Each Scale Item

		Expert ratings of ability to capture variability				
Item	Aspect of CE	Very Small	Small	Somewhat	Large	Very large
A	Selecting	-	1	6	15	8
B	Rating belief or emotion	-	1	8	13	8
C	Exploring	1	1	4	14	10
D	Alternative Explanation	-	1	4	17	6
E	Developing an experiment	-	2	6	11	11
F	Evaluating an experiment	1	2	2	13	11
G	Collaboration	-	2	5	11	12

Note. $N = 30$.

The pattern of results in Table 14 indicates that, while the mean score on this question was 80.6%, 77.3% of respondents rated the ability of each scale item to capture variability to either a large or very large extent, indicating strong endorsement, but just short of the required 78% required by the CVI.

Clarity of the scale items.

Question Three focused on the clarity of the items. Scores on this question were lower than for Q1 and Q2; however, experts still strongly endorsed the items' clarity overall (*M* = 3.82, or 76.4%, *SD* = 0.62). Table 15 shows the number of expert researchers who endorsed each level of clarity for each item.

Table 15

Number of Experts Endorsing Each Level of Item Clarity for Each Scale Item

		Expert ratings of item clarity				
Item	Aspect of CE	Very Small	Small	Somewhat	Large	Very large
A	Selecting	-	-	11	9	2
B	Rating belief or emotion	-	2	6	11	2
C	Exploring	-	1	8	8	5
D	Alternative Explanation	-	2	3	13	4
E	Developing an experiment	-	2	4	8	6
F	Evaluating an experiment	-	-	5	11	6
G	Collaboration	-	1	4	12	5

Note. N = 22.

Overall, 67.5% of experts rated the clarity of each scale item to either a large or very large extent. The pattern of results evident in Table 15 indicates that, while the mean score on this question was 76.4%, rating was spread mainly between the *Somewhat* and *Large* ranges. These results are consistent with experts' feedback comments, discussed in the qualitative results section below, which were mainly focused on the clarity of the items and anchor points.

Clarity of the item anchor points.

Question four focused on the clarity of the item anchor points. Mean scores on this question were also lower than for Q1 and Q2, but still indicated strong endorsement of anchor point clarity (M = 3.79, or 75.8%, SD = 0.64). Table 16 shows the number of expert researchers who endorsed each level of clarity of the anchor points. Although the mean score on this question was 75.8%, 63.8% of experts rated the ability of each scale item to capture variability to either a large or very large extent. As can be seen from Table 16, experts' ratings were concentrated in the *Somewhat* and *Large* ranges.

Table 16

Number of Experts Endorsing Each Level of Anchor Point Clarity for Each Scale Item

Item	Aspect of CE	Expert ratings of anchor point clarity				
		Very Small	Small	Somewhat	Large	Very large
A	Selecting	-	2	6	13	1
B	Rating belief or emotion	-	1	9	10	2
C	Exploring	-	1	8	10	3
D	Alternative Explanation	-	2	7	10	3
E	Developing an experiment	-	1	4	10	7
F	Evaluating an experiment	-	2	4	11	5
G	Collaboration	-	1	7	8	4

Note. N = 22.

Ecological validity.

Question five sought feedback on the degree to which each scale item reflected an aspect of CE that respondents routinely used in their professional practice. This was aimed at checking the ecological validity of each item. Respondents strongly endorsed each scale item as describing an aspect of CE they routinely used in their professional practice (*M* = 4.07, or 81.2%, *SD* = 0.73). Table

17 shows the number of experts endorsing each level of ecological validity for each scale item.

Table 17

Number of Experts Endorsing Each Level of Ecological Validity for Each Scale Item

		Expert ratings of ecological validity				
Item	Aspect of CE	Very Small	Small	Somewhat	Large	Very large
A	Selecting	-	2	3	11	6
B	Rating belief or emotion	1	2	5	8	6
C	Exploring	-	2	3	8	9
D	Alternative Explanation	-	1	3	9	9
E	Developing an experiment	1		1	10	10
F	Evaluating an experiment	1	1	1	11	8
G	Collaboration	-	1	4	8	9

Note. N = 22.

As can be seen from Table 17, 79.2% of respondents endorsed each scale item as representing an aspect of CE they used in their professional practice to either a large or very large extent.

It is noted here that respondents strongly endorsed designing empirical tests (90.9%) and evaluating empirical tests (86.4%) as elements of their professional practice to a large or very large extent. This is interesting because, as will be discussed in Chapter 8, reporting on Study 3: CE as a predictor of therapeutic outcome, these aspects of CE were very infrequently detected during rating for Study 3.

In summary, mean scores on expert feedback questions were high to very high. The CVI cut-off was met for content validity and ecological validity, despite the higher standard of CVI used here compared with that recommended in the literature. This demonstrates a very high level of expert endorsement for the content validity and ecological validity of the CES. Ability to capture variance (.773) nearly met the CVI cut-off of .78. Mean scores for ratings of clarity of items and anchor points were still high (M = .76, SD = 0.63), but did not meet the CVI marker. Experts focused on these clarity issues in their feedback comments, discussed next.

Qualitative Results

In addition to rating the CES items, experts were invited to comment on each item and on the scale as a whole. Experts' comments are presented here in two levels. Comments pertaining to the scale overall or to more than one item are presented first, followed by comments focused on specific items only. After each comment or group of comments, responses to comments and changes made to the scale as a result are discussed.

For clarity, scale items are referred to by name (e.g., 'Item A'). Anchor points of the items are referred to by item letter combined with the anchor point number (e.g., Item A anchor point 4 is referred to as 'A4'). The full text of the questionnaire

version of the CES is presented as part of the Expert Feedback Questionnaire in Appendix C. The CES final version is presented in Appendix A.

Respondents made 104 comments in total. At least one comment was made by 23 of the 30 respondents (76.7%). The number of comments per respondent was normally distributed (*M* = 3.35, *SD* = 2.7). Researchers were 74.2% of the sample and made 76% of the comments, indicating an even distribution of comments between respondent groups. A breakdown of the number of comments for each scale item is presented in Table 18.

Table 18

Number of Comments for Each Scale Item

Scale Item	Number of Comments Total | Researchers | Practitioners
A - Selecting a focus of discussion to explore	16 | 12 | 4
B - Rating the degree of emotion or belief	15 | 12 | 3
C - Exploring the focus of discussion	11 | 9 | 2
D - Developing an alternative interpretation	14 | 10 | 4
E - Designing an empirical test	11 | 9 | 2
F - Evaluating an empirical test	11 | 9 | 2
G - Collaboration	15 | 11 | 4
‘Other Comments’	11 | 9 | 4

Level 1: Comments pertaining to the scale as a whole, or to multiple items.

Positive feedback.

Participants commented that the scale was "well considered", "well written", and "took a useful approach" to measuring the construct. Positive feedback also related to the focus on "specificity and thoroughness" in doing and rating CE, the "very thorough descriptions of key elements of CT and collaborative empiricism", and the helpfulness of examples in clarifying concepts and anchor points. Respondents commented that they found the scale a "really good approach to tapping this dimension of CBT"; in which "generally the questions seem to capture the construct of collaborative empiricism well with good face validity"; and that "once it is fully developed and tested further, (the scale) will be an excellent addition to the literature and to further research".

Difficulty distinguishing between anchor points three and four.

The most frequent critical comment related to difficulty distinguishing between anchor points three and four on the empirical items (A – F). Five respondents commented on this issue, and generally repeated these comments for each empirical item. This repetition makes sense, as the underlying rating scheme is the same across the empirical items. Two respondents commented that the phrasing "SOME PROBLEMS" (in anchor three) and "MINOR PROBLEMS" (in anchor four) "were potentially difficult to discriminate", and "may be difficult for raters to distinguish in practice". One respondent commented that the "anchors were clear" but the "distinctions in some of the middle items seem small". In a related comment, one respondent commented that the phrase "The client's EXPERIENCE WAS THE BASIS FOR…", which appeared in anchor points three and four, was confusing.

The above comments were addressed by clarifying the distinction between anchor points three and four. In each scale item, anchor point three was changed from "SOME PROBLEMS" to "PROBLEMS WITH EMPIRICISM that REDUCED the empiricism". This change aimed to highlight the qualitative difference between anchor points three and four, that is, whether problems related to empiricism or not. The aim was also to remove the potential for confusion regarding the quantitative distinction between 'some' and 'minor' problems by removing the term 'some'.

In addition, the phrase "The client's EXPERIENCE WAS THE BASIS FOR", which one respondent found confusing, was replaced wherever it occurred with more specific wording that detailed the level of use made of the client's experience that would be required to endorse that anchor point. For example, in Item C, anchor C2, the phrase "a thought WAS DISCUSSED, but WAS NOT THE BASIS FOR EXPLORING" was replaced with "The CLIENT'S EXPERIENCE was MENTIONED but was NOT DISCUSSED OR FOCUSED ON". Similarly, the phrase "The client's EXPERIENCE WAS THE BASIS FOR" was replaced in C3 with "The CLIENT'S EXPERIENCE was EXPLICITLY USED to" and in C4 with "The CLIENT'S EXPERIENCE was EXPLICITLY USED in a SPECIFIC WAY".

Examples illustrating anchor points.

Five respondents commented on the examples illustrating the anchor points. Some respondents commented that the examples were "clear" and "made the items easier to understand". Another reported that "the examples were a bit confusing" and another that they "suggested verbally fluent clients in white collar employment". In response to these comments, each example was reviewed and several were edited to increase clarity. For example, in anchor D4, the text "some concern of the client's about the test remained unaddressed; solutions were developed for anticipated

problems but the client remained only partially confident)" was replaced by "the client expressed some doubt or reservation about the alternative interpretation, which was not explored; failing to offer or request feedback". These changes replaced descriptions of inner states of the client (e.g., 'partially confident') with examples of directly observable behaviour for raters to identify (e.g., 'expressed some doubt or reservation…, which was not explored'). An example involving a client who anticipates difficulty engaging socially at a party ('Some people will talk to me at the party.' vs. 'At least 3 people will talk to me after I say hello.') was not changed as it was thought that, although it might reflect a verbally fluent demographic, it did not unbalance the scale overall.

Seven respondents suggested that adding more examples "may be helpful" to "increase the clarity" of the items or anchor points, "even if it is somewhat repetitive". Two of these comments focused on the scale in general and five related to specific items or anchor points: A4 (2 comments), B5 (2 comments), and Item C (1 comment). These remarks were counterposed by other comments that the items were "too wordy" and "dense" and that the examples were helpful, but risked overburdening the rater. During initial pre-questionnaire testing, raters had reported anecdotally that having examples for every anchor point was too much information to hold in mind during rating. To address and reconcile these comments, the general principle was adopted to balance the increased clarity provided by more examples, with the increased amount of information that raters would need to hold in mind when rating the scale. In practice, this resulted in two or three examples for each scale item, focused on the more conceptually dense middle anchor points, rather than at every anchor point.

Clarity of key concepts.

One respondent commented that the meaning of 'experience' was difficult to discriminate from 'interpretation' or from 'a specific event in the person's life'. This comment was addressed by expanding the definition of experience in the 'Key Constructs' section to include clients ongoing 'lived' experience of their emotions, behaviour, cognitions, and cognitive processes. The term 'experience' was further distinguished from 'interpretation' by adding detail to relevant examples in the anchor points. For example, anchor A3, which read: "(e.g., failing to distinguish experience from interpretations: 'My boss ignores me in the corridor because he thinks my work is no good')" was augmented with terms further describing the level of experience needed to endorse this anchor point: "experience was vague, general, ambiguous, misinterpreted, or was sourced from the therapist, rather than from the client."

Clarification of wording.

One respondent commented that "the interaction of cognition and experience are central, although sometimes difficult to separate clearly". It was unclear if this comment related to a specific issue. It was addressed by providing additional guidance to raters during training in the form of a diagram strengthening the distinction between the client's experience and the current focus of discussion, and highlighting that a core aspect of CE is the interaction of clients' cognition and cognitive processes with the identification and the discussion/exploration of that experience. This diagram is presented in Appendix F.

Complexity of the scale and reliability of ratings.

Three respondents commented that the scale "seemed valid" and "accurate" but expressed concern that it "might be difficult to rate reliably" due to complexity or level of detail. This concern was echoed in a related comment that the scale was likely

to be useful in research and training, but might be too difficult for "your average trainee" to use without training. The complexity of the scale reflects a design decision to attempt, for the first time, to accurately capture CE in detail within a session. As such, the impact of complexity on the reliability of rating was continuously assessed but the scale was not edited in response to these comments. The concern was, however, noted for careful review once the reliability of rating had been assessed during pilot testing.

Number of conceptual elements varying between anchor points.

One respondent commented that Item A "tries to capture variability on a range of items from explicitness, degree of prioritization of most important, resolution of problems etc.". The respondent suggested that it might be easier to provide examples for the item if the introduction to the item "provided more detail about the sub-elements of empiricism measured". This comment raises the design issue of the number of concepts varying between each anchor point on an item, and role and number of examples accompanying the anchor points. This point is discussed here with respect to Item A but applies equally to all empirical items on the scale.

The basic structure of the empirical items is that each item varies on a single primary concept, the use of the client's experience. Specific anchor points then have examples designed to clarify and further distinguish the level of empiricism matching that anchor point. It is not the case that every possible manifestation of CE at each of the five levels in an item is spelled out for the rater. In Item A, for example, the only concept that varies between anchor points A1, A2, and A3 is the use of the client's experience, which ranges from *not identified*, to *mentioned*, to *explicitly used*. Anchor point A4 includes examples to illustrate potential 'minor problems' that would reduce the quality of empiricism, such as 'failing to check for other clinically important

beliefs' and 'variability of focus on the task'. The intention of these examples is to help the rater identify the level of problems with empiricism required to endorse this anchor point, and to help distinguish anchor point A4 from A3.

This structure reflects a design decision to allow the scale to rate empiricism in multiple contexts, at multiple levels, within a therapy session. An alternative approach might be to have a larger number of empirical items, or have multiple sub-elements for the existing items, each of which would vary on a single sub-element of empiricism, for example, degree of importance, explicitness, etc. One difficulty with this approach, however, is that it risks overburdening the rater with detail related to multiple aspects of empiricism varying in a single item. Another difficulty is that specifying components of empiricism in particular contexts may reduce the range of contexts in which empiricism can be rated. As the CES is the first measure focused on CE, it was decided that maintaining the flexibility of the measure, by not listing specific sub-elements of CE in specific contexts, would better enable the CES to capture CE in multiple contexts, at multiple levels in a session, including in unanticipated contexts where empiricism may be used in therapy. Accordingly, it was decided to vary each item on a single primary concept, 'the use of the client's experience', but augment this with a small number of examples in the anchor points to aid the rater in differentiating the level of empiricism matching each anchor point.

The response to this comment is linked with the response discussed above to the issue of clarifying the distinction between anchor points three and four. Anchor points on each scale item were edited to highlight that the main concept that varies between anchor points is the 'use of the client's experience'. This involved replacing references to 'empiricism as the basis of' in each anchor point with graded descriptions of the levels of 'use of the client's experience'. Table 19 presents a

comparison of the anchor points in Item A between the questionnaire and rating versions of the scale.

Table 19

A Comparison between Questionnaire and Final Versions of the CES of the 'Use of the Client's Experience' in Item A

	Scale version	
Anchor	Questionnaire	Final
A1	NO REFERENCE	NOT IDENTIFIED
A2	PASSING REFERENCE, but EXPERIENCE WAS NOT THE BASIS for selecting	MENTIONED but was NOT USED
A3	WAS THE BASIS for … SOME PROBLEMS which reduced the quality of the empirical data	EXPLICITLY USED… but PROBLEMS WITH EMPIRICISM that REDUCED the empiricism
A4	WAS THE BASIS for… MINOR PROBLEMS, which did not reduce the use of experience	EXPLICITLY USED in a SPECIFIC WAY, but MINOR PROBLEMS, which DID NOT REDUCE the empiricism
A5	WAS THE BASIS for … NO PROBLEMS, or any difficulties WERE RESOLVED	NO PROBLEMS, or any difficulties WERE RESOLVED

Level 2: Comments pertaining to particular scale items only.

Item A. Selecting a focus of discussion to explore.

One respondent suggested that anchor A1 in Item A (Selecting) might include "thought chosen from an pre-populated list" as an example of a poor use of the client's experience in selecting a focus of discussion. The respondent noted that he or she "sees this frequently in bad CBT". This is an excellent suggestion for an anchor point, and was used as an example during rater training. It was not, however, added to the final version of the scale due to concerns of overloading raters with too much information to process during rating.

A similar issue was raised by another respondent who commented that "it was somewhat difficult to answer this question (Q1) without a clinical example to work from". Clinical examples were used in rater training, but detailed clinical examples were omitted from the scale to avoid overwhelming the rater with too much detail.

Item B. Rating the degree of emotion or belief.

One respondent asked, regarding the relationship between Item A and Item B, whether "if no ratings are sought, but the thought is identified for evaluation, then that would be an example of Item A, is that correct?" This is correct, and demonstrates that the respondent had understood the intent of the item; however, to ensure item clarity, anchor B1 was edited to read: 'whether a rating was sought, but not completed', and B2 to read: 'a rating was made but not used to quantify the degree of belief or emotion'. This meant that if no rating of belief or emotion was sought, Item A would be rated, but if an attempt was made to rate belief or emotion, Item B would be rated. The respondent's question was used during rater training to test raters' understanding of the relationship between Item A and B, for example, "If no ratings are sought, what item is rated?"

One respondent remarked that "Item B in my view is not so important for collaborative empiricism, but rather has to do with the scaling of conviction in beliefs". Item B does indeed focus on the scaling of conviction (and strength of emotion) associated with beliefs, and so the comment was taken to support the clarity of the item, as the respondent had correctly paraphrased the item's intent. It was interesting to note, however, the respondent's understanding that the scaling of conviction of beliefs is 'not so important' for CE. Rating the strength of belief and emotion is a quintessentially empirical process that is regularly featured in CBT training manuals, including seminal discussions of CE in A. T. Beck et al. (1979) and Beck (1995, 2011) that underpin the CES. The centrality of the rating of beliefs as a core aspect of CE was reiterated recently by Dr. Judith Beck during an expert panel discussion on CE in which she participated at the 7th International Congress of Cognitive Psychotherapy (N. Kazantzis et al., in press).

In a related comment, another respondent remarked that, regarding rating the strength of emotion associated with a 'hot thought', "I know what to do on a tape for credentialing, we all do" but many therapists "skip this rating step but will tell you that both they and the client know the centrality of a hot thought and how hot that thought is, but not using numbers." This comment suggested that, despite the recommendation in CBT training manuals to quantify emotion, therapists might do relatively little of this. The frequency of ratings of emotion or belief in the present sample will be evaluated in the examination of the relationship between CE and therapeutic outcome in Chapter 8.

Finally, one respondent commented, "I might be confused rating this item if the therapist asked which thought made the client most anxious but did not rate it on a

scale." To address this issue, the scale was edited to provide additional guidance in the instructions for rating the item:

> "This item is rated when a rating is made of the degree of emotion or belief associated with the focus of discussion. Do not rate this item when there is only a description of emotion or belief, without an attempt to quantify degree."

Item C. Exploring the focus of discussion.

One respondent commented that Item C focuses on both the evaluation of the validity of a client's beliefs and also on the functionality of cognitions and cognitive content. This is correct and in fact true of each item on the scale. This respondent commented that the focus on functionality "sounds like a revised understanding" and does not relate to the greater focus on validity that was "characteristic of earlier cognitive therapy". The respondent further remarked that functionality had been "always available" as an element "but not always specifically discussed".

Although no changes were made to the scale in response to this comment, it raises an important point that relates to the conception of CE underlying the scale. Seminal discussions of CBT such as A. T. Beck et al. (1979) and J. S. Beck (1995), go further than simply evaluating beliefs for validity, to consideration of the functionality and helpfulness of cognitions and cognitive processes for the client. This inclusion of functionality and helpfulness as a dimension of evaluation was also advocated by Professor Dobson during the development of the scale (personal communication, October 6, 2011) and Dr Jim Overholser (personal communication, July 26, 2011). Accordingly, the CES was developed based on a conception of CBT in which the functionality and helpfulness of the client's beliefs is considered in tandem with validity.

Another respondent suggested that this item might additionally capture whether the questioning was “in depth and broad enough to bring in a range of relevant data”. The concepts of depth and breadth are captured to some extent in the negative in anchor point C3, “experience was vague, ambiguous, (and) misinterpreted”. This issue relates to the design decision, discussed above, to vary each item on the single main category of the use of the client’s experience, and to include relevant aspects of CE in the anchor point descriptions to aid the rater in distinguishing levels of empiricism. In addition, various other descriptors for these concepts were trialled in pre-testing, with a view to capturing the degree of depth and breadth of exploration of the focus of discussion; however, raters found these concepts difficult to identify reliably during pre-testing.

Item D. Developing an alternative interpretation.

Five experts queried the ability of Item D to measure collaboration. This is noteworthy because Item D was designed to measure empiricism, not collaboration. Experts commented, for example, that the item did not “emphasize reflective questions” or “Socratic dialogue” and might “address further the collaborative nature of developing alternatives”.

One explanation for experts evaluating this item in terms of its ability to measure collaboration rather than empiricism might be that, as they worked through the feedback questions in order, they may not have anticipated that collaboration is specifically measured in the last item, Item G, and so taken the opportunity to point out that it could also be usefully employed, and measured, in developing alternative explanations. Another, perhaps more likely, explanation is that experts were sensitized to the quality of collaboration because the examples used to illustrate empiricism on anchor points D2 and D3 were low on collaboration. For example,

anchor D2 in the questionnaire version read, 'Therapist: "Has that happened every single time?" Client: "No, not every time I guess." Therapist: "Okay, so it's not inevitable then?'. Low quality collaboration (low because the therapist is 'doing the work') was employed deliberately in these examples to illustrate that empiricism is evaluated independently and irrespective of the quality of collaboration. The number of comments regarding poor collaboration, however, suggests that this example was confusing. Several steps were taken to address this issue. The independence of the empiricism and collaboration items was highlighted to raters during training and practice was given in rating the empirical items independently of collaboration. In addition, the example in anchor D2 was edited to be more neutral in terms of collaboration, with the therapist refraining from drawing a conclusion (Therapist: "How else could you think about this situation?" Client: "I guess it's not always so bad." Therapist: "Ok"). Finally, to balance the level of collaboration in the examples, a new example was added to anchor D4 demonstrating better collaboration:

> (Therapist: "So, you didn't fail?" Client: "It wasn't very good, but I did manage to do it." Therapist: "Let's look a bit more at how you managed to do it." Client: "I guess I stuck at it, even though I felt it was pretty bad. I did do that better than before." Therapist: "What might that mean for the idea 'you're a failure'?)

Item E. Designing an empirical test.

Three respondents commented positively that this item "very clearly" differentiated levels in the anchors points. One respondent speculated this was "perhaps because inevitably tied more to behavioural criteria".

Two respondents commented on the choice of concepts varying across anchor points E1 – E5. One respondent commented that the process of designing a test is

difficult to capture "due to the complexity and possible variability" of possible manifestations in therapy. Another suggested that the item might try and capture the additional dimensions of "whether a central belief is tested" and whether the test "really tests the belief."

These comments relate to the purpose of the item and to the difficulty of capturing a complex and variable behaviour such as developing an empirical test of a cognition or cognitive process. These issues are related and will be addressed together. In the above comments, respondents focused on the ability of Item E to measure competence in designing empirical tests. The primary purpose of the item, however, is to evaluate the use of empiricism in designing a test, rather than comprehensively assess the quality of the test itself. Of course, these issues are related, as, for example, a test that does not make good use of the client's experience is unlikely to be a good test. The emphasis in respondents' comments on how the item evaluated the quality of an empirical test per se, suggested that the focus of the item (on empiricism) needed to be clearer.

Capturing empiricism in complex behaviours inevitably involves a trade-off between the number and complexity of scale items and the level of reliability obtainable by raters. This balance was set in the questionnaire version of the scale by focusing Item E on assessing the use of criteria for outcomes of the test, with further discrimination between anchor points included in the descriptions of 'some problems' and 'minor problems'. Criteria for outcomes of the test was chosen because it has been (1) discussed in CBT manuals as a key aspect of a good test (e.g., Beck, 1995, 2011), and (2) it was deemed likely that a good use of empiricism in a test would need clear criteria to evaluate the possible outcomes of the test, and conversely that a test with clear criteria for its possible outcomes would be well-thought out in other ways.

An alternative approach would have been to include additional scale items, each targeting individual aspects of designing an empirical test (e.g., centrality of belief, whether the belief was adequately tested, etc.) It was thought, however, that it would be too difficult for raters to reliably track and rate multiple sub-components simultaneously.

To address the issues raised in these comments, Item E was edited to clarify that the primary concept varying between anchor points was 'the use made of the client's experience'. This was added to each anchor description and highlighted using capital letters. The focus on 'criteria for outcomes of the test' was moved to an explicit but supporting role. These changes brought Item E into line with the other empirical items, which focus explicitly on the use of empiricism. Also, in line with changes to the other empirical items, the term 'SOME PROBLEMS' was replaced with 'PROBLEMS WITH EMPIRICISM that REDUCED the empiricism'.

Item F. Evaluating an empirical test.

Experts responded positively that this was a "very good, very clear, important item", anchor points for this item "seem clearer", and commented that "again this is strong. Clear examples except for anchor point 5, makes it clear what is being looked for".

One respondent commented that "anchor 5 seems unusually aspirational". The list of attributes in anchor F5 were derived directly from the descriptions of a good empirical test in Beck et al. (1974) and J. S. Beck (1995). As such, anchor point A5 was deemed to represent a high but reasonable standard of the use of empiricism in evaluating an empirical test and was not changed in the final scale version.

One respondent noted that "there is maybe an inevitable confounding of this variable with the last one" because poor criteria in developing a test may hinder

evaluation of the test. This comment correctly identifies the potential dependence of Item F (evaluating a test) on Item E (developing a test). These items are not necessarily dependent, however, as it is possible to have developed a poorly designed test but evaluate the results of it well. Accordingly, the item was not edited in response to this comment.

Finally, another respondent commented that "criteria seems emphasized more than other relevant aspects". In related comments, two respondents wondered if the item might also measure additional aspects such as whether the therapist dealt with negative events that happened in the context of designing the test or whether the "behavioural experiment affected experience or belief (change)".

These comments relate to the issue discussed above in relation to Item E regarding the primary focus of the empirical items on the use of client's experience. Behaviours such as 'dealing with negative events' may be relevant to the evaluation of the quality of an empirical test; however, the primary purpose of the empirical items is to evaluate the use of empiricism, rather than competence in CBT more generally. This is achieved in Item F primarily by varying across anchor points 1 – 5 on the quality of the use of the client's experience. As discussed above, however, it is not possible to completely separate competence in empiricism from competence in CBT. This is because empiricism is affected by factors, such as core counselling skills, that affect the quality of therapy.

In the CES, the solution chosen for this issue is to have the empirical items vary on a single primary construct of 'the use of the client's experience', but incorporate the effect that issues with general competence in CBT will have on the quality of that empiricism by including examples in the anchor points. For example, the impact on empiricism of issues with general counselling competencies is included

in anchor point 4 (‘minor problems’) and, by the absence of problems, in anchor point 5. Anchor point 4 describes minor problems that do not directly impact on the use of the client’s experience, but which may affect the overall quality of the therapeutic work and consequently the quality of empiricism. As with item E above, these comments indicated that the focus of the items on empiricism needed to be made clearer. Accordingly, the same solution was implemented as per Item E. Item F was edited to clarify that the primary concept varying between anchor points is ‘the use made of the client’s experience’. This was highlighted in capital letters in each anchor description, and the focus on ‘criteria for outcomes of the test’ was moved to an explicit but supporting role.

The notion that core counselling skills underpin specific CBT competencies has been articulated in some detail in the UK CBT Competencies Framework (Roth & Pilling, 2008). The Framework also identifies attributes of the relationship in CBT. Elements of collaboration in CE are implicit in the framework, such as ‘sharing responsibility for session structure and content’ and the ‘ability to monitor therapy and guide outcome’. Elements of empiricism are also inherent in some competencies listed, such as the ‘ability to help the client reality test automatic thoughts and images’. Despite their inclusion, these aspects of CE remain partial, implicit, and fragmented in the Framework. Collaborative empiricism is not discussed and its core components are not articulated or related to specific competencies or interventions.

Item G. Collaboration.

Respondents commented positively that “this item reflects collaboration nicely and scales it well” and, that the terms ‘Balance and Imbalance’ are “used well” in the item to describe the therapeutic relationship, and in accordance with previous use in the literature.

Three respondents raised concerns that the quality of collaboration between client and therapist should be assessed relative to the stage of therapy. For example, respondents commented that balance may be considered "to mean equal share" but may fluctuate "depending on which stage of therapy is being considered". One respondent noted that "at first as the client is learning, the therapist should probably be more active" but that by the end of treatment, the client "should be able to go through the process with minimal help. A 50/50 balance is not always optimal."

The idea that CE may change across the course of therapy was a core concept underpinning the design of the scale; however, the above comments indicated that this was not sufficiently clear in the questionnaire version of the scale. To address this, the introduction to Item G was edited to specifically incorporate the above points:

> "Ideal balance means the client and therapist share the work to the extent that they are able given the stage of therapy. For example, early in therapy, the client may contribute less to the design of an empirical test, as this is a new activity for the client, but the client can still fully share the work by making contributions, involvement in shared decisions, and mutually responsive interactions."

In a related point, one respondent noted that, in order to maintain the therapeutic relationship, therapists sometimes allow clients room to "dominate the discussion (at least temporarily)". The respondent suggested that while therapists can always try to increase the client's involvement, clients cannot always moderate their behaviour and so it would be "unfortunate to indicate a low score in those situations." This respondent noted, however, that the key point here is whether the purpose of the scale is to rate therapist competence in CE or to rate "the existence of the CONDITION of CE at a given moment or in a given session, then my concern is not

as relevant" (emphasis in original). The purpose of the scale is to rate the quality of CE present in the session. This was made clearer in the rating version of the scale by including explicit instructions to raters to rate the quality of CE actually present in each rated interaction. In addition, to make it easier for raters to identify collaborative behaviours in a therapy session, three key components of collaboration in CE were listed in the instructions of Item G: "... Sharing the work involves shared decisions, mutually responsive interactions, and contributions from both parties (engagement, participation, effort)."

Focusing on the amount of CE actually manifested in the session, rather than CE 'attempted' or 'fostered', also addresses a comment made by another respondent, "What about when the client leads?" Whether the client or therapist leads the rater will still rate the quality of CE attained during any particular interaction.

Discussion

Standards of Expert Consensus

This study used a methodology consistent with the Delphi method and incorporating the Content Validity Index to conduct a post-development evaluation of the CES. Assessing the content validity and ecological validity of the scale was particularly important for a measure that may be used in clinical contexts and in the training and evaluation of therapists. For both the Delphi method and the CVI, the standards of expert consensus were set higher here than those normally adopted in the literature. Delphi studies have typically included 15-20 experts (Boulkedid et al., 2011; Ludwig, 1997), whereas the present research succeeded in attracting the participation of 30 experts. Similarly, for 10 experts, the CVI requires at least 78% of ratings to be 3 or higher on a 4-point Likert scale, to ensure statistically significant consensus on content validity (Lynn, 1986). This standard was significantly exceeded

here by requiring 78% of 30 experts, rather than the statistically required 10, to agree on rating the CES items at least 4 on a 5-point scale, rather than 3 on a 4-point scale. By setting a very high level of agreement to define expert consensus, these standards provided a high degree of confidence in experts' feedback.

Expert Feedback

Experts' ratings on all feedback questions were high, with low variability between scores, demonstrating consistently high ratings across all five areas assessed. In addition, there was close agreement between researchers and practitioners across all ratings, suggesting that the CES represents a theoretical understanding and a professional practice of CE that was shared by both groups.

Experts' ratings for content validity and ecological validity strongly exceeded both the Delphi and CVI standards. The Delphi standard was also exceeded for the ability of CES items to capture variance, and narrowly missed on the CVI (77.3%, with 78% needed). These results indicate a strong endorsement of the content validity and ecological validity of the CES, and a solid endorsement of the ability to capture variance.

Experts also rated the clarity of items and anchors points highly (M = 3.82 = 76.4% and M = 3.73 = 74.6%, respectively). These results indicate solid support for the clarity of the CES, but are below the 80% cut-off required to meet the Delphi standard. CVI scores of 67.5% and 63.8% were obtained for the clarity of the items and anchor points respectively, reflecting the lower range of scores making up the 74.6 – 76.4% mean scores on the feedback questions. In the Delphi method, items that do not meet the Delphi cut-off of 80% may be sent for re-examination by the group if expert consensus is between 70 – 79% (A. P. Morrison & Barratt, 2010). In the present context of a post-development evaluation study, items were not sent for re-

examination by experts. Instead, experts' comments, which were focused on the clarity of the items and anchor points, were used to revise items with mean scores between 70-79%.

Experts made 104 constructive critical comments, generally clustered around a small group of themes shared by several respondents. That 30 experts identified similar issues indicates that the issues identified were central and important. Experts were asked to provide feedback comments immediately after rating each item. It seems reasonable to assume that, for example, having just rated the validity of an item, experts would be likely to comment on any issues they had with validity. Instead, there were no globally negative evaluations or concerns expressed regarding the structure or validity of the scale. Experts focused on specific issues, primarily the number and placement of examples, difficulty distinguishing between anchor points, and clarity of expression. Comments on the overall design of the scale were all positive. Experts were also asked for feedback comments on the scale as a whole. They took this opportunity to make comments, but again focused on issues of clarity, further suggesting they were in general agreement with the design of the scale. The close consistency between experts' quantitative ratings and their feedback comments suggests that experts were being critical and careful in their responses, rather than responding randomly or withholding criticism. Taken together, these results suggest that experts did not have issues with the validity or structure of the scale.

In summary, experts were willing and able to give critical feedback on the scale, but were focused on specific issues, largely regarding clarity, rather than issues of validity or scale design. The pattern of responding suggests that if experts had concerns about the validity of the scale, they would have expressed them. The lack of

such comments indicates a strong endorsement of scale validity and structure, and is highly consistent with the pattern of experts' responses to the feedback questions.

Addressing Experts' Feedback

The expert feedback obtained provided sufficient basis for revision of the scale. Every comment was carefully reviewed. Adjustments were made to the scale in response to all comments except three. In these three cases the issue was flagged for further evaluation during pilot testing. No significant revisions to the scale structure were required to address experts' comments. The lack of any significant revisions required further supports the endorsement by experts of the overall scale structure. It is argued here that the detailed responses made to expert feedback improved the scale from the questionnaire version in preparation for pilot testing. This is likely to have raised the mean scores for the clarity of items and anchor points from 74.6 – 76.4 to over the 80% required by the Delphi standard, although it is acknowledged that, due to time constraints, this was not assessed via another round of expert feedback.

Implications for Previous Research

Support for theory.

As is the case with many aspects of CBT process, there is no previous empirical research on CE. The present research is the first to develop a measure of the construct. Despite this, the strong endorsement provided by experts for the content validity and ecological validity of the CES provides support for the theoretical work underpinning the design of the scale.

Limitations

In assessing the content validity of the CES, an alternative approach would have been to specifically ask experts to comment on aspects of the construct they felt were missing from the scale. The trade-off with this approach is the length and

complexity of the feedback questionnaire. In order to keep the questionnaire to a reasonable length, and thereby maximize the number of experts providing feedback, experts were not asked in a separate question about missing aspects of the CE construct.

While acknowledging this limitation, it is argued here that it is unlikely that experts' feedback was affected, for two reasons. Firstly, CE is not a new concept in CBT. In designing the CES, sufficient guidance was available in the literature from the past 30 years to specify the CE construct. This guidance was grounded in close analyses of CE in two seminal works on CBT process, A. T. Beck et al. (1979)and J. S. Beck (1995), both widely recognized as authoritative and which themselves span 20 years of the literature. This information was communicated to experts in the introduction to the scale and in the supporting theoretical paper (included with the feedback questionnaire). It seems reasonable to assume that expert CBT researchers and practitioners therefore had a ready grasp of the construct, sufficient to judge whether aspects were missing. Secondly, experts had numerous opportunities to comment on missing aspects if they chose. Experts were asked after rating each scale item to comment on the item, then on the scale overall, and given a further opportunity to make 'any other comments' at the end of the questionnaire. It would have been possible for experts to comment at any of these points on aspects of CE they felt were missing. Experts made 104 comments on the scale, indicating that they were quite willing to provide feedback; however, none of these comments suggested an aspect of the construct was under-represented or missing.

As acknowledged above, a potential limitation was the lack of a second round of expert feedback on the revised items. A second round is common in the Delphi method but is also regularly omitted (Fish & Busby, 1996; Powell, 2003; S. J. Young

& Jamieson, 2001). It was not possible here due to the time constraints of the research project. A second round would have enabled confirmation that experts' comments had been suitably addressed. Despite this, two points suggest that this was not a significant limitation of the study. First, experts' comments were focused on improving the clarity of items that they had already rated highly for clarity (*M* = 74.6 – 76.4). Experts saw room for improvement, but not that the items were fundamentally flawed or inoperable. Addressing incremental change is likely to be easier to do and to evaluate as having been done, compared with significant revision, particularly to items that have been rated highly for validity and ability to capture variance. Second, experts provided extensive guidance in their comments to address the issues of clarity they raised. It is argued here that the detailed response to 104 feedback comments is likely to have been sufficient to revise the clarity of items from a mean score 74.6 – 76.4 to over the 80 mark required by a second round of expert review.

Future Research

Post-development evaluation of psychotherapy process measures.

Existing process measures have rarely employed a post-development, consensus driven, expert evaluation of the measure during construction. It will be argued here that there are important advantages to the post-development evaluation of process scales that warrant its consideration in future research.

The method utilized with the CES represents a procedure for obtaining post-development expert consensus, grounded in a combination of two established methods, the Delphi method and the CVI. Employing the Delphi method in scale development typically involves experts rating whether attributes of a construct, from a given list, are deemed essential to the construct or not part of the construct. A problem

with applying this method to the development of a measure of psychotherapy process is that the generated list of attributes may differ from the established understanding in the literature. Conceptualization of a therapy process is necessarily relative to the theoretical orientation within which that process is understood; however, experts' may differ in theoretical orientation (Kazantzis, Busch, Ronan, & Merrick, 2007). Even if experts identify their orientation or have published on a specific orientation, their understanding of a psychotherapy process construct may differ from their stated orientation (Cook, Biyanova, Elhai, Schnurr, & Coyne, 2010; Hollanders & McLeod, 1999; Thoma & Cecero, 2009). Grounding a new scale in the literature, where possible, is therefore crucial for therapy process measures because attributes of therapy process differ between different therapeutic orientations and may differ between therapists. In the case of CE, collaboration is understood as an 'active sharing of the work' of therapy; a specific understanding that is different, for example, from collaboration in the WAI and the CTRS.

Divergence from the seminal literature may raise doubts about whether a new scale adequately captures the construct as it has been defined and theorized in the literature. Divergence from the literature may also limit the relevance of a scale for previous research that is based on the established definition. An example of this issue can be seen in the results of the post-development evaluation of the CES. Expert consensus on ecological validity was 81.4% overall. In contrast, consensus on ecological validity for Item B. Rating emotion or belief, was rated at 74.6%. One respondent suggested that rating the degree of emotion associated with a belief should not be considered a part of CE. Six experts responded that rating emotion was 'somewhat' an aspect of their clinical practice (i.e., rated 3 out of 5). Another commented that practitioners know they should rate emotion ("for accreditation") but

typically don't. These results suggest that for some experts the rating of emotion was perhaps a 'non-core' aspect of CE, and in one case a contentious addition to the construct. This is in contrast to both the research and clinical literature, in which rating emotion has been consistently described as a core aspect of CE (A. T. Beck et al., 1979; J. S. Beck, 1995, 2011). Had expert respondents been asked to nominate which aspects of the CE construct should be included in the CES, some of them may have omitted or marginalized the rating of emotion. This in turn could have lead to confusion regarding core aspects of CE, or differences in the measure from the seminal understanding in the literature. This has direct consequences in the present context because, as we shall see in the results of the examination of CE as a predictor of therapeutic outcome in Chapter 8, only a single example of rating emotion was present in this sample. This would not have been revealed if the CES did not include an item measuring the rating of emotion. This issue demonstrates the utility of post-development evaluation of a psychotherapy process measure, in which the measure has been designed from established understandings in the literature, and then evaluated post-development by experts, compared with the more typical Delphi study in which experts select aspects of a construct for inclusion in a measure.

Conclusion

The present research sought quantitative and qualitative feedback on the CES from 30 expert CBT researchers and practitioners. Experts strongly and consistently endorsed the scale for content validity, ecological validity, and ability to capture variance in CE. Experts also provided a high level of endorsement for the clarity of the scale items and anchor points, and contributed 104 comments focused on improving the clarity and usability of the scale. Despite the large number of detailed comments, experts did not suggest significant alterations to the conceptual framework

or rating structure of the scale, nor identify issues with validity or missing aspects of the construct. Expert feedback was clear and specific, and was addressed comprehensively and in detail. Revision of the scale is therefore likely to have improved clarity and usability, without changing the basic parameters endorsed by experts. The strong endorsement of the scale by experts provides support for the theoretical work underpinning the scale. Although not specifically asked to suggest aspects of CE missing from the scale, experts had numerous opportunities to comment on missing elements but did not. Experts were not asked to review the revised scale, but it was argued that experts' comments provided sufficient guidance to have addressed the issues regarding clarity they raised. It was suggested that future research could usefully employ a post-development, consensus driven, expert evaluation of the measure during construction. The results of this process provided a solid basis for proceeding to the training of raters and pilot testing of the scale.

Chapter 6. Common Method Section for Empirical Studies 2 and 3

Aims and Outline

For ease of reference, and in order to avoid repetition, this chapter presents elements of the Method that are common to Studies 2 and 3. This includes details of participants, therapists, therapy protocol, measures, and descriptive statistics for archival data. Elements of the Method specific to individual studies are presented in the chapter reporting on each study. The titles of the empirical studies 1 – 3 are presented Table 20.

Table 20

Titles of Empirical Studies 1 – 3

Study number	Empirical study title
1	Expert Review of the Collaborative Empiricism Scale
2	Pilot Studies A and B
3	Collaborative Empiricism as a Predictor of Therapeutic Outcome

Common Method for Empirical Studies 2 and 3

Participants

Audio recordings of CBT for depression and data for depression severity used in this study originated in a large, multisite, NIMH funded study of the efficacy of CBT for depression, conducted by Jacobson, Dobson, and colleagues (Jacobson et al., 1996). The Jacobson et al. study received ethics approval from each of the participating institutions. The present research received ethics approval from the La Trobe University Faculty of Science, Technology, and Engineering Human Ethics Committee (approval number FHEC09.R59). A formal data access and publication protocol was also agreed between the La Trobe University Cognitive-Behaviour

Therapy Research Unit (CBTRU), Professor K. S. Dobson, and the student researcher. Text of the authorship and dissemination agreement is presented in Appendix I.

Grateful acknowledgement is given here for the invaluable assistance of Professor K. S. Dobson, a principle researcher in the Jacobson et al. (1996) study, in obtaining use of these data and audio recordings, as well as for helpful discussions regarding the methodology of rating. Data for ratings of the working alliance and therapist competence for this sample were gratefully obtained from two fourth year undergraduate projects also conducted at the CBTRU, under the supervision of the CBTRU Director, and supervisor of the present research, Dr. Nikolaos Kazantzis.

The sample consisted of a subset (44 of 50) of the participants from the cognitive therapy condition of the Jacobson et al. (1996) study. Six participants from the original Jacobson et al. study were excluded as the audio recordings of their sessions were unclear. Participants were 44 individuals (33 female (75%), 11 male (25%)) aged between 21 to 60 years (M = 38.6 years). All participants had completed high school, 27% had completed college, and had 12% completed post college education. Eighty per cent of participants were referred to the study from Group Health Cooperative, a large health maintenance organization in Washington State, and 20% recruited from public service announcements. All participants met criteria for Major Depressive Disorder in the Diagnostic and statistical manual of mental disorders (3rd ed., rev.; DSM–III–R; American Psychiatric Association, 1987). Diagnoses were made using the Structured Clinical Interview for DSM-III-R (Spitzer, Williams, & Gibbon, 1987). All participants scored 20 or higher on the Beck Depression Inventory (BDI; A. T. Beck et al., 1979; A. T. Beck, Ward, Mendelson, Mock, & Erbaugh, 1961) and scored 14 or higher on the 17-item version of the

Hamilton Rating for Depression (HRSD; Hamilton, 1960; Hamilton, 1967). In order to reduce potential confounds of the therapy protocol, participants were excluded if they had current diagnoses of a psychotic disorder, psychotic depression, bipolar disorder, panic disorder, current substance abuse or dependence, past diagnoses of schizophrenia or schizophreniform disorder, organic brain syndrome, mental retardation, or if hospitalization was required for current risk of suicide or psychosis.

Therapy Protocol

Treatment was provided by four experienced CBT therapists. Therapists had a mean age of 43.5 years (range 37 – 49 years), a mean post-degree clinical experience of 14.8 years (range 7 – 20 years), and a mean length of cognitive therapy practice of 9.5 years (range 8 – 12 years). Therapists underwent an additional year of training using a cognitive therapy condition training manual based on Beck et al.'s (1979) original cognitive therapy manual (Center for Clinical Research, 1996). Adherence to the cognitive therapy treatment protocol was measured by Professor K. S. Dobson, who provided supervision for all four therapists, screening a random selection of 20% of recorded sessions using a modified version of the National Institute of Mental Health Collaborative Study Psychotherapy Rating Scale (CRPRS; S. D. Hollon et al., 1988; S. D. Hollon, Evans, Elkin, & Lowery, 1984). Therapists were immediately alerted to any protocol violations, and any difficulties with the protocol were reviewed, while the clinical stage of the trial was underway, in monthly meetings between all therapists and the chief investigators N. S. Jacobson and K. S. Dobson. Therapist competence in CBT was measured by Professor Dobson rating a random selection of sessions using the Cognitive Therapy Rating Scale[2] (CTRS; Vallis et al.,

[2] The Cognitive Therapy Rating Scale (CTRS) is also referred to in the literature, chiefly for historical reasons, as the Cognitive Therapy Scale (CTS). The Beck Institute currently refers to the scale as the 'Cognitive Therapy Rating Scale' and that name is used here. The Cognitive Therapy

1986; J. E. Young & Beck, 1980b). The CTRS has shown adequate reliability and validity as a measure of therapist competence in CBT (K. S. Dobson, Shaw, & Vallis, 1985; Vallis et al., 1986). The mean competency scores for all sessions were above 40, the conventional cut-off for competence on the CTRS (Vallis et al., 1986). Therapists in the sample used in the present study obtained mean scores on the CTRS of 45.16, 44.01, 47.91, and 46.17.

Participants completed at least 12 of 20 sessions of CBT for depression and 40 of the 44 participants (90.1%) completed the full 20 sessions. The therapy employed a wide range of standard interventions in cognitive therapy, including identification and examination of automatic thoughts, underlying assumptions, core beliefs and schemas; the development of alternative beliefs, assumptions, and core beliefs; and the examination of the short versus long term advantages of beliefs. A range of common behaviourally focused techniques were also used, including behavioural monitoring and activation to increase pleasure and mastery, scaffolding and structured increase in difficulty of challenging behaviours, and rehearsal of difficult or stressful situations in session.

Measures

The Collaborative Empiricism Scale.

The CES is presented in Appendix A.

The Beck Depression Inventory.

The Beck Depression Inventory (BDI; A. T. Beck et al., 1979; A. T. Beck et al., 1961) is a 21-item client report measure of depression severity that has been widely used in clinical work and research. It has demonstrated excellent psychometric properties (A. T. Beck, Steer, & Garbin, 1988), is sensitive to clinical change (Hill &

Rating Scale (CTRS) is distinct from the Cognitive Therapy Scale – Revised (CTS-R; Blackburn et al., 2001).

Lambert, 2004), and stable over time (Barkham, Mullin, Leach, Stiles, & Lucock, 2007). The BDI was administered at pre-treatment, at the start of every therapy session, at post-treatment, and at follow-up 6, 12, 18, and 24 months post-treatment.

The Working Alliance Inventory.

The Working Alliance Inventory – Short Revised – Observer version (WAI-SR-O; Kazantzis, Osborne, & Cronin, 2012) is an adaptation of the Working Alliance Inventory – Short Revised (WAI-SR; Hatcher & Gillaspy, 2006). The WAI-SR-O consists of the 12 items of the WAI-SR, with the wording of items adapted to the perspective of observer raters rather than the therapist. Each item records the frequency and intensity of negative interactions between client and therapist on a 5-point Likert scale. The WAI-SR-O also differs from the WAI-SR in that it is designed for use in rating CBT sessions and provides explicit guidelines for raters. Guidelines for raters were expanded from guidelines provided for the Working Alliance Inventory – Observer version (WAI-O; Raue, Castonguay, & Goldfried, 1993).

The Cognitive Therapy Rating Scale.

The Cognitive Therapy Rating Scale (CTRS; J. E. Young & Beck, 1980a) is a measure of therapist competence in CBT. The scale consists of 11 items, divided into two subscales of General Therapy Skills and Specific Cognitive Behavioural Therapy Skills. Each item is rated on a 7-point Likert scale ranging from ‘poor’ to ‘excellent’. The CTRS has demonstrated good internal consistency (K. S. Dobson et al., 1985) and good inter-rater reliability (Vallis et al., 1986); however, some studies have reported limitations in the CTRS (J. P. Barber et al., 2003).

Chapter 7: Pilot Studies A and B

Aims and Outline

This chapter reports on two pilot studies of the CES. The aims of Pilot Study A were to evaluate inter-rater reliability for the CES, and refine the rating methodology. Adequate inter-rater agreement was attained by the end of training for both Group 1 (postgraduate) raters and Group 2 (undergraduate) raters. A problem was identified whereby errors in dividing a session into segments to be rated reduced the reliability of subsequent ratings of empiricism and collaboration for that session. To address this, a change in the rating methodology was evaluated in a second pilot study, and a final version of the scale produced, for use in Study 3: CE as a predictor of therapeutic outcome (Chapter 8).

Method

Participants, Therapists, and Therapy Protocol

To avoid repetition of content, details of participants, therapists, and the therapy protocol are described in Chapter 6: Common Method Section for Empirical Studies 2 and 3.

Raters

Eight raters took part in Pilot Studies A and B. Raters were recruited into two groups. Group 1 consisted of three postgraduate psychology students (two female, one male). All postgraduate raters were engaged in the development of measures of CBT process, and two were also undertaking postgraduate training as clinical psychologists and CBT therapists. The roles of the postgraduate raters were to contribute their experience of scale development and clinical work in CBT to the training of raters and pilot rating of the scale, and to provide a baseline for comparison of the Group 2 raters.

Group 2 consisted of five raters, who had varying degrees of psychology and CBT training. Three had completed honours level (four years undergraduate) psychology training, one of whom had completed a brief CBT training course, and two were current honours (fourth year) students in psychology. All Group 2 raters had completed an evidenced-based practice subject, which covered the cognitive model, fundamentals of CBT, the evidence base for CBT, and an introduction to clinical interventions. One rater from Group 2 (rater 'R4') voluntarily withdrew from the pilot study near the end of the three days of rater training.

Measures

The Collaborative Empiricism Scale (CES).

The CES incorporated revisions in response to the expert feedback described in Chapter 4. The complete text of the CES appears in Appendix A.

Rater Training Materials

Training was conducted using a purposely developed PowerPoint presentation. The training PowerPoint is presented in Appendix J. Training first outlined the cognitive model, defined CE and related core constructs, and provided clinical vignettes to illustrate collaboration and empiricism in CBT. The student researcher prepared criterion ratings of therapy sessions for training, by using the CES to rate randomly selected sessions from the Jacobson et al. (1996) archival recordings. A diagram was developed illustrating the use of the client's experience at each level of empiricism from 1 – 5 on the CES (See Appendix F). Training also employed brief audio segments from a CBT training DVD authored by Dr. Christine Padesky and published by the American Psychological Association.

Procedure

Rater training.

Training and pilot rating took place in a conference room at La Trobe university. Training was conducted by the student investigator with additional input from the research supervisor, Dr. Nikolaos Kazantzis. Group 1 (postgraduate) and Group 2 (undergraduate) raters participated in comprehensive training over three days. Training days 1 and 2 took place over one weekend and day 3 the following Saturday. First, the cognitive model underpinning CBT was reviewed. Raters were then asked to brainstorm and discuss together their understanding of CE, collaboration, and empiricism in CBT. The aim was to activate raters' existing knowledge of CE, and relevant related constructs, and to motivate participation in the training process. A trainer led group discussion of the CE construct followed, with clinical examples of CE in cognitive therapy. Raters were then asked to read the 'Instructions to raters' provided with the CES. Elements of the scale were then introduced progressively. Finally, each item on the scale was discussed and practice exercises conducted to familiarize raters with each item and the rating process.

Trials 1 – 5.

Ratings in Pilot Study A involved a sequence of five *trials*. Each trial began by training raters on an aspect of the scale, followed by written and audio examples and discussion, and then rating of an audio recording of a therapy session or training DVD. Progressively more elements of the scale were introduced across the series of trials. In Trial 1, raters were required to identify only the focus of discussion in the session and to record the time when it changed. In Trials 2 and 3, raters identified the focus of discussion and identified the appropriate scale item to be rated for that segment. In Trials 4 and 5, the full scale was rated, including identifying the focus of

discussion, selecting the matching empirical item (A, B, C, D, E, F) to rate, and rating empiricism and collaboration (Item G) for each focus of discussion. After each trial, the reliability of rating was assessed, and corrective feedback given to raters individually and as a group. A summary of the scale elements rated in each trial is presented in Table 21.

Table 21

CES Elements Rated in Each Trial

Trial	CES elements rated
1	Focus of discussion only
2	Focus of discussion and identifying item (A, B, C, D, E, F) to rate
3	Focus of discussion and identifying item (A, B, C, D, E, F) to rate
4	Full scale. Empiricism and collaboration for each segment
5	Full scale. Empiricism and collaboration for each segment

Results

Data Structure of the CES

Ratings of the CES produce two levels of data. At the first level, the CES produces nominal data, in the form of the start time for each focus of discussion (each segment of the session) and a one-line description of each focus of discussion. At the second level, the CES produces interval data, in the form of ratings of empiricism and collaboration for each segment. Each segment is rated for empiricism, using the appropriate empirical item (A – F), and also collaboration using Item G. An example rating sheet showing the two-level data structure of the CES is presented in Table 22.

Table 22

Rating Sheet Showing the Two-Level Data Structure of the CES

Segment Start	Item (A–F)	Emp. (1-5)	Collab. (1-5)	Focus of Discussion
0.08	C	2	4	"You were right last week it was really stressful."
0.52	A	2	2	"So… I'm interested in hearing kind of what's transpired and how you're doing with that… and whatever you would like us to focus in on…"
1.01	C	2	4	"Well, A. moved in."
3.59	C	2	4	"I feel threatened by him."
5.16	C	3	4	"A. and I are going to be really good together, I can see that."

As seen in Table 22, Level 1 nominal data, consisting of each segment in the session and the related focus of discussion for that segment, run down the page. Ratings of empiricism and collaboration for each segment (the Level 2 interval data) run across the page.

Inter-rater Agreement

Choice of index for inter-rater agreement.

Agreement between raters at the first level of nominal data (agreement on segmenting the session) was calculated using percentage agreement (Kolbe & Burnett, 1991). Percentage agreement is the number of agreements between raters divided by the total number of ratings made (correct and incorrect). It is a widely used

index of agreement between ratings of nominal data (Tinsley & Weiss, 2000). Percentage agreement is quick to calculate and easy to interpret, which were important advantages during training as agreement was used to inform feedback to raters between ratings. Percentage agreement has the disadvantage that it does not account for chance agreement between raters, which may lead to overestimation of agreement (Hayes & Hatch, 1999). A standard recommendation in the literature is that .70 is an acceptable lower bound for percentage agreement (Hartmann, 1977; Stemler, 2004). This level applies to ratings where a choice is made from a small number of predefined options, such as selecting from the anchor points on a Likert scale. In Pilot Study A, however, raters were firstly required to listen to the free flow of a therapy session and segment the session into separate focuses of discussion. Therapeutic discourse can be semantically rich and complex because therapy is typically focused on a specific therapeutic purpose and is time limited (Weck, Bohn, Ginzburg, & Stangier, 2011). This complexity makes segmenting a discourse on the basis of its semantic content difficult. To compensate for this complexity, the minimum level of agreement needed to progress to the next trial in pilot training was set at .60.

Two other indices of agreement commonly used in psychological research are Cohen's Kappa (J. Cohen, 1960) and Fleiss' multi-rater Kappa (Fleiss, 1971). Cohen's Kappa does account for the effect of chance agreement between raters, but was not appropriate for the current study as it can only be calculated for comparisons between two raters (J. Cohen, 1968). Fleiss' multi-rater Kappa is a modification of Cohen's Kappa that can be used to calculate agreement between more than two raters. Both Kappa statistics, however, are highly dependent on the symmetry of the marginal distribution ('marginal dependence') (Agresti, 2002). Marginal dependence

means that in nominal data with two categories, such as data representing the presence or absence of a focus of discussion in the first level of the current study, Kappa may be strongly affected if the ratio of present to absent ratings (the 'symmetry') differs from 1. It is highly likely that the assumption of symmetry is not met in clinical data (Scott, 1955). In light of these considerations, percentage agreement was chosen as the index of agreement for nominal data.

Agreement between raters for the second, interval level of data (rating empiricism and collaboration for each segment) was calculated using intraclass coefficients[3] (ICCs: Shrout & Fleiss, 1979). Intraclass coefficients are a widely used measure of inter-rater reliability in the psychological literature (McGraw & Wong, 1996). Intraclass coefficients were calculated using the SPSS version 21 two-way random effect, absolute agreement ICC (2, *k*) procedure, where *k* = number of raters (Shrout & Fleiss, 1979). The ICC (2, *k*) statistic is appropriate when raters are selected from a larger population of raters, so that results of analyses are generalizable to raters from comparable populations (McGraw & Wong, 1996). This model also accounts for systematic variability between raters.

Trials 1 – 5

Nominal Data: Splitting the session into segments.

Percentage agreement was used to measure agreement between ratings of the start time of each focus of discussion, and a brief one line description of each focus of discussion. Rating the start time of each focus of discussion is equivalent to raters' division of the therapy session into segments, where each segment contains a single focus of discussion.

[3] In Study 3 (Chapter 8), Finn's *r* is used as a measure of inter-rater reliability in addition to ICCs. Finn's *r* was not used during pilot evaluation or rater training because inter-rater reliability was examined immediately after each rating trial, and tools for calculating Finn's *r* quickly were not available.

Table 23 presents the percentage agreement between each rater and the student researcher's criterion rating, for segmenting the session into focuses of discussion, for each Trial 1 – 5. Rater R4, who withdrew near the end of training, was excluded from all analyses, but is included in Table 23 for comparison purposes.

Table 23

Percentage Agreement for Segmenting Into Focuses of Discussion, for Trials 1 – 5

Trial	All raters without R4[a] (*n*=7)	All raters[b] (*n*=8)	Group 1 (Postgraduate) (*n*=3)	Group 2 without R4 (*n*=4)	Group 2 (*n*=5)
1	51.4	49.7	69.0	38.2	38.1
2	65.3	63.1	70.3	61.6	58.7
3	65.4	66.1	63.8	66.6	67.5
4	54.0	54.8	61.6	48.3	50.7
5	69.2	61.9	81.8	62.9	54.0

Note. [a]R4 = Rater 4 (from Group 2). [b]All raters = Groups 1 and 2 combined.

Trial 1.

As seen from Table 23, the Group 1 (postgraduate raters) rated at an acceptable level on their first attempt, whereas the Group 2 (undergraduate) raters rated poorly. This result shows that the postgraduate group were able to reliably divide a therapy session into segments and describe the focus of discussion of the segment with relatively little training. In contrast, the Group 2 (undergraduate) raters had considerable difficulty with this task, rating at less than chance.

Following Trial 1, the group discussed their experiences of rating. Raters reported that they had most difficulty identifying when the focus of discussion changed to a new focus (i.e., splitting the session into segments). This self-report was consistent with their ratings, in which the most common mistake was adding extra

segments, by incorrectly identifying additional focuses of discussion. Feedback was given to raters to help them determine when the current focus of discussion changed to a new one. Practice test ratings were then conducted, followed by group discussion. When raters were no more than one point apart on their test ratings, training proceeded to Trials 2 and 3.

Trials 2 and 3.

In Trials 2 and 3, raters were required to again identify the focus of discussion for each segment. In addition, they were required to select which empirical item (A – F) should be used to rate each segment. As seen in Table 23 (above), percentage agreement for Group 1 (postgraduate) raters remained approximately static, whereas Group 2 (undergraduate) raters improved across ratings 2 and 3 to a level comparable to the postgraduate group. The combined level of agreement of all raters at the end of Trial 3 (65%) was deemed sufficient to proceed to Trial 4.

Trial 4.

Trials 4 and 5 involved a considerable increase in difficulty compared with Trials 2 and 3 because raters were required to rate the full scale. This entailed firstly rating the start time and focus of discussion for each segment and selecting the matching empirical item (A – F), as per Trials 2 and 3, and then additionally rating empiricism (Items A – F) and collaboration (Item G) for each segment they had identified.

As seen in Table 23, in Trial 4, percentage agreement was approximately stable for Group 1, but fell considerably for Group 2. Raters reported that they found rating the full scale difficult because they were still considering a rating for one segment when ‘the next segment was upon them’.

Feedback for raters after Trial 4 concentrated on issues related to choosing the correct empirical item (A – F) to rate. It is noted here that these issues did not arise during Trials 2 and 3, in which selections of the item to rate were made with adequate reliability for both Groups 1 and 2.

Raters had difficulty with Items A, C, and D. With Item A ('Selecting a focus of discussion to explore'), raters had difficulty distinguishing situations where the focus of discussion flowed into the next focus of discussion without a selection being made. For example, if the client was discussing one issue and said, "it's the same with…" and proceeded to discuss a second issue, without any explicit reflection or discussion of the choice of topic. Clarification was given that Item A is rated only when an explicit, deliberate choice is made regarding a focus. With Item D. 'Developing an alternative interpretation', raters reported difficulty distinguishing when to rate Item C ('Exploring') followed by Item D, versus when to rate only Item D. Clarification was given that the first case (C then D) is rated when the client and therapist are exploring an issue and then move to developing an alternative interpretation. For example, Items C then D would be rated if the client and therapist are discussing the client's social anxiety (e.g., "colleagues in a work meeting think I am a bad presenter because they are staring at me") and the client and therapist then move to developing an alternative interpretation of the 'staring'. Item D would be rated alone if the client and therapist moved straight to discussing an alternative to a thought on a homework sheet, for example.

Further training was given in selecting items by working through two example thought records and discussing these as a group. This allowed raters to learn about other raters' thought processes while rating, and to get feedback on rating difficulties

in real time. Once raters were proficient at selecting items to rate, training proceeded to Trial 5.

Trial 5.

As seen in Table 23, in Trial 5, reliability for the Group 1 (postgraduate) raters improved strongly, and Group 2 honours level raters improved to an acceptable level. These results demonstrate that postgraduate raters were able to rate well initially, and subsequently improve to a high level of agreement by the end of Pilot Study A. Group 2 honours level raters were able to reach an acceptable level when rating only nominal data, but were not able to improve on this when required to rate the full scale, despite three days of training.

Correct versus incorrect ratings.

Raters' performance is illuminated further by examining the pattern of correct and incorrect ratings across each Trial 1 – 5. Ratings of agreement on dividing the session into segments by focus of discussion could be scored in one of three ways: *Correct* – where the rating matched a segment in the criterion rating; *Incorrect* – where there was a segment in the criterion rating that should have been rated but the rater missed it; and *Incorrect-Additional* – where the rater created an additional rating that did not match any segment in the criterion rating. Percentage agreement was calculated by dividing the number of Correct ratings by the total number of ratings (Correct, Incorrect, and Incorrect-Additional). Table 24 shows the number of Correct and Incorrect-Additional ratings for each Trial 1 – 5.

Table 24

Number of Correct and (Incorrect-Additional) Ratings by Rater for Each Trial

Rater Group	Rater	Trial 1	2	3	4	5
Group 1	R1	4 (0)	14 (2)	6 (4)	5 (1)	10 (0)
	R2	4 (4)	10 (0)	5 (2)	6 (1)	Absent
	R3	4 (3)	12 (1)	6 (4)	5 (0)	8 (0)
Group 2	R4	3 (4)	9 (3)	5 (2)	6 (2)	2 (0)
	R5	3 (4)	13 (7)	8 (4)	6 (0)	9 (2)
	R6	3 (5)	14 (6)	5 (2)	3 (0)	6 (2)
	R7	3 (4)	11 (3)	8 (5)	4 (3)	9 (0)
	R8	4 (5)	15 (6)	8 (4)	4 (1)	6 (0)
Criterion Rating		4 (0)	16 (0)	8 (0)	8 (0)	11 (0)

Note. The number of incorrect ratings is given by subtracting the number of Correct ratings for Raters 1 – 8 from the number of Correct ratings in the Criterion Ratings.

The impact of Incorrect-Additional ratings on percentage agreement can be seen from Table 24. In Trial 1, the Group 1 postgraduate raters' overall percentage agreement was 69% (see Table 23 above). Table 24 shows that all raters in (postgraduate) Group 1 got all four ratings in the criterion rating correct, but raters R2 and R3 added seven Incorrect-Additional ratings between them, giving a mean percentage agreement for Group 1 of (100 + 50 + 57.1)/3 = 69%. The impact of Incorrect-Additional ratings is even clearer in Trial 2. The mean number of Correct

ratings was nearly identical for Group 1 (*M* = 12, *SD* = 2) and Group 2 (*M* = 12.4, *SD* = 2.4) but the Incorrect-Additional ratings were far lower for Group 1 (*M* = 1, *SD* = 1) compared with Group 2 (*M* = 5, *SD* = 1.9). These results show that in Trials 1 and 2, in which raters had to rate only the focus of discussion and select the correct empirical item (A – F) to rate, Group 1 (postgraduate) raters showed higher agreement with the criterion rating because of lower Incorrect-Additional ratings compared with Group 2 (undergraduate) raters.

Table 24 also shows that the number of Incorrect-Additional errors improved for both groups over the five Trials. In Trial 4, which was the raters' first attempt at rating the full scale, both groups did poorly but this was caused by low Correct rather than Incorrect-Additional ratings (Group 1 *M* = 5.3, *SD* = 0.6 Correct; Group 2 *M* = 4.6, *SD* = 1.3 Correct). Raters were making far fewer Incorrect-Additional ratings but either missed focuses of discussion completely or rated the right segments but rated them incorrectly.

Table 25 shows the percentage agreement for ratings of correctly identified segments in the criterion rating, across Trials 2 – 5. Rating a correctly identified segment means that raters had already correctly identified a segment that existed in the criterion rating, and then rated that (correctly identified) segment using an empirical item A – F.

Table 25

Percentage Agreement by Item A – F for Correctly Identified Segments

Item	Trial 2	Trial 3	Trial 4	Trial 5
A	77.1 (6)	71.9 (4)	81.3 (2)	60 (5)
B	87.5 (1)			
C	81.3 (4)	87.5 (4)	54.2 (6)	60.7 (4)
D	70.0 (5)			85.7 (2)
Total Instances	16	8	8	11

Note. Numbers in brackets refer to number of instances of each CES Item rated. Items E – F were not present in the session rated.

As seen in Table 25, percentage agreement for Item C in Trials 2 and 3 was high compared with low agreement in Trial 4 and adequate agreement in Trial 5. Item

A had similar agreement in Trial 5, compared to earlier trails. In Trials 2 and 3, raters were required to rate only the start time and focus of discussion for each segment, compared with Trials 4 and 5 in which they rated the full scale. Table 25 shows that in Trials 2 and 3, when raters correctly identified a session to rate, they subsequently chose the correct item (A – F) to use to rate that segment for empiricism with a high degree of accuracy. Whereas in Trials 4 and 5, after correctly identifying a segment, they were much less accurate in choosing the correct item to rate. This result shows the negative impact on agreement when raters were required to rate the full scale in Trials 4 and 5. That is, adding the extra task of rating empiricism and collaboration for each segment adversely affected agreement on the task of correctly selecting a segment to rate.

Interval Data: Rating empiricism and collaboration for each segment.

Trials 4 and 5 involved rating the full scale, including ratings of empiricism and collaboration for each segment. Table 26 presents ICCs for ratings of empiricism and collaboration across Trials 4 and 5.

Table 26

ICCs for Empiricism and Collaboration Across Pilot Ratings

Trial	Empiricism All raters except R4	Empiricism All raters	Collaboration All raters except R4	Collaboration All raters
4	.558	.756	.674	.694
5	.754	.690	.626	.670

As seen in Table 26, by Trial 5, ICCs were adequate for both empiricism and collaboration. It is important to note that Table 26 presents ICCs for ratings of

empiricism and collaboration only for the segments that were correctly identified by the raters. Segments that were incorrectly identified were invalid, and consequently ratings of empiricism and collaboration for these segments were also invalid. Table 23 shows that in Trial 5, Group 2 rated 62.9% of segments correctly, that is, 27.1% of segments were incorrectly identified and consequently had invalid ratings of empiricism and collaboration. The ICCs reported in Table 26 apply to the 62.9% of segments identified correctly. This demonstrates the substantial effect of errors in segmentation on ratings of empiricism and collaboration.

Discussion

The aims of Pilot Study A were to assess whether raters could achieve adequate reliability using the CES, and to identify any difficulties rating the scale. It was hypothesized that postgraduate raters, with experience developing CBT process scales and/or clinical training in CBT, would be better raters than undergraduate raters without this level of experience. This hypothesis was supported. Group 1 (postgraduate) raters, who had process research experience and/or clinical training, were able to reliably divide a therapy session into segments, at an adequate level of agreement in their first trial, and attain a high level of agreement by the end of training. Group 2 undergraduate raters initially rated with lower agreement but attained an adequate level of agreement by the end of training. These results suggest that postgraduate raters with clinical training or experience in developing psychotherapy process scales are better raters of psychotherapy processes compared with undergraduate raters without such training or experience. This is consistent with research showing that evaluating therapist competence is enhanced if raters have expertise in the specific treatment modality being evaluated (J. Waltz, Addis, Koerner, & Jacobson, 1993), and that reliable judgments of psychotherapy processes

are possible between experienced raters with a shared understanding of the therapeutic orientation (Messer, 1991). These results demonstrate that good reliability can be attained for dividing a session into segments, but suggest that this may require some familiarity with therapy process.

The Segmentation Problem

Pilot Study A revealed an issue with segmenting sessions into separate focuses of discussion. In Trials 2 and 3, Group 2 raters divided sessions into segments with adequate agreement, similar to the Group 1 (postgraduate) raters. In contrast to this, in Trial 4, Group 2 raters had difficulty with the same activity of splitting the session into segments. The difference between Trials 2 and 3 compared with Trial 4 is that Trial 4 included the extra task of rating empiricism and collaboration for each segment. The drop in percentage agreement when rating Trial 4 suggests that the extra cognitive load of simultaneously rating each segment for empiricism and collaboration made segmenting more difficult. Group 2 raters improved in Trial 5, but remained at only an adequate level at the end of training and did not improve past the level attained when segmenting the session alone in Trial 3.

The final percentage agreement obtained by Group 2 raters for dividing the session into segments (62.9%) is problematic because errors in segmenting the session directly reduce the reliability of ratings of empiricism and collaboration for those segments. If segmentation is incorrect for any reason, subsequent ratings of empiricism and collaboration for that segment are rendered invalid. This would perhaps be acceptable at the 81.8% agreement for segmenting attained by the postgraduate raters; however, at the 62.9% agreement attained by the Group 2 raters, the impact of incorrect ratings of nominal data on subsequent ratings of empiricism and collaboration would be considerable.

A Response to the Segmentation Problem.

Segmenting a session is difficult because therapy is semantically complex on multiple levels. Small segments of therapy sessions can be meaning rich, containing complex markers of therapist adherence and competence (Weck et al., 2011) and key elements of core client dynamics (Mintz & Luborsky, 1971; Stiles et al., 2006). This density and complexity makes identifying the focus of discussion for a small part of the flow of therapy discourse challenging. It seems reasonable to assume that this difficulty would be increased for raters with relatively little experience of therapy sessions.

In discussion with Professor K. S. Dobson regarding this issue (personal communication, October 6, 2011), Professor Dobson suggested that, as the focus of the current study was to investigate collaborative empiricism in CBT, rather than divide sessions into discrete semantic units, a change in the rating methodology might overcome the segmentation problem. Professor Dobson suggested that the student researcher rate all sessions, to serve as criterion ratings, and supply the start time of each segment in the criterion ratings to raters. As raters listened to a session, they would select the appropriate item to rate each segment, and then rate empiricism and collaboration for that segment. This removes the requirement for raters to divide the session into segments while simultaneously rating each segment for empiricism and collaboration. It was hypothesized that allowing raters to focus on rating empiricism and collaboration for each segment would result in higher reliability of rating. This methodology was tested in Pilot Study B.

Conclusion

Pilot Study A evaluated the reliability of the CES and sought to identify problems with rating. The reliability of dividing sessions into segments was found to

differ based on rater group. Group 1 (postgraduate) raters showed an acceptable level of agreement in their first trial and attained a high level of agreement by the end of training. Group 2 (honours level undergraduate) raters started with low agreement but improved to an acceptable level. Adequate reliability was attained for ratings of empiricism and collaboration for segments that had been identified correctly; however, the study identified an issue whereby errors in dividing a session into segments invalidated subsequent ratings of empiricism and collaboration for those segments, resulting in lower reliability overall for ratings of empiricism and collaboration. In keeping with the primary aim of the project, to investigate CE in CBT for depression, a change in the rating methodology was proposed, in which the student researcher would rate all sessions and pre-identify the start time of each segment. Raters would then be provided with a rating sheet containing each segment start time and would be required to rate collaboration and empiricism for each segment. This change was trialled in Pilot Study B.

Pilot Study B: Comparative Evaluation of the Pre-segmented Rating Method

Aims and Outline

The aim of Pilot Study B was to evaluate a revised method for rating the CES by comparing it with the original method from Pilot Study A. In the original *Rater-segmented* method from Pilot Study A, raters were required to divide the session into segments based on separate focuses of discussion, and rate each segment for empiricism and collaboration as they went. In the new *Pre-segmented* method, raters were provided with a rating sheet with the segment start times filled-in. They were required to choose the correct item to rate each segment, and then rate each segment for empiricism and collaboration. The revised method allowed raters to focus on the primary purpose of the study, rating CE, rather than simultaneously segmenting a therapy session into semantically related units. It was hypothesized that raters using the Pre-segmented method would achieve a higher reliability of ratings of empiricism and collaboration, and that raters using the original Rater-segmented method would show poorer reliability, as they had done in Pilot Study A.

Method

Participants

Participants were drawn from the same archival audio recordings of therapy sessions used in Pilot Study A, recorded as part of a study of the efficacy of CBT for depression (Jacobson et al., 1996).

Raters

Raters were the same four Group 2 (honours level undergraduate) raters from Pilot Study A.

Measures

The CES was used unchanged from Pilot Study A. Only the method of rating the scale was altered to test the new Pre-segmented rating method.

Procedure

The four Group 2 (undergraduate) raters were randomized into two groups of two raters. The groups were then randomly assigned to either the Rater-segmented or the Pre-segmented condition. Four randomly selected therapy sessions (S1 – S4), two from session 3 and two from session 15, were rated by the student researcher, to serve as criterion ratings. Rating sheets for the Rater-segmented condition were blank, as in Pilot Study A. Rating sheets for the Pre-segmented condition were identical to those provided for the Rater-segmented condition except that the start time of each segment was filled in.

Raters were blind to the aims and hypotheses of the study. Ratings were conducted on the same day, but the two rater groups had no communication with each other during rating and were located at different campuses of La Trobe University. Raters in the Rater-segmented condition were told only that another two pilot ratings were being conducted. Raters in the Pre-segmented condition were given revised rating sheets, with the segment start times filled in, and asked to rate collaboration and empiricism for each identified segment.

A series of four Trials were conducted. Each Trial consisted of one rater from each condition rating the same randomly allocated therapy session. Table 27 presents the order of ratings for each Trial (6 – 9). Trials were numbered 6 – 9 to distinguish them from Trials 1 – 5 in Pilot Study A.

Table 27

Order of Rating of Trials in Pilot Study B

	Rater-segmented Method 1		Pre-segmented Method 2	
Trial	Rater 5	Rater 6	Rater 7	Rater 8
6	S1			S1
7		S2	S2	
8		S3		S3
9	S4		S4	

Note. S1 – S4 denotes the sessions rated.

Results

All statistical analyses were conducted using SPSS Version 21. Inferential statistics were evaluated against an alpha level of .05.

Inter-rater Agreement

The accuracy of ratings in each condition were determined by calculating agreement with the criterion rating. For the Rater-segmented condition, percentage agreement was calculated for agreement on segmenting the session, and ICCs were calculated for agreement on ratings of empiricism and collaboration for each segment. For the Pre-segmented condition, ICCs were calculated for agreement on ratings of empiricism and collaboration for each segment. Table 28 presents percentage agreement and ICCs for Trials 6 – 9. Means (and standard deviations) for percentage agreement and ICCs for the Rater-segmented and Pre-segmented conditions are presented in Table 29.

Table 28

Percentage Agreement and ICCs for Rater-segmented and Pre-segmented Groups for Trials 6 – 9

		Rater-segmented		Pre-segmented	
Trial	Agreement	Rater 5	Rater 6	Rater 7	Rater 8
6	Percentage agreement	60			N/A[a]
	Empiricism	.44			.69
	Collaboration	.40			.64
7	Percentage agreement		65	N/A	
	Empiricism		.36	.64	
	Collaboration		.10	.68	
8	Percentage agreement		48		N/A
	Empiricism		.31		.79
	Collaboration		.41		.80
9	Percentage agreement	.59		N/A	
	Empiricism	.29		.69	
	Collaboration	.52		.69	

Note. [a] Percentage agreement for segments in the Pre-segmented condition was not calculated as the sessions were pre-segmented.

Table 29

Means (and Standard Deviations) of Percentage Agreement and Intraclass Coefficients for Rater-segmented and Pre-segmented conditions

	Rater-segmented	Pre-Segmented
Percentage Agreement	58 (7.2)	N/A[a]
ICCs for Empiricism	.35 (.07)	.70 (.06)
ICCs for Collaboration	.36 (.18)	.70 (.07)

Note. [a] Percentage agreement was not calculated for segments in the Pre-segmented condition as the sessions were pre-segmented.

As seen from Table 29, for the Rater-segmented method, the mean percentage agreement for dividing the session into segments was comparable with that obtained in Pilot Study A (Pilot Study A, M = 62.9%; Pilot Study B, M = 58%). Raters were just below the minimum acceptable level of 60% agreement for segmenting sessions into focuses of discussion. Mean ICCs for empiricism and collaboration in the Pre-segmented condition were twice that of the Rater-segmented condition. As the data were non-normally distributed, a Mann-Whitney U test was used to compare ICCs between groups. For empiricism, the test indicated that ICCs in the Pre-segmented condition (*Mean Rank* = 6.50, n = 4) were significantly higher than the Rater-segmented condition (*Mean Rank* = 2.50, n = 4), U = 0, z = -2.323 (not corrected for ties), p = .029 two-tailed, Cohen's r = .821, by convention a large effect (J. Cohen, 1988). Similarly, for collaboration, a Mann-Whitney U test indicated that ICCs in the Pre-segmented condition (*Mean Rank* = 6.50, n = 4) were significantly higher than the

Rater-segmented condition (*Mean Rank* = 2.50, n = 4), U = 0, z = -2.309 (not corrected for ties), p = .029 two-tailed, Cohen's r = .816, also a large effect.

Discussion

The hypothesis that raters using the Pre-segmented method would achieve higher reliability of ratings of empiricism and collaboration compared with raters using the Rater-segmented method was supported. The difference in reliability of ratings using the two rating methods was large. Raters in the Pre-segmented condition rated at an adequate level of reliability, and at twice the level of raters in the Rater-segmented condition. Raters in the Pre-segmented condition also rated empiricism and collaboration more reliability than in any trial in Pilot Study A. This improvement in reliability occurred despite a one week gap between training in Pilot Study A and the trials in Pilot Study B. These results demonstrate that removing the need to divide each session into segments enabled undergraduate raters to rate collaboration and empiricism to an adequate level of reliability and improve their reliability from Pilot Study A.

In contrast, raters in the Rater-segmented group showed borderline acceptable agreement when dividing sessions into segments, but poor reliability of ratings of empiricism and collaboration for each segment. This pattern of results is highly similar to that found in Pilot Study A. In Pilot Study A, undergraduate raters in Trial 3 showed adequate agreement regarding dividing sessions into segments, when that was the single required task, but reliability dropped sharply (to chance levels) in Trial 4 when they were required to segment the session and simultaneously rate empiricism and collaboration for each segment.

Undergraduate raters in Pilot Study B, using the Rater-segmented method, were not able to reliably segment a session and simultaneously rate empiricism and

collaboration for each segment. In contrast, postgraduate raters in Pilot Study A were able to segment sessions to a high degree of agreement (81.8%), and simultaneously rate empiricism and collaboration to an adequate level of reliability; and undergraduate raters in Pilot Study B were able to achieve adequate reliability of ratings of collaboration and empiricism when not required to simultaneously divide the session into segments. Taken together, the results of Pilot Study B support the utility of the Pre-segmented method to allow undergraduate raters to use the CES to rate CE in sessions of CBT for depression. On the basis of these results, the Pre-segmented method was adopted and the study of CE as a predictor of therapeutic outcome commenced.

Implications for Future Work

Postgraduate raters in the present study were a specialized group with particular experience related to the development of CBT process scales. To a lesser extent, undergraduate raters were also specialized in that they had all participated in an introductory subject on the CBT model and its role in treatment. The role of experience in the ability to rate CBT processes has implications for future research, training in CBT, and the evaluation of therapist competence. In a research context, future research may benefit from considering raters' experience with CBT process and process scales, in order to improve reliability of ratings of CBT processes. In a clinical context, past experience rating CBT processes may enhance the ability to reliably focus on the key elements of a CBT session and identify specific CE events within the flow of a therapy dialogue. The skills involved in identifying CE events underpin clinical awareness of collaboration and empiricism, and possibly other attributes of the therapeutic relationship. Enhancing the ability to focus on the key content of a session could have application in the training of CBT therapists, for

example in the UK's Improving Access to Psychological Therapies (IAPT) program, the Beck Initiative's 'Partnership to implement cognitive behavioural therapy in a community health setting', or, in the Australian context, for the evaluation of trainee competence in post-graduate clinical training programs. Increased awareness of collaboration and empiricism, and the ability to focus on key session content, could help trainee therapists work better with their clients to develop shared case conceptualizations and test these out via empirical feedback, and also better tailor CBT to the specific social and cultural settings of clients from community health settings.

Future research could also explore differences in clinical trainees ability to segment sessions and identify CE events, and the relationship between these skills and scores on a measure of general therapist competence, such as the CTRS.

Conclusion

The aim of Pilot Study B was to compare a revised, Pre-segmented, method for rating the CES with the Rater-segmented method used in Pilot Study A. Raters in the Pre-segmented condition were able to reliably rate empiricism and collaboration using the CES, at a level twice that of raters in the Rater-segmented condition. In addition, raters in the Pre-segmented condition achieved a higher reliability of rating than in any of their ratings in Pilot Study A. Reliability of rating in the Rater-segmented condition was poor. These results demonstrate that adequate reliability of rating can be achieved on the CES, by undergraduate raters, using the Pre-segmented rating method. Future research may benefit from consideration of raters experience rating process scales. Future research could also usefully explore whether skills involved in identifying CE events enhance clinical trainees awareness of therapy processes, relationship attributes, and other therapist competencies.

Chapter 8: Study 3: CE as a Predictor of Therapeutic Outcome

Aims and Outline

After the CES had been assessed by experts and revised to incorporate their feedback, and had demonstrated adequate reliability in the pilot studies, the scale was used to rate sessions of CBT for depression. The aims of Study 3 were to (a) explore the characteristics of CE in the sample, (b) determine whether CE could predict reduction in depression severity at the end of therapy, and at four time points (6, 12, 18, and 24 months) post-therapy, (c) determine whether CE could predict symptom reduction after controlling for the working alliance and therapist competence, (d) examine the relationship between the duration of segments in a session and the quality of CE of those segments, and (e) explore the characteristics of CE within sessions.

Specific aims and hypotheses

As a result of the adequate reliability attained with the CES in pilot testing, it was predicted that raters would continue to improve the reliability of their rating over the course of the 132 sessions rated in Study 3, and demonstrate good reliability overall. Based on the view in the literature that CE is a fundamental element of the change mechanism in CBT, or a key mediator of change, it was hypothesized that CE would predict reduction in depression severity at termination, and continue to do so at 6, 12, 18, and 24 months post-therapy. It was further hypothesized that CE would predict change in depression severity beyond that accounted for by pre-therapy depression, the working alliance and general therapist competence. From the understanding that CE is a skill that takes time to acquire and to manifest in a session, it was hypothesized that longer segment durations would on average be related to higher segment CE scores.

Method

Participants, Therapy Protocol, and Measures

To avoid repetition of content, details of participants, therapists, and the therapy protocol are described in Chapter 6: Common Method Section for Empirical Studies 2 and 3.

Raters

Raters for Study 3 were drawn from two groups. Two raters were selected from the four Group 2 undergraduate raters in Pilot Studies A and B. These raters were chosen on the basis of the reliability of their rating, and their overall understanding of the CE construct. Another four undergraduate raters, designated Group 3, were recruited using the same procedure used to recruit the Group 2 raters. Group 3 raters had completed the same evidenced-based practice subject as the Group 2 raters. This subject covered the cognitive model and fundamentals of CBT, the evidence base for CBT, and an introduction to clinical interventions. Training for Group 3 took place in a conference room at La Trobe University, and used the same materials and procedure used to train the Group 2 raters. Raters in Group 3 were trained to use the new Pre-segmented rating method tested in Pilot Study B. Group 3 raters were given rating sheets with the start times of each segment filled in and required to select the correct CES item and use it to rate each segment for collaboration and empiricism. Raters in Group 3 benefitted from experience gained in training the raters in Groups 1 and 2. Training for Group 3 was completed in one day. Raters were judged to have attained an adequate standard of rating when their mean ICCs for empiricism and collaboration were at least .70. This was reached for Raters 9 and 10 after two rating sessions. As in earlier training, Finn's *r* was not used as an index of inter-rater reliability during training as no rapid method for calculating it was

available, and it was necessary to give raters immediate feedback on their reliability after each training rating. After training, raters were selected for rating in Study 3 based on their reliability and familiarity with the concepts underpinning the scale. The final group of raters for Study 3 consisted of Raters 7 and 8 from Group 2 (from the Pilot Studies) and Raters 9 and 10 from Group 3.

Procedure

Each rater was randomly assigned 2 – 4 sessions to rate at a time. Raters worked on laptops in an office at La Trobe University. They used headphones to listen to digitized copies of the original audio recordings of therapy sessions and recorded their ratings directly into Word document rating sheets. Reliability was initially calculated after each rating and corrective feedback given to raters as required. Once it was clear that reliability was adequate, reliability was calculated after each group of four ratings was completed.

Results

Data Structure of the CES

The base unit of data in the CES is the *segment*. In the process of rating, each session is divided into a series of contiguous segments. A segment is defined to be the duration of an individual focus of discussion. A focus of discussion is a discrete slice of therapy discourse in which the client and therapist are focused on a specific activity. Each focus of discussion belongs to a single segment, and each segment contains a single focus of discussion. Each segment is rated once for empiricism using the empirical item (A – F) matching the type of segment, and once for collaboration using the collaboration item (Item G). The CES thus provides a rating of empiricism and collaboration for each segment in a therapy session. Examples of focuses of discussion for a segment include setting an agenda, discussing an event from during

the week, deciding what to do for homework, and rating the degree of emotion associated with a thought.

Segment and Session scores.

The empiricism and collaboration scores for each segment are referred to as the *segment empiricism* and *segment collaboration* scores. The segment empiricism and segment collaboration scores were non-normally distributed, and consequently the median of the segment scores was the appropriate summary statistic to represent the segment scores per session. Accordingly, the median of the segment empiricism and collaboration scores for a particular session was used to represent the overall empiricism and collaboration scores for that session. The total CE score for a session was obtained by adding the session empiricism and session collaboration scores for that session.

Preliminary Data Analysis

All analyses were conducted using SPSS Version 21.0. Inferential statistics were evaluated against an alpha level of .05. The data were initially checked for errors, missing values, and outliers. There was one out of range value detected for the BDI. As the original BDI rating sheets were unavailable for consultation, and scores on the BDI for this client were normally distributed (M = 31.33, SD = 4.48), this value was replaced with the mean BDI score for this client (Tabachnick & Fidell, 2007).

Missing Data.

There were no missing values for any CES or BDI variables. There were three missing variables for the CTRS and three for the WAI. As no pattern could be detected in these missing values, and their number was small, they were deleted pairwise in all analyses (Tabachnick & Fidell, 2007).

Outliers.

Outliers were defined in terms of the interquartile range (IQR) (i.e., the middle 50% of scores) of the variable in question (Tukey, 1977). As the data were drawn from a clinical population, 'mild outliers' between 1.5 - 3 times the interquartile range were retained, as these values are likely to represent true variability in a clinical population. Cases with values three times the interquartile above or below the median (MEDIAN ± 3*(IQR)) were deemed to be 'extreme outliers' and were replaced with the value corresponding to either Q1 + 3×IQR or Q3 - 3×IQR respectively (Dawson, 2011; Tabachnick & Fidell, 2007).

Characteristics of the Independent and Dependent Variables Over Time

Prior to investigating the relationship between CE and outcome, the characteristics of each independent and dependent variable over time were examined. Table 30 presents means (and standard deviations) for the CES, WAI, CTRS, and BDI at each time point, early, middle, and late (sessions 3, 9, and 15 respectively).

Table 30

Means (and Standard Deviations) for CES, WAI, CTRS, and BDI at Sessions 3, 9, and 15

Variable	Time point				
	Pre-therapy	Session 3	Session 9	Session 15	Post-therapy
CES Empiricism	-	2.66 (0.50)	2.49 (0.54)	2.73 (0.42)	-
CES Collaboration	-	3.56 (0.48)	3.74 (0.42)	3.53 (0.51)	-
CES Total	-	6.22 (0.66)	6.23 (0.60)	6.26 (0.58)	-
WAI	-	21.51 (5.98)	17.95 (3.86)	20.65 (6.17)	-
CTRS	-	49.67 (5.62)	47.06 (7.52)	44.94 (5.67)	-
BDI	29.2 (6.3)	22.2 (9.0)	14.7 (9.8)	12.2 (9.3)	8.1 (7.3)

As seen in Table 30, the working alliance (WAI) was lower at session 9, and therapist competence (CTRS) declined progressively over the course of therapy. BDI dropped steadily from a high at the beginning of therapy to a low at termination. Collaborative empiricism was stable across time points, whereas empiricism and collaboration varied reciprocally over time.

Response Range

Figures 6, 7, and 8 show the frequency of responses across all segments in all session for Segment Empiricism, Segment Collaboration, and Segment CE respectively.

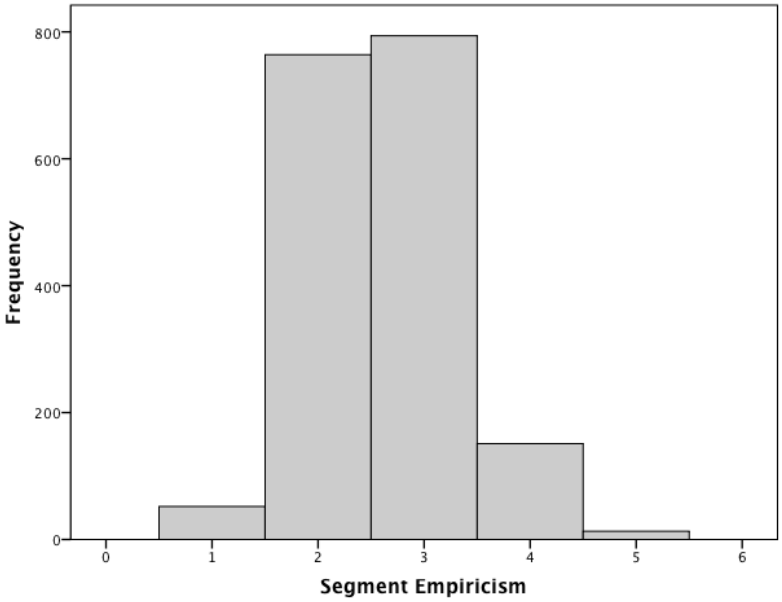


Figure 6. Frequency of Segment Empiricism across all segments in all sessions.

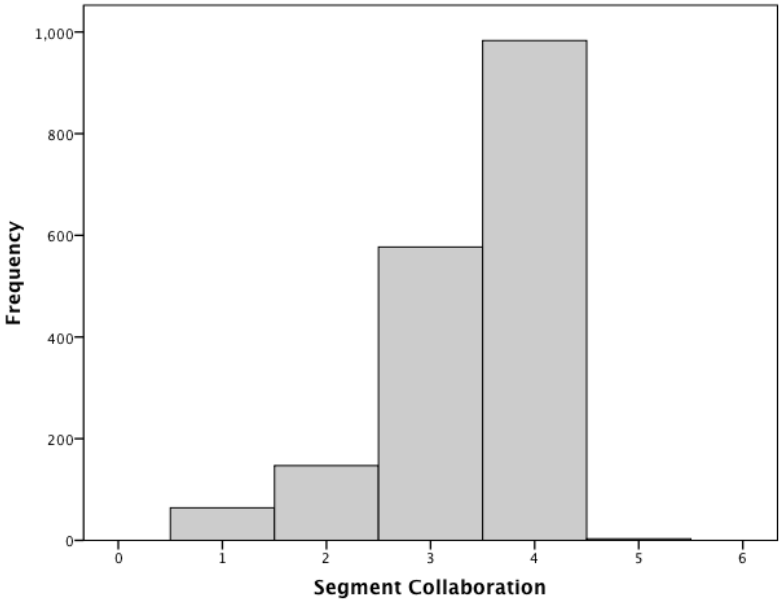


Figure 7. Frequency of Segment Collaboration across all segments in all sessions.

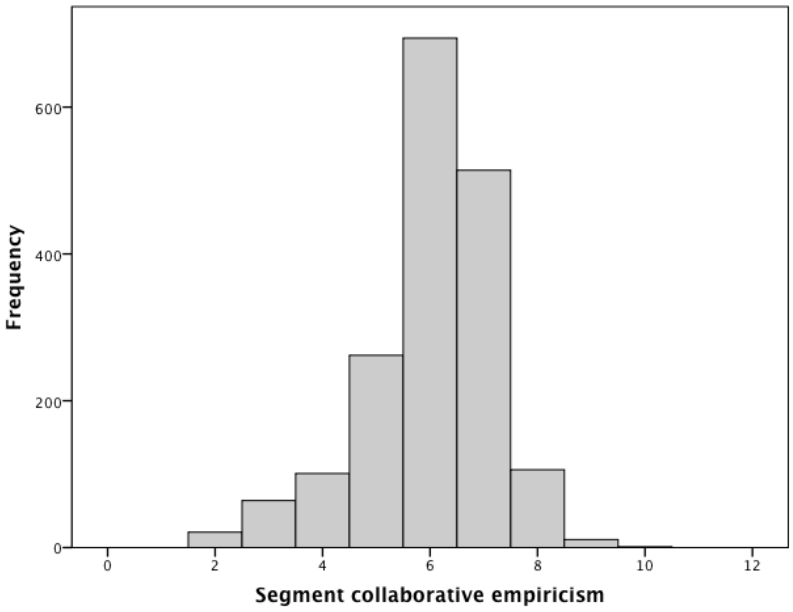


Figure 8. Frequency of Segment Collaborative Empiricism across all segments in all sessions.

As seen in Figures 6, 7, and 8, responses for Segment Empiricism, Segment Collaboration, and Segment Collaborative Empiricism had a restricted range. For Segment Empiricism, 87.9% of responses were either 2 or 3 (out of a range of 5), for Segment Collaboration, 87.9%[4] of responses were either 3 or 4 (out of a range of 5), and for Segment CE, 82.9% of responses were either 5, 6, or 7 (out of a range of 10). Restricted range is likely to reduce the effect size of analyses involving regression or correlation (Lahey, Downey, & Saal, 1983; Sackett, Lievens, Berry, & Landers, 2007; Sackett & Yang, 2000; Vallis et al., 1986). Regression and correlation explain variability in one variable in terms of variability in another. If variability is restricted, it may be more difficult to identify statistically significant correlations, and observed correlations are likely to be lower than in reality (Kaplan & Saccuzzo, 2013). This can consequently impact on calculations of inter-rater reliability and correlations to determine concurrent validity (Tinsley & Weiss, 2000).

[4] Percentages were coincidentally the same for Segment Empiricism and Segment Collaboration.

Reliability

Choice of reliability indices.

Inter-rater reliability is a measure of the consistency of measurements made by several raters measuring the same construct. Reliability of ratings was calculated using intraclass correlation coefficients (ICCs: Shrout & Fleiss, 1979). Intraclass correlation coefficients are widely used in psychological research (McGraw & Wong, 1996), and are recommended as the most appropriate measure of inter-rater reliability for ordinal and interval measurements (Tabachnick & Fidell, 2007; Tinsley & Weiss, 2000). Intraclass correlation coefficients were calculated using the SPSS version 21 two-way random effect, absolute agreement ICC (2, *k*) procedure, where *k* = number of raters.

One disadvantage of ICCs is that they are likely to be reduced when variance between ratings is low (Fagot, 1991; Selvage, 1976; Whitehurst, 1984). In this case, it is recommended to use an alternative index of inter-rater reliability (Tinsley & Weiss, 2000). Finn's *r* is an index of inter-rater reliability developed for use when within-rater variance is low (Finn, 1970; A. P. Jones, Johnson, Butler, & Main, 1983; Tinsley & Weiss, 2000; Whitehurst, 1984). In the present study, reliability will be calculated using both ICCs and Finn's *r*. In the literature, inter-rater reliability is deemed adequate for ICCs above .70 and good for ICCs above .80 (Fleiss, 1986; Kline, 1999). Intraclass correlations coefficients above .75 have been recommended for scales in health research (Streiner & Norman, 1995). Finn's *r* is considered good above .80 and excellent above .90 (Lombard, Snyder-Duch, & Bracken, 2002).

Inter-rater reliability results.

The reliability of ratings was assessed by calculating inter-rater reliability between raters and the criterion rating for each session (rated by the student

researcher). Intraclass correlation coefficients and Finn's *r* for ratings of empiricism and collaboration across all ratings are presented in Table 31. Table 32 presents ICCs and Finn's *r* for each Item (A – F) and Table 33 presents ICCs and Finn's *r* for each rater.

Table 31

Intraclass Correlation Coefficients (ICCs) and Finn's r for Empiricism and Collaboration Across all Ratings

	ICCs	ICC 95% CI	Finn's *r*
Empiricism	.77	[.77, .79]	.958
Collaboration	.79	[.77, .81]	.952

Note. *N* = 1774.

Table 32

Intraclass Correlation Coefficients (ICCs) and Finn's r for Empiricism and Collaboration for Each Item (A – F)

	n	ICCs	ICC 95%CI	Finn's *r*
Item A				
Empiricism	247	.76	[.746, .770]	.952
Collaboration	247	.77	[.759, .783]	.947
Item C				
Empiricism	1472	.75	[.744, .754]	.959
Collaboration	1472	.76	[.757, .768]	.953
Item D				
Empiricism	53	.79	[.759, .820]	.964
Collaboration	53	.75	[.713, .757]	.967

Note. For Item B and Item E, *n* = 1; For Item F, *n* = 0.

Table 33

Intraclass Correlation Coefficients (ICCs) and Finn's r for Empiricism and Collaboration for Each Rater

	n	ICCs	ICC 95%CI	Finn's *r*
Rater 7				
Empiricism	293	.77	[.758, .781]	.958
	293	.75	[.735, .760]	.953
Collaboration				
Rater 8				
Empiricism	813	.74	[.732, .745]	.952
	813	.76	[.752, .765]	.953
Collaboration				
Rater 9				
Empiricism	251	.74	[.743, .771]	.951
	251	.73	[.724, .742]	.950
Collaboration				
Rater 10				
Empiricism	417	.76	[.750, .769]	.956
	417	.80	[.790, .810]	.968
Collaboration				

As seen in Table 31, Finn's *r* was above .95, for both empiricism and collaboration, across all ratings, and ICCs were above 0.75. Tables 32 and 33 show that both Finn's *r* and ICCs maintained this standard across all scale items and all

raters. All 95% Confidence intervals for ICCs were small, indicating that point estimates of reliability could be held with a high degree of confidence.

Concurrent Validity

Predicted Correlations with the CTRS.

Correlations with CES Empiricism.

In order to evaluate concurrent validity for the CES, correlations were examined with the CTRS. From the descriptions attached to the CTRS item anchor points, and the rating instructions in the CTRS manual, it was hypothesized that items 8, 9, and 10 on the CTRS would be most likely to correlate highly with CES Empiricism. Items 8, 9, and 10 focus on the implementation of cognitive-behavioural techniques. Cognitive-behavioural techniques on the CTRS are defined as techniques such as the thought record and behavioural experiment, identifying shifts in mood, and ascertaining the meaning of an event for the client (J. E. Young & Beck, 1980b). All of these techniques are related strongly to exploring cognitions or examining evidence, and should therefore involve empiricism. Item 8 ('Focusing on key cognitions or behaviours') measures the therapist's skill at eliciting and focusing on key thoughts, assumptions, behaviours, etc. that are highly relevant to the client's problems and offer good promise for progress. Item 9 ('Strategy for change') is operationalized as the therapist following a consistent strategy that seems promising and 'incorporates the most appropriate cognitive-behavioural techniques'. Item 10 ('Application of cognitive-behavioural techniques') focuses on the therapist's skill in employing cognitive-behavioural techniques. It also seems likely that Item 1 and Item 6 on the CTRS may correlate with empiricism. Item 1 measures the quality of agenda setting, and Item 6 measures the pacing and efficient use of time. Setting an agenda that focuses on key problems and then efficiently using session time to explore these

problems should facilitate empiricism, and as a result moderately correlate with CES Empiricism. It was thought possible that the CTRS Item 7 (Guided Discovery) may correlate with the CES Empiricism score because one of the four anchor points in Item 7 focuses on 'examining evidence, and considering alternatives', but due to the focus on the therapist's skills in CTRS Item 7 this prediction was considered speculative.

It was hypothesized that Item 3 'Understanding' (a measure of the therapist's empathy and listening skills); Item 4 'Interpersonal Effectiveness' (a measure of the therapist's warmth, concern, confidence, genuineness, and professionalism); and Item 5 'Collaboration' (a measure of the 'therapist's ability to collaborate with the patient') would differentially show low correlations with CES Empiricism as these items have little content focused on empiricism.

Correlations with CES Collaboration.

The CES and the CTRS conceptualize collaboration differently. The CES focuses on the quality of shared work and mutual engagement of the dyad, whereas the CTRS focuses on the therapist's attempts at collaboration with the client. For example, feedback on the CTRS is focused on the therapist's ability to elicit the client's reactions to the session. Feedback on the CES focuses on whether either the client or therapist seek clarification or confirmation about any aspect of the session. If, for example, the therapist is unilaterally directing the session, but also seeks feedback periodically regarding the client's reactions to the therapist's agenda, this would score highly on the CTRS but poorly on CES Collaboration. Similarly, collaboration on the CTRS is weighted towards the therapist's 'attempts to set up collaboration with the patient', whereas collaboration on the CES focuses on the quality of actively shared therapeutic work. A therapist who frequently sought

agreement with the client regarding goals or tasks of therapy, but did not allow the client time to actively contribute to the design and evaluation of therapeutic interventions, could score highly on the CTRS, but would score low on CES Collaboration. Accordingly, it was hypothesized that CES Collaboration and CTRS Item 2 (Feedback) and CTRS Item 5 (Collaboration) would not be significantly correlated.

Table 34 presents Pearson product-moment correlation coefficients for correlations between CES Empiricism and CES Collaboration and each item on the CTRS.

Table 34

Correlations Between CES Empiricism and CES Collaboration and Items on the CTRS

	CTRS Items (Pearson's r, *p*)										
CES	1	2	3	4	5	6	7	8	9	10	11
Emp	.418	.275	.181	-	.194	.318	.258	.494	.395	.399	.286
p	.005*	.074	.246	.022	.212	.038*	.095	.001*	.009*	.008*	.063
				.887							
Col	-.310	-.344	-	.145	-	-.316	-	-.286	-.349	-.258	-
p	.041*	.024*	.115	.353	.043	.039*	.099	.063	.022*	.095	.208
			.464		.782		.526				.181

Note. * = $p < .05$. Emp = CES Empiricism; Col = CES Collaboration.

As seen in Table 34, the hypothesized pattern of differential correlations was observed. CES Empiricism was moderately correlated with CTRS Items 8, 9, 10, 1, and 6, and these correlations were statistically significant. CES Empiricism was not correlated with CTRS Items 3, 4, and 5.

Also as hypothesized, CES Collaboration was not correlated with CTRS Item 5 (Collaboration); however, contrary to prediction, CES Collaboration was correlated with CTRS Item 2 (Feedback). Unexpectedly, the correlation with CTRS Item 2 was negative, indicating that as Feedback on the CTRS increased, CES Collaboration was on average decreased.

Collaborative Empiricism as a Predictor of Reduction in Depression Post-Therapy

A primary aim of the present research was to investigate whether CE could predict depression severity, after controlling for pre-therapy depression, the working alliance, and therapist competence. A linear mixed effects model (Singer & Willet, 2003) was used to examine the relationship between CE and symptom reduction over five post-therapy time points (0, 6, 12, 18, and 24 months post-therapy). Linear mixed effects models explicitly model change over time, offer flexibility in the covariance structure of repeated measures, generalize to non-normal data, and are robust to missing data (Tabachnick & Fidell, 2007; West, Welch, & Galecki, 2007). For comparison purposes, a hierarchical linear regression analysis of the same relationship is presented in Appendix K. The linear mixed model has the advantage that it accounts in a single analysis for the relationship between CE and depression across the five time points post-therapy (0, 6, 12, 18, and 24 months), whereas the hierarchical linear regression is restricted to 0 months post-therapy only. The linear regression has the advantage that it provides an effect size format (ΔR^2) that is readily compared with effect sizes for other processes variables in the literature. The results of the two analyses are in close agreement.

Session sampling strategy.

The session sampling strategy refers to the choice of sessions from the sample to represent CE in the whole sample. In investigating the relationship between a dependent variable such as symptom reduction and a process construct of interest, such as CE, studies have frequently sampled a single session to represent the construct of interest across a course of therapy (McCabe & Priebe, 2004). Single session studies have frequently sampled session three (e.g., Derisley & Reynolds, 2000; Krupnick et al., 1996; Levin, Henderson, & Ehrenreich-May, 2012). A rationale given for this is that by session three the alliance and other relationship based constructs will have stabilized to a level representative of the therapy overall (Lambert, 2004). The utility of single session sampling is called into question by research showing that the alliance is a fluid construct that changes between sessions and within a single session (Arnkoff, 2000; Leahy, 2008; Safran et al., 2011). Further, CE is a skill that is understood to develop with instruction and practice over a course of therapy (J. S. Beck, 1995, 2011). The common dictum in CBT that the client should 'become their own therapist' by the end of treatment (e.g., Ayers, 2007) also reflects an understanding of the therapy as, in part, a process of skill acquisition. This suggests that the assumption that a single session is representative of the overall therapy may be less likely to hold for CE because, as a skill based construct, it is purposefully developed over the course of therapy using a combination of psychoeducation, in session practice, and homework.

Recent research has taken a broader sample of the available sessions to best represent a course of therapy. For example, in a study of the relationship between therapist adherence and competence in CBT and symptom reduction, Boswell and colleagues used a random 15% of sessions in the sample to represent the course of

therapy (Boswell et al., in press). Following this approach, the present study used the average amount of CE that each client and therapist experienced across the three sessions sampled, represented by the mean of the CES scores across all three time points. The average CES score across all sessions sampled has the advantage that it draws from three times the data compared with a single time point. It is also less sensitive to variations related to a specific time point. Accordingly, the average CES score across all sessions was used to represent CE in the sample. In the present data, the mean CES score was stable across time (session 3 $M = 6.216$, $SD = .659$; session 9 $M = 6.227$, $SD = .605$; session 15 $M = 6.261$, $SD = .576$), indicating that the mean CES score did not obscure mean differences between time points.

Research has demonstrated that the working alliance and therapist competence fluctuate between sessions and within a single session (Castonguay et al., 2006; Safran et al., 2011; Stiles et al., 2004). Accordingly, to represent competence and alliance across the sample, and avoid undue influence of localized fluctuations, the working alliance was represented in the regression model by the mean value of the WAI for each client across all time points. Therapist competence was similarly represented by the mean value of the CTRS for each client across all time points.

The relationship between CE and depression severity over time post-therapy.

As a first step in the analysis of the relationship between CE and depression severity, the relationship between CE and depression severity over time post-therapy (0, 6, 12, 18, and 24 months) was examined visually. Figure 9 depicts the relationship between mean BDI over time post-therapy.

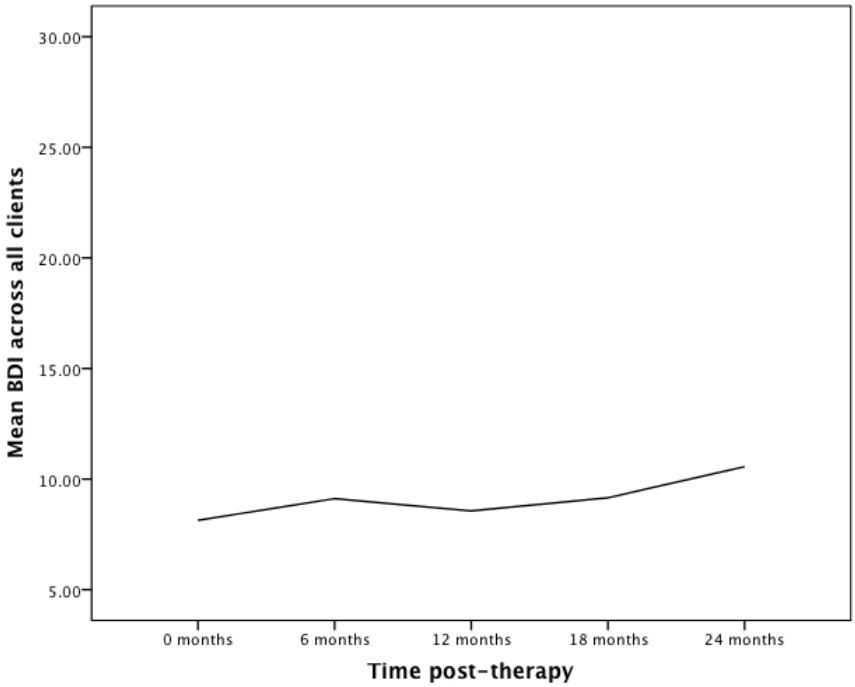


Figure 9 Mean BDI over time post-therapy.

Figure 9 suggests that mean BDI does not vary significantly in the 24 months post-therapy. Figure 10 plots the mean BDI against CES levels for each time point (0, 6, 12, 18, 24 months) post-therapy, and Figure 11 plots the mean BDI across time for each CES level in the sample.

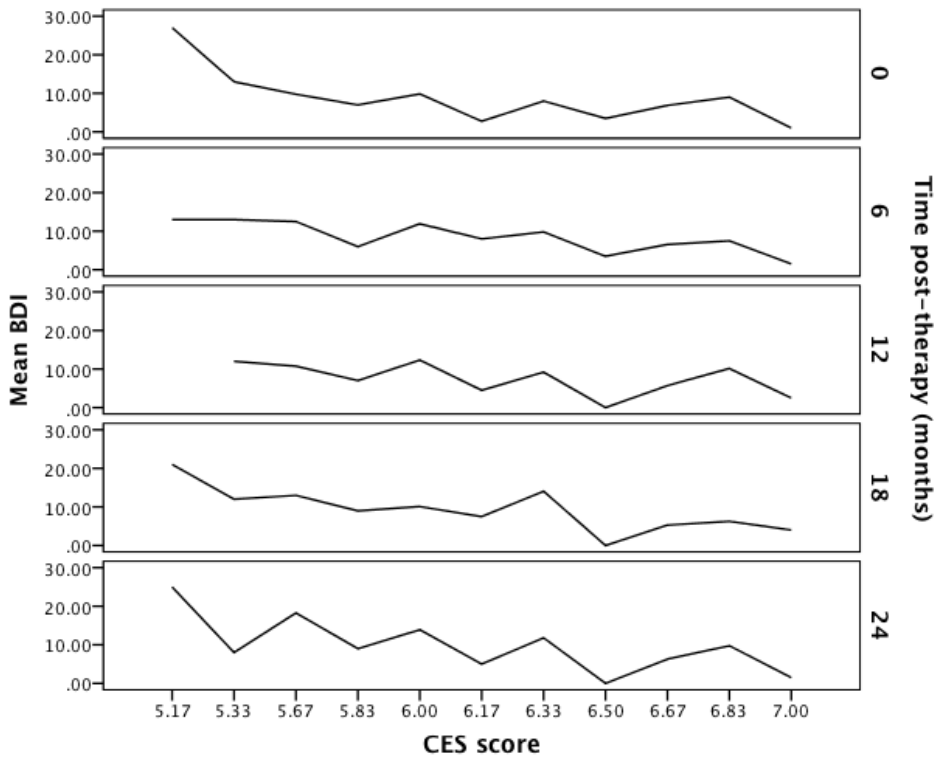


Figure 10. Mean BDI versus CES levels for each time point post-therapy.

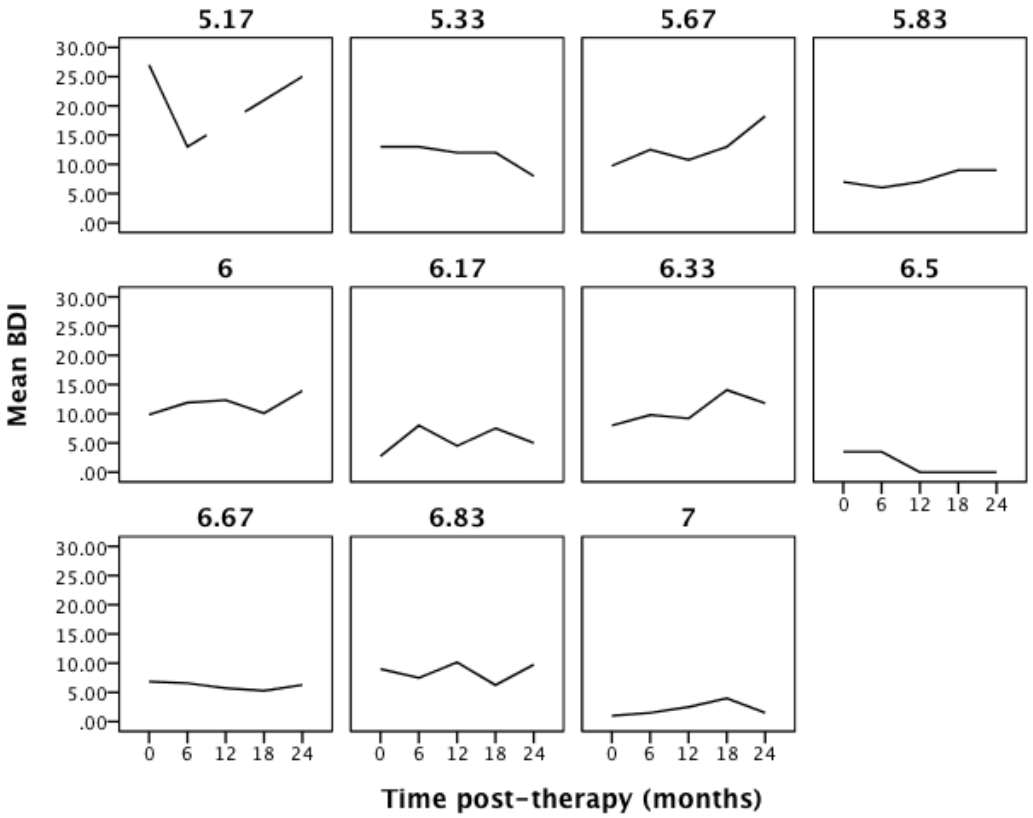


Figure 11. Mean BDI versus time. Each graph represents a CES level in the sample.

Figure 10 shows that (a) for each time point post-therapy, BDI levels on average decrease linearly as CES levels increase, and (b) this relationship between BDI and CES is similar at each time point. Figure 11 shows the effect of CES on the relationship between BDI and time post-therapy. For the highest level of CES scored (CES = 7), BDI was low at termination (0 months) and stayed low. Conversely, for each step CES level was lower, BDI was generally higher at termination and stayed higher across the following 24 months.

Analysis of the relationship between CE and depression severity over time.

A linear mixed effects model was used to examine the relationship between CE and depression severity at 5 time points (0, 6, 12, 18, and 24 months) post-therapy. The model is described by Equation 1:

$$BDI_{ij} = \beta_0 + \beta_1 Time_{ij} + \beta_2 PreBDI_j + \beta_3 WAI_j + \beta_4 CTRS_j + \beta_5 CES_j + \beta_6 PreBDI_j \times Time_{ij} + \beta_7 WAI_j \times Time_{ij} + \beta_8 CTRS_j \times Time_{ij} + \beta_9 CES_j \times Time_{ij} + \varepsilon_{ij} \quad (1)$$

In Equation 1, *i* (= 1, 2, 3, 4, 5) represents each six month time point post therapy (0, 6, 12, 18, and 24 months) and *j* represents the jth client (*j* = 1, 2, …, 44). *Time* is a continuous variable that represents the value of time. The $PreBDI_j$ term represents the pre-treatment BDI score for each client minus the mean pre-treatment BDI score (i.e. PreBDI centred around the mean). Similarly, the WAI_j term represents the WAI score for each client centred around the mean, the $CTRS_j$ term represents the CTRS score for each client centred around the mean, and the CES_{ij} represents the CES score for each client centred around the mean. The *PreBDI* × *Time* term represents the interaction between pre-treatment BDI level and time. Similarly, the *WAI* × *Time* term represents the interaction between WAI level and time, the *CTRS* × *Time* term represents the interaction between CTRS level and time, and the *CES* × *Time* term

represents the interaction between CES level and time. The parameters $\beta_0 - \beta_9$ represent the fixed effects in the model. The parameter β_0 represents the mean BDI when all other parameters in the model are equal to zero, that is, the initial BDI level (i.e. $Time_{ij}$ = 0 months post-therapy) for clients with mean CES and mean PreBDI levels. The parameter β_1 represents the effect of time on the mean BDI level for clients *j* with mean pre-treatment BDI level and mean CES level. The parameter β_2 represents the change in the initial post-therapy BDI level when PreBDI increases by 1 unit, keeping parameter β_3 , β_4, and β_5 levels fixed. Similarly, the parameters β_3 , β_4, and β_5 represent the change in the initial post-therapy BDI level when WAI, CTRS, and CES increase by 1 unit respectively. Parameters β_6 , β_7 , β_8, and β_9 represent the interaction effects between time and pre-treatment BDI levels, time and WAI levels, time and CTRS levels, and CES levels respectively. Parameter β_9 allows the model to determine whether the relationship between CES and post-treatment BDI significantly changes over time. The model assumes that the distribution of the vector of residuals for client *j* is multivariate normal with mean 0 and unstructured covariance matrix (**R**). The unstructured covariance matrix means that the variance of the post-therapy BDI levels for each client can differ at each time point post-therapy, and the covariance between each pair of post-therapy BDI scores can also differ.

The SPSS Version 21.0 MIXED procedure was used to compute the linear mixed effects model given in Equation 1 above. The fixed effects for each parameter in the model are presented in Table 35.

Table 35

Fixed Effects for the Linear Mixed Effects Model of PreBDI, WAI, CTRS, CES, and BDI Post-Therapy

Parameter	Coefficient	95% CI	*p*
Intercept	8.0	[6.40, 9.59]	.000
Time	0.064	[-0.0002, 0.128]	.051
PreBDI	0.20	[-0.067, -0.469]	.137
WAI	0.53	[-0.014, 1.065]	.056
CTRS	0.12	[-0.321, 0.569]	.576
CES	-4.23	[-8.05, -0.469]	.029
Time * PreBDI	0.0066	[-0.0047, 0.018]	.243
Time * WAI	0.019	[-0.0026, 0.040]	.084
Time * CTRS	0.0024	[-0.015, 0.20]	.777
Time * CES	-0.53	[-0.204, 0.097]	.477

As seen in Table 35, CES was the only statistically significant predictor of BDI in the model. The results indicate that for a 1 point improvement in CE, depression severity dropped by 4.23 points, after accounting for the effect of pre-therapy depression, the working alliance, and therapist competence.

Table 35 also shows no effect for time in the model. This indicates that the relationship between CE and depression severity at 0 months post-therapy did not change significantly over the following 24 months. It is noted that for the Time parameter $p = .051$; however, the parameter coefficient was 0.064, indicating that even if an effect of time was considered in the model due to the marginal p value, the effect would be very small. The interaction terms with time were also all non-

significant, indicating that the relationships between pre-therapy depression, working alliance, therapist competence, and CE with post-therapy BDI did not change significantly over time.

Assumption testing of the linear mixed effects model.

The linear mixed effects model assumes that the distribution of the vector of residuals associated with client *j* is multivariate normal, with mean 0 and an unstructured covariance matrix (**R**) (Tabachnick & Fidell, 2007; West et al., 2007). Inspection of histograms showed that the assumption of normality of residuals was satisfied at each time point. The residual covariance matrix (**R**) for the linear mixed effects model of CES and BDI post-therapy is presented in Table 36.

Table 36

*Residual Covariance Matrix (**R**) for the Linear Mixed Effects Model of PreBDI, WAI, CTRS, CES and BDI Post-Therapy*

	Time (months) post-therapy				
Time (months)	0	6	12	18	24
0	40.69	25.12	18.57	21.24	36.27
6		38.44	25.52	23.53	36.90
12			38.56	24.28	22.18
18				41.04	32.87
24					66.01

Table 36 shows that the variability of BDI levels is not constant over time post-therapy, and the correlations between each pair of BDI scores varies. This justifies the use of an unstructured covariance matrix for residuals (**R**).

Collaborative empiricism as a predictor of depression in the next session.

To examine the proximate effect of CE on depression, hierarchical linear regression was used to examine whether CE and depression severity at each time point (early, middle, late) could predict depression severity at the next session. At each time point, inspection of the normal probability plot of standardized residuals and the scatterplot of standardized residuals against standardized predicted values indicated that the assumptions of normality, linearity, and homoscedasticity of residuals were met. Similarly, at each time point, the Mahalanobis distance did not exceed the critical χ^2 for $df = 2$ (at $\alpha = .001$) of 13.816, indicating that multivariate outliers were not a concern. All predictors in each model had tolerances greater than .933, indicating that multicollinearity would not affect the interpretability of the regression analysis.

Three hierarchical linear regressions were conducted, one at each time point (early, middle, late). In each regression, depression and CE at the predictor session (e.g., session 3) were used to predict depression at the next session (i.e., session 4). At session 3, in step 1 of the hierarchical regression, depression at session 3 accounted for a statistically significant 42.8% of the variance in depression at session 4, $R^2 = .428$, adjusted $R^2 = .414$, $F\ (1, 41) = 30.627$, $p = .000$. In step 2, CE accounted for an additional, non-significant, 6.0% of the variance in depression at session 4, $\Delta R^2 = .006$, $\Delta F\ (1, 40) = 0.434$, $p = .514$. The full model of two predictors explained a statistically significant 43.4% of the variance in depression at session 4, $R^2 = .434$, adjusted $R^2 = .405$, $F\ (2, 40) = 15.319$, $p = .000$. Regression coefficients and part correlations (sr^2) for each predictor in the session 3 model are presented in Table 37.

Table 37

Unstandardized (B) and Standardized (β) Regression Coefficients, and Squared Semi-Partial Correlations for the Session 3 Model Predicting Depression at Session 4

Predictor	*B* [95% CI]	*β*	sr^2	*p*
Step 1				
Session 3 depression	0.693 [0.440, 0.946]	.654	.654	.000
Step 2				
Session 3 depression	0.694 [0.439, 0.948]	.654	.654	.000
Collaborative empiricism	-1.132 [-4.604, 2.340]	-0.078	-.078	.514

As seen in Table 37, both the Step 1 and Step 2 models were significant predictors of depression at session 4; however, depression at session 3 was the only significant predictor in either model. This indicates that CE at session 3 did not significantly explain more variance in depression at session 4 than that explained by depression at session 3. As can be seen in Table 37, the *B* coefficient for depression at session 3 was 0.694. This indicates that for a 1 point increase in depression at session 3, depression at session 4 was predicted to rise by 0.694 points. These results indicate that depression at session 3 was a good predictor of depression at session 4, whereas CE at session three was not. This pattern of results was similar at sessions 9 and 15. Analyses for sessions 9 and 15 are presented in Appendix G. These results indicate that CE did not predict depression severity in the next session.

Segment Level Data Analyses

The CES provides data on the start time of each segment in a session and scores for empiricism and collaboration for each segment. The number of segments

per session is a measure of the number of focuses of discussion in that session. Analysis of the number of segments enables examination of the relationship between the quality of CE in a session and the number of separate focuses of discussion in that session. It was hypothesized that there would be a non-linear relationship between the quality of CE in a session and the average number of focuses of discussion in that session. Specifically, it was hypothesized that doing too many things in a session, indicated by a high number of segments, would reduce the overall quality of CE for that session.

Segment characteristics.

Number of segments per session.

The characteristics of the sample were first explored. In total, 1774 segments were rated across the 132 sessions in the sample. Each segment was rated for both empiricism and collaboration, giving a total of 3548 segment ratings. The number of segments per session ranged from 6 – 25 (M = 14.6, SD = 4.0). Kolmogorov-Smirnov tests showed that the number of segments per session was normally distributed at each time point: early, $D(44) = 0.110$, $p > .200$, middle $D(44) = 0.076$, $p > .200$, and late $D(44) = 0.098$, $p > .200$. Figure 12 shows the distribution of the frequency of segments per session across item types (A – F).

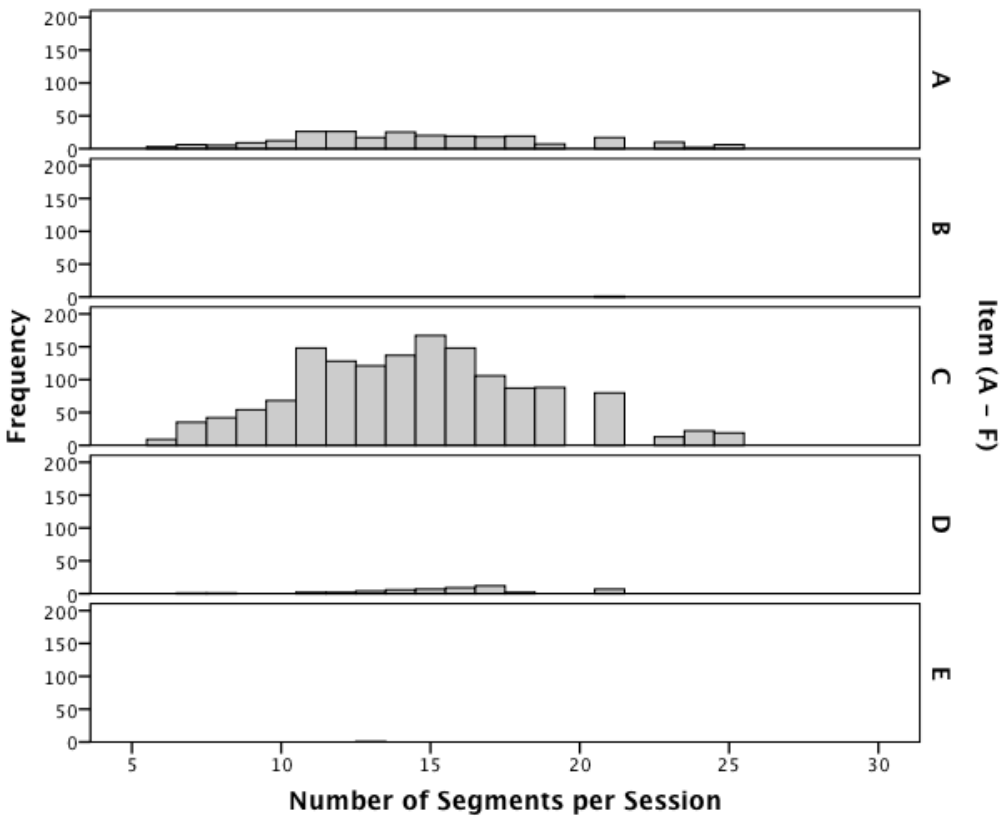


Figure 12. Frequency of segments per session by item type (A – F)

Segment duration.

Segment duration ranged from 2 – 1880 seconds (*M* = 204.40 sec, *SD* = 210.91 sec). The means and standard deviations of segment duration (in seconds) for each empirical item (A – F) are presented in Table 38.

Table 38

Means (and Standard Deviations) of Segment Duration for Empirical Items (A – F)

Scale item (A – F)	Segment duration (seconds) *M* (*SD*)
A. Selecting	48.7 (65.4)
B. Rating the degree of emotion or belief	46.0
C. Exploring	229.4 (216.0)
D. Developing an alternative interpretation	241.5 (204.7)
E. Developing an empirical test	45.0
F. Evaluating an empirical test	-

Note. For Items B and E, $n = 1$.

As seen in Table 38, the mean duration of segments corresponding to Item C ('Exploring') and Item D ('Developing an alternative interpretation') were approximately equivalent in duration and had large standard deviations. Segments corresponding to Items C and D were both four times longer on average than segments corresponding to Item A ('Selecting'). Segments matching Item A tended to be brief (<1 minute). Typically, the content of Item A segments related to choosing agenda items for the session, or consisted of the therapist suggesting a focus of discussion (e.g., reviewing homework), and rhetorically asking whether the client agreed (e.g., T: 'How about we review your homework?' C: 'Ok'). These types of interactions scored low for empiricism and collaboration.

Segment empiricism and collaboration.

Table 39 shows the frequency and percentage of segments matching each empirical item (A – F).

Table 39

Frequency and Percentage of Segments Matching Each Empirical Item (A – F)

Item Type	Frequency	Percentage
A. Selecting	247	13.9
B. Rating emotion or belief	1	0.1
C. Exploring	1472	83.0
D. Developing an alternative interpretation	53	3.0
E. Developing an empirical test	1	0.1
F. Evaluating an empirical test	0	0
Total	1774	100.0

As seen in Table 39, Item C (‘Exploring’) was the most commonly rated item. This was expected as Item C covers the widest range of behaviours. It is notable that Items B, E, and F were essentially missing from the sample. Means (and standard deviations) for segment empiricism, segment collaboration, and segment CE are presented in Table 40.

Table 40

Means (and Standard Deviations) for Segment Empiricism, Segment Collaboration, and Segment Collaborative Empiricism Scores by Item Type (A – F)

		Segment scores		
Item	*n*	Empiricism	Collaboration	CE
A	247	2.21 (0.68)	2.51 (1.12)	4.72 (1.62)
B	1	2	3	5
C	1472	2.65 (0.69)	3.55 (0.62)	6.20 (0.99)
D	53	3.36 (.62)	3.47 (0.58)	6.83 (0.87)
E	1	2	3	5
F	0	-	-	-

Note. Raw scores are reported for Items B and E as $n = 1$ for both items.

As seen in Table 40, segments corresponding to Item A scored lowest for empiricism, Item C scored higher, and Item D highest. Item A was lower for collaboration, whereas Items C and D were comparable. Segment CE scores were lowest for Item A, higher for Item C, and highest for Item D. These results show that 'selecting a focus of discussion' (Item A) was lowest in both empiricism and collaboration. Segments matching Item A characteristically involved quick choices of what to discuss next. These were typically led by the therapist, with little or only moderate input from the client (e.g., the client was invited to agree with the therapists choice of topic). In contrast, segments rated with Item C ('Exploring') involved more empiricism and were much more collaborative, and segments rated with Item D ('Developing an alternative explanation') were highest overall in CE.

Segment duration and collaborative empiricism.

Figure 13 shows a scatterplot of segment CE scores versus segment duration, for each time point (early, middle, late).

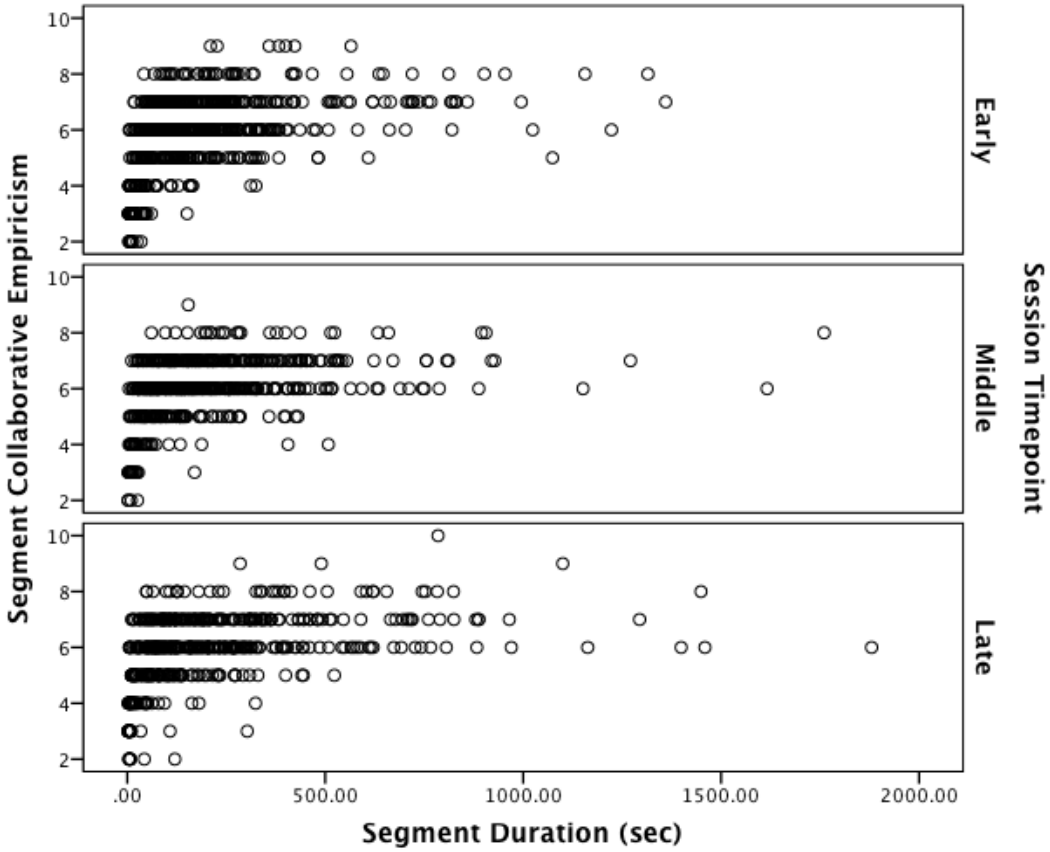


Figure 13. Segment CE scores versus segment duration (in seconds), for each time point (early, middle, late).

As seen in Figure 13, segment CE scores appear to be in a logarithmic relationship with segment duration. Figure 13 shows that, on average, CE scores rise quickly with longer segment durations, and the rate of rise tapers off as segment durations increase.

A fixed effects model was used to further examine the relationship between CE and segment duration within sessions. The model uses ordinary least squares regression to model the relationship between segment CE and segment duration. The assumption for ordinary least squares of equal variance across all error terms is met in

the model because the variances of the error terms $\varepsilon_{ij} - \bar{\varepsilon}_j$ are approximately constant and covariances approach zero (Kennedy, 2008; Tabachnick & Fidell, 2007). The model is described by Equation 2:

$$CES_{ij} - \overline{CES}_j = \beta_1(DUR_{ij} - \overline{DUR}_j) + \beta_2(DUR_{ij} - \overline{DUR}_j)^2 + \varepsilon_{ij} - \bar{\varepsilon}_j$$

(2)

In Equation 2, i (= 1, 2, 3, … , T_j) represents the ith segment, T_j represents the number of segments per client j, and j represents the jth client (j = 1, 2, …, 44). The CES_{ij} term represents the CES score for the ith segment for the jth client . The $\overline{CES}_j$ term represents the mean CES score for each client j averaged over each segment for that client. The $CES_{ij} - \overline{CES}_j$ term therefore represents the CES score for segment i for client j, centred around the mean of the segment CES scores for that client. Similarly, the $(DUR_{ij} - \overline{DUR}_j)$ term represents the duration of the ith segment for client j, centred around the mean of the segment durations for that client, and the $(DUR_{ij} - \overline{DUR}_j)^2$ term represents the quadratic effect of duration on CES. This term was included because the logarithmic relationship between segment CES and segment duration revealed in Figure 13 above suggests a quadratic relationship between segment duration and segment CES. The $\varepsilon_{ij} - \bar{\varepsilon}_j$ term represents the error term for segment i for client j, centred around the mean of the error terms for that client. The parameter β_1 represents the change in mean centred CES when mean centred duration increases by 1 unit. The parameter β_2 represents the quadratic effect of mean centred duration on mean centred CES.

To investigate the relationship between segment CE and segment duration, the SPSS Version 21.0 REGRESSION procedure was used to compute the fixed effects model given in Equation 2 above. The overall model accounted for a statistically significant 25.9% of the variance in segment CE, R^2 = .259, F (2, 629) = 109.986, p =

.000. Inspection of the normal probability plot of standardized residuals and the scatterplot of standardized predicted values against standardized residuals indicated that the assumptions of normality and homoscedasticity of residuals were met. The Mahalanobis distance exceeded the critical χ^2 for *df* = 2 of 13.82 (at α = .001) for 9 of the 632 cases (0.014%), indicating that these cases were multivariate outliers. The analysis was repeated with these cases removed, with no changes in statistical significance and only very minor adjustments to parameters. Accordingly, as these were clinical data, these cases were retained unchanged. Fixed effects for each parameter in the model are presented in Table 41.

Table 41

Fixed Effects for the Model of Segment CES and Segment Duration Within Sessions

Parameter	Coefficient	95% CI	*p*
Intercept	.152	[0.56, 0.247]	.002
Segment Duration mean centred	0.276	[0.238, 0.314]	.000
Segment Duration mean centred Squared	-0.016	[-0.020,-0.011]	.000

As seen in Table 41, both segment duration and segment duration squared were significant predictors of segment CES. This confirms the logarithmic nature of the relationship between segment duration and segment CES seen in Figure 13 above.

Characteristics of Collaborative Empiricism Within Sessions

It was also of interest to explore the characteristics of CE within sessions. Figures 14 and 15 show scatterplots of each rating of empiricism and collaboration respectively, by Item (A – F), within all sessions. Figure 16 shows a scatterplot of each rating of collaborative empiricism, by Item (A – F), within all sessions.

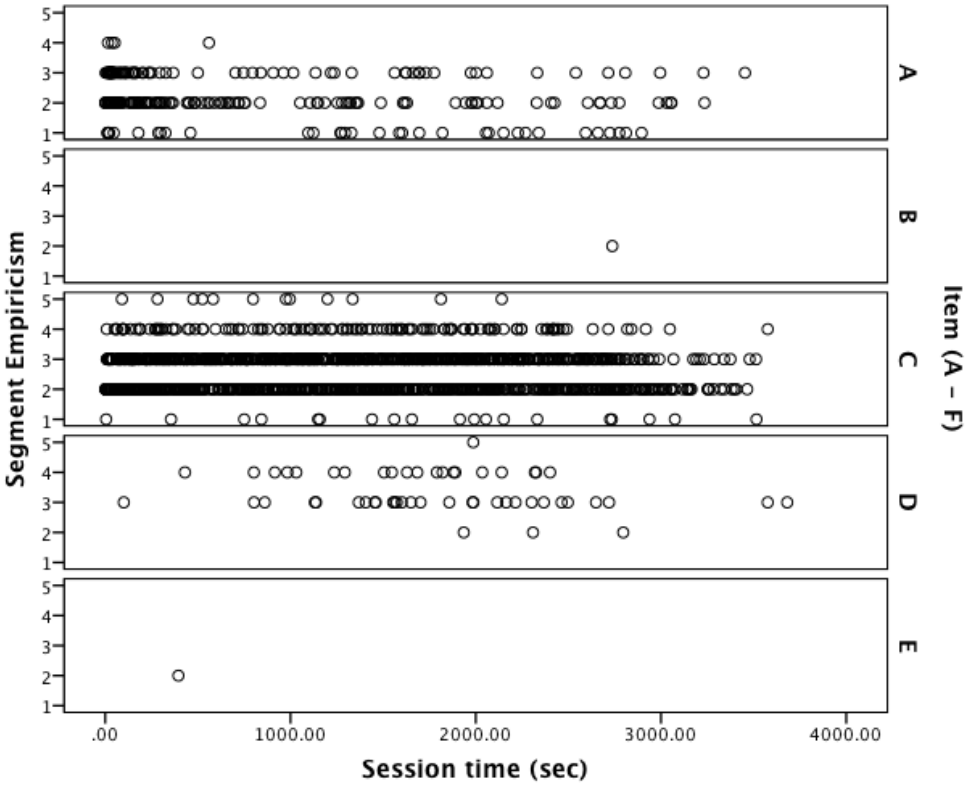


Figure 14. Scatterplot of each rating of empiricism, by Item (A – F), within all sessions.

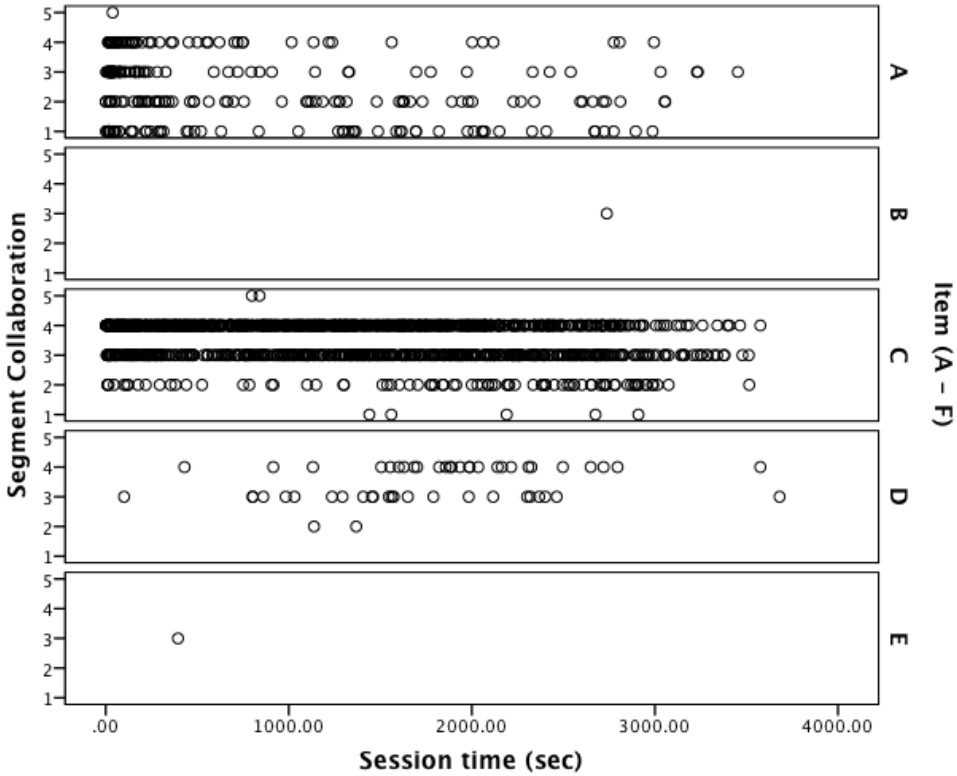


Figure 15. Scatterplot of each rating of collaboration, by Item (A – F), within all sessions.

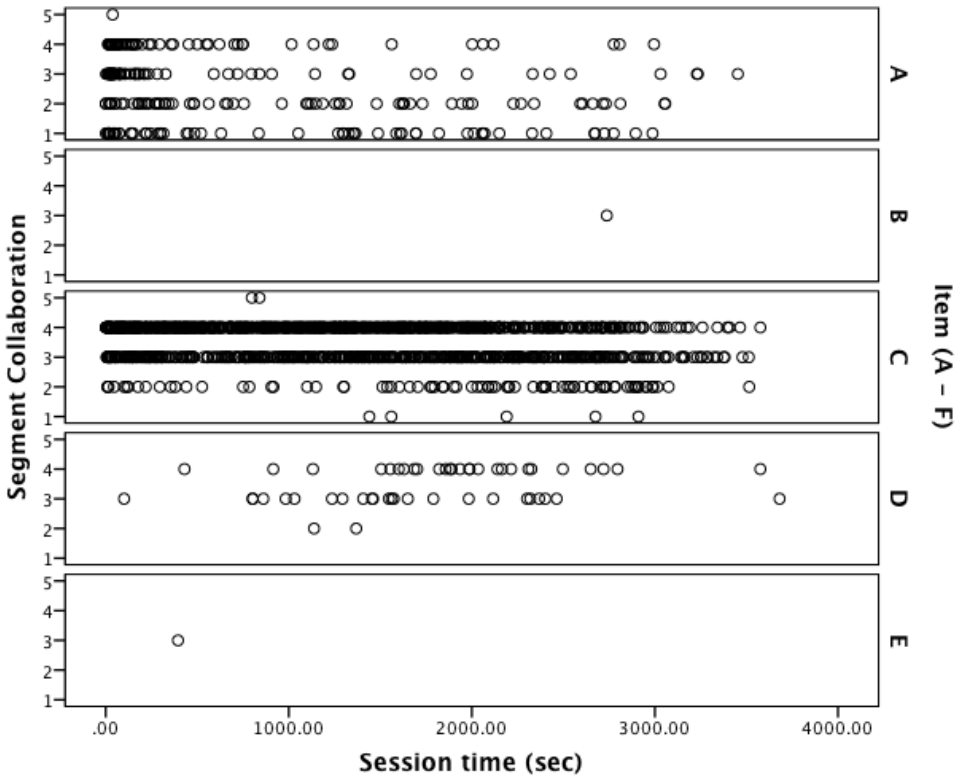


Figure 16. Scatterplot of each rating of collaborative empiricism, by Item (A – F), within all sessions.

Figure 14 shows differences in the pattern and range of scores within sessions for segments matching different types of empirical events. Segments matching Item A (‘Selecting a Focus of Discussion’) show many ratings of ‘1’ for empiricism, indicating that the client’s experience was not mentioned during the segment. Segments matching Item A also show few ratings of ‘4’ and none of ‘5’ for empiricism. This shows that selecting a focus of discussion was in general low in empiricism and that frequent selections were made with no reference to the client’s experience at all. In contrast, segments matching Item D (‘Developing an Alternative Explanation’) showed no scores of ‘1’ and only three scores of ‘2’ for empiricism, indicating that this activity was generally moderate to high in empiricism. Segments matching Item C (‘Exploring the Focus of Discussion’) were overall between scores for segments matching Item A and Item C.

Figure 15 shows that collaboration was more evenly distributed across item types, although the pattern for empiricism was also observed here. Several segments matching Item A (‘Selecting’) showed poor collaboration (score of ‘1’), whereas only two of the segments focused on Item D (‘Developing an Alternative Explanation’) scored lower than 3. The patterns revealed for empiricism in Figure 14 and collaboration in Figure 15 are carried over to CE in Figure 16.

Discussion

The main aims of Study 3 were to determine if the CES could be reliably rated in a larger sample, assess concurrent validity, and use the measure to examine whether CE predicts reduction in depression in a sample of CBT for depression. A secondary aim was to explore the process of CE within a session.

Psychometric Evaluation

Reliability.

The hypothesis that a measure of CE could be reliably rated was supported. Finn's *r*, a measure of reliability that is robust to restricted variability of ratings, was above .95 across all ratings, and this level was maintained across each scale item, and each rater. Intraclass correlation coefficients were also good across all ratings, and for each scale item, and each rater, indicating consistently good reliability and confirming Finn's *r*. These results demonstrate the reliability of the CES for this sample.

Concurrent validity.

CES Empiricism.

The concurrent validity of the CES was assessed by examining correlations between CES Empiricism and CES Collaboration and selected items of the CTRS. The hypothesis that CES Empiricism would be moderately correlated with CTRS Items 8, 9, 10, 1, and 6 was supported. The hypothesis that CES Empiricism would not correlate with CTRS Items 3, 4, and 5 was also supported. The speculative hypothesis that CES Empiricism would correlate with CTRS Item 7 was not supported. Although CES Empiricism was correlated with CTRS Item 7 at .258, this correlation was not statistically significant, $p = .095$.

CTRS Items 8, 9, 10 focus on the implementation of cognitive-behavioural techniques, including identifying key cognitions and using empirically based interventions such as the thought record and the behavioural experiment to explore these cognitions. CTRS Items 1 and 6 focus on structuring the session to focus on the most important cognitions and allow sufficient time to work on them productively. These items are the most related to empiricism on the CTRS. The moderate correlation between CES Empiricism and each of these items indicates that CES

Empiricism is tapping into these empirically loaded items, but is distinct from the specific competencies captured by them. CTRS Items 3, 4, and 5 focus on the therapist's 'understanding', 'interpersonal effectiveness', and 'collaboration' respectively. These attributes are low in empiricism. The lack of correlation between CES Empiricism and these CTRS items indicates that CES Empiricism is not tapping into these concepts. This pattern of differential correlations between CES Empiricism and the CTRS supports the concurrent validity of the CES.

CES Collaboration.

The hypothesis that CES Collaboration would not be correlated with the CTRS collaboration item (Item 5) was supported. Collaboration on the CES and the CTRS are very different constructs. For example, a frequently observed situation in the present sample consisted of high consensus but low shared work. High consensus in a session will score high on the CTRS collaboration item, but will score low on CES Collaboration unless there is shared work involving active involvement, balanced contributions, and shared decision making. The lack of correlation between CES Collaboration and the CTRS collaboration item supports the divergent validity of the CES.

CES Collaboration was unexpectedly correlated with CTRS Item 2 (Feedback). The direction of this correlation was negative, indicating that, on average, CES Collaboration was higher when CTRS Feedback was lower. This finding may be due to different operalizations of feedback on the CES and the CTRS. On the CTRS, feedback is focused solely on the therapist's ability to elicit the client's 'reactions to the session' and adjust his or her responses accordingly. On the CES, feedback focuses on whether either the client or therapist seek clarification or confirmation about any aspect of the session. If, for example, the therapist is unilaterally directing

the session, but also seeks feedback periodically regarding the client's reactions by seeking acknowledgement of the client's agreement, this would score highly on the CTRS Item 2 (Feedback) but poorly on CES Collaboration. This may explain the negative correlation observed. It is also possible that the restricted response range on the CES may have effected the size and direction of the observed correlations (Lahey et al., 1983; Sackett et al., 2007; Sackett & Yang, 2000; Vallis et al., 1986).

The Relationship Between Collaborative Empiricism and Depression Severity Post-Therapy

The hypothesis that CE would predict reduction in depression severity above that explained by pre-therapy depression, the working alliance, and therapist competence was supported. A one point increase in CE significantly predicted a 4.23 point reduction in depression severity post-therapy. This result shows that a difference in CE scores the size of the range of CES scores in this sample (1.83 points), would predict a decrease in depression severity on the BDI of 7.74 points. Using the standard clinical interpretation of scores on the BDI, the difference between mild depression (below 19 points) and severe depression (above 29 points) is 10 points. The range of CE in this sample is therefore associated with an improvement in depression equivalent to a reduction from the low severe range on the BDI to two points above mild depression, a clinically significant improvement. The clinical significance of this result may be further increased due to the importance and difficulty of treating depression (Ferguson, 2009; Thompson, 2002).

The hypothesis that CE would continue to predict depression severity across the 24 months post-therapy was supported. The relationship between CE and depression severity remained constant across the 24 months post-therapy. This result is consistent with previous findings of the so called 'prophylactic effect' of CBT, in

which CBT is associated with a reduction in relapse rates post-therapy (Cuijpers et al., 2013; Evans et al., 1992; S. D. Hollon et al., 2005).

The results obtained here demonstrate empirically, for the first time, the importance of CE as a psychotherapy process. These results support the fundamental importance ascribed to CE in the seminal CBT literature of the past 40 years (A. T. Beck et al., 1979; J. S. Beck, 2005, 2011; Kuyken et al., 2009; Persons, 2008; van Oppen, 2004).

Collaborative empiricism as a predictor of depression in the next session.

Collaborative empiricism was not found to predict depression in the next session. This may be due to the sensitivity of depression to session-by-session fluctuations in process variables. Previous research has found evidence for such fluctuations in treatment (e.g., 'sudden gains') (DeRubeis et al., 2005; Tang & DeRubeis, 1999; Tang, DeRubeis, Beberman, & Pham, 2005; Thomas & Persons, 2013). A second explanation relates to session sampling. Collaborative empiricism was sampled at three time points across the course of therapy. The nature of CE as a skill that develops across therapy suggests that sampling every session may have revealed a trend in the relationship between CE and depression that was not seen by sampling three sessions. A related point concerns the assessment timeframe of the BDI. The BDI-I is a self-report measure of depression over the past week. A week between ratings of depression allows for potentially many other effects on the client's mood. These effects may confound or obscure the effect of a process variable, such as CE, on BDI.

Lack of relationship between of therapist competence and outcome.

No relationship was found between therapist competence and outcomes in this sample. Variability in the relationship between therapist competence and outcome has

been observed in previous research (e.g., Boswell et al., in press; Webb, DeRubeis, & Barber, 2010; Webb et al., 2012), and the effect of therapist competence on outcome has been found to depend on client characteristics such as degree of anxiety and chronicity of depression (Strunk, Brotman, DeRubeis, & Hollon, 2010). More specifically, the lack of relationship between competence and outcomes in this sample may be due to a close focus in the original study on the adherence of therapists to the therapy protocol (Jacobson et al., 1996). Therapists in the original Jacobson et al. study received training for a year on the therapy protocol and were closely monitored during the trial, with good results for adherence, suggesting that the focus on adherence was a success. Although a distinct construct, adherence overlaps with competence (Jacques P. Barber, Crits-Christoph, & Luborsky, 1996). A consequence of the close focus on adherence, is that therapists' competence may have been restricted in range by the need to closely adhere to the protocol. Restriction in range may be responsible for the lack of statistical association between competence and outcomes (Lahey et al., 1983; LeBreton, Burgess, Kaiser, Atchley, & James, 2003; Tinsley & Weiss, 2000). The focus on adherence in this sample may limit the external validity of the present data, a common problem for clinical research (Finger & Rand, 2008; K. H. Morrison, Bradley, & Westen, 2003).

Collaborative Empiricism Within Sessions

Segment duration and CE.

A further aim was to examine CE within individual therapy sessions. The finding of a positive logarithmic relationship between segment duration and CE within each client-therapist dyad shows that segments of very short duration have generally poor CE, and that CE increases quickly with duration and then increases more slowly for segments of very long duration. This demonstrates that CE will on

average be maximized in a session by concentrating effort on a smaller number of separate focuses of discussion, and conversely, that CE is likely to be less than optimal if there are too many separate focuses of discussion in a session. This is consistent with clinical recommendations in training to avoid attempting too much in a single session (e.g., J. S. Beck, 2011).

Characteristic differences in CE between types of session activities.

Different types of in-session activities were characteristically associated in this sample with different patterns of empiricism and collaboration. Activities involving selecting (Item A) were in general low in empiricism and collaboration, activities involving exploring an issue or cognition (Item C) were higher, and activities involving developing an alternative interpretation (Item D) were highest in CE. It makes sense that an activity in which the client is encouraged to develop an alternative interpretation of a thought or belief (Item D) would be generally highest in empiricism, followed by the less structured and more general activity in exploring issues (Item C). Also, therapists are generally trained specifically to be collaborative and empirical when developing alternative interpretations in thought records, by eliciting alternatives from the client (J. S. Beck, 1995, 2005, 2011; Padesky & Greenberger, 1995). The results obtained reflect this pattern, further supporting the validity of the CES.

Implications for Research

The therapeutic relationship in CBT.

The results obtained here raise the question of how the relationship in CBT is defined, and how it is differentiated from the working alliance. The working alliance is the most frequently employed account of the therapeutic alliance in psychotherapy process research (Horvath, 2001; Horvath & Greenberg, 1989, 1994; Horvath &

Luborsky, 1993; Horvath & Symonds, 1991; Martin et al., 2000). It has been regarded as an important psychotherapy process variable (Horvath, 1994; Horvath & Greenberg, 1994; Horvath & Luborsky, 1993; Horvath & Symonds, 1991), perhaps the most important (Lambert & Barley, 2001; Lambert & Ogles, 2004), because it has been reliably associated in meta-analyses with between 4 – 7% of the variance in psychotherapy outcomes (Horvath et al., 2011; Martin et al., 2000; Shirk & Karver, 2003; G. S. Tryon, Blackwell, & Hammel, 2007, 2008). The hierarchical linear regression reported in Appendix K allows ready comparison of an effect size for CE as a predictor of reduction in depression severity with published effect sizes for the working alliance and other process variables in the literature. The effect size obtained in the present sample for CE as a predictor of reduction in depression (8.8%), is at least comparable with that of the working alliance (4 – 7%), and at best double in size. The effect size obtained for CE is also comparable with effect sizes reported in meta-analyses of other relationship attributes as predictors of outcome, including empathy (9.6%) (Elliott, Bohart, Watson, & Greenberg, 2011), cohesion in group therapy (6.25%) (Burlingame, McClendon, & Alonso, 2011), and positive regard (7.3%) (Farber & Doolin, 2011). That the effect size obtained for CE here is equivalent to other predictors of outcome suggests that the list of evidence-based relationship attributes may not be complete, and that CE is of comparable importance to those previously studied.

The working alliance.

As discussed in earlier chapters, collaboration is characterized in the working alliance and the WAI as consensus on the goals of therapy and agreement on the tasks required to achieve those goals. As such, the working alliance does not capture collaboration as the active sharing of creative therapeutic work in CE, nor the

empiricism that is central to the CE construct. The CES was specifically designed to measure empiricism, and those aspects of collaboration in CE that are not captured by the working alliance. The result that CE significantly predicts depression post-therapy, with a clinically significant effect size, after accounting for pre-therapy depression, the working alliance, and therapist competence, suggests that our understanding of the therapeutic relationship in CBT needs to be expanded to account for CE. Previous research has frequently assumed that the therapeutic relationship has been adequately accounted for by the working alliance. The results obtained here suggest that this assumption may be unwarranted. If so, it is possible that accounting for the effect of CE on outcomes, beyond that of the working alliance, may alter the conclusions of previous research.

The APA Task Force on Evidence-Based Therapy Relationships.

The results obtained here are also relevant for the findings of the APA Task Force on Evidence-Based Therapy Relationships (Norcross, 2001; Norcross & Lambert, 2011). The Task Force conducted a series of meta-analyses of psychotherapy outcome research studies, to identify evidence-based attributes of the therapeutic relationship. Based on the evidence available at the time, the Task Force concluded that (a) the therapy relationship accounts for psychotherapy outcomes at least as much as particular treatment methods, and (b) not including the relationship in efforts to promote evidenced-based practice is “seriously incomplete and potentially misleading” (Norcross & Wampold, 2011, p. 98). The Task Force recommended that future research should review the Task Force’s findings and include new elements of the psychotherapy relationship (Norcross & Wampold, 2011). The present research has taken up the Task Force’s recommendation to investigate new elements of the relationship in CBT.

From the data available to it, the Task Force found empirical support for three aspects of collaboration in psychotherapy: ‘collecting client feedback’, ‘goal consensus’, and ‘collaboration’. As discussed in Chapter 2 of this thesis, the studies reviewed by the Task Force operationalized collaboration and goal consensus in terms of the agreement or compliance of the client with the agreed upon goals of therapy, and the tasks needed to reach these goals. Also, ‘collecting client feedback’ was operationalized as the therapist collecting outcome data from symptom measures for the purpose of evaluating the effectiveness of interventions. These conceptions of collaboration do not capture the active sharing of the work, and wide-ranging concept of feedback, that distinguish collaboration in CE. As a result, the findings reported here, of a relationship between CE and therapy outcome, relate to different aspects of collaboration (and also empiricism), and are distinct from the Task Force’s findings.

The Core Competencies Framework.

The UK Core Competences Framework (Roth & Pilling, 2008) is a set of clinical practice guidelines designed to specify the competencies required to deliver effective CBT for people with depression and anxiety disorders. The Framework has been endorsed by the UK National Health Service as a reference for the specific activities required to carry out CBT in accordance with best practice (Roth & Pilling, 2007). The Framework has also been frequently recognized in the literature as providing a ‘comprehensive definition of CBT competence’ (Muse & McManus, 2013). It is notable in this context that empiricism and active collaboration as defined in CE are both absent from the Framework. The Framework lists examples of basic CBT competencies related to collaboration (e.g., sharing responsibility for session structure), and has a section focusing on Socratic Dialogue and guided discovery; however, empiricism is almost completely absent. Empiricism is implicit in several of

the therapist behaviours described (e.g., 'the ability to use thought records', 'identifying key cognitions') but these behaviours need not be empirical, and the issue of the quality, or even presence, of empiricism as an independent construct in these or other behaviours is not discussed. In particular, the specific behaviours captured by several of the CES empirical items, such as rating beliefs and developing or evaluating empirical tests, are absent. It is also notable that the case examples given to illustrate the Competencies Framework would probably score highly on the Cognitive Therapy Rating Scale, and yet empiricism is ignored in them. The Task Force's recommendation to explicitly address competencies that promote evidence-based attributes of the therapeutic relationship suggests that the Core Competencies Framework may be usefully updated with reference to CE.

Collaborative Empiricism as a Mediator of Skill Acquisition and use in CBT.

Previous research has examined the relationship between the client's use of therapy skills and therapy outcomes. The use early in therapy of CBT techniques involving concrete skills, such as identifying negative automatic thoughts, has been shown to predict reduction in depression, whereas more abstract techniques such as psychoeducation regarding the CBT model have not (DeRubeis & Feeley, 1990; Feeley, DeRubeis, & Gelfand, 1999). The degree to which clients are able to learn and implement CBT skills in session has also been related to reduction in depression severity post-therapy (Barber & DeRubeis, 1992, 2001), and reduction in relapse rates at 12 months post-therapy (Strunk, DeRubeis, Chiu, & Alvarez, 2007).

What is not clear from this research is what contributes to the development and implementation of these therapy skills. The theoretical model of CE proposed in Chapter 1 suggests that CE increases intrinsic motivation for change by meeting the basic needs for autonomy, competence, and relatedness described by Self-

Determination Theory. The model proposes that CE does this via actively shared work with the client's intrinsic data. This actively shared work takes precisely the form of the concrete skills involved in the collection and evaluation of empirical data (e.g., identifying negative automatic thoughts, developing alternative interpretations of events) described in the research above. The finding here that CE predicts reduction in depression post-therapy suggests that CE may mediate the acquisition and employment of concrete skills in CBT.

Implications for Practice

The APA Task Force recommended that guidelines for practice should explicitly address therapist competencies that promote evidence-based attributes of the therapeutic relationship, and provide competency based training in elements of the relationship found to be demonstratively effective (Norcross & Wampold, 2011). The results reported here provide the first empirical confirmation of CE as an evidence-based attribute of the therapeutic relationship. This suggests that CE may be an important therapist competency for inclusion in future training and clinical guidelines.

A review of the audio of segments matching Item A ('Selecting') revealed that these segments were generally low in CE because therapists were highly directive, made unilateral decisions, and did not give the client time to contribute. This suggests that clinical training may usefully review skills with collaboration and empiricism when applied to the activity of selection, for example of goals, interventions, or agenda items for the session.

It was found that Items B, E, and F were essentially missing from the sample, indicating that although clients and therapists did thought record work, and developed alternative explanations for thoughts, they did not rate the degree of emotion associated with thoughts, or develop empirical tests of the thoughts examined. This

was surprising as these behaviours are consistently identified as core aspects of CE in the seminal literature (e.g., A. T. Beck et al., 1985; A. T. Beck et al., 1979; J. S. Beck, 2005, 2011) and these items were scored very highly for content validity in the expert review of the scale. The absence of the development and evaluation of empirical tests is particularly surprising as these are frequently described elements of core CBT techniques such as cognitive restructuring and testing cognitions in homework assignments (Bennett-Levy et al., 2004; Kazantzis, Deane, & Ronan, 2000; Kazantzis et al., 2005). The lack of these core aspects of CE in the sample suggests that it may be helpful for clinical training to evaluate whether key components of CE are present in practice. In addition, the lack of core aspects of CBT practice in the sample suggests that the restricted range of scores on the CES is an accurate reflection of characteristics of the sample, rather than a failure of raters to use the full response range of the scale.

Limitations.

A limitation of the present research was the restricted variance in CE scores in the sample. As discussed above, it seems reasonable to conclude that the low variance in CE scores reflects the CE skills and training of the therapists in the sample. Nevertheless, the restricted range of CE scores may have reduced the size of correlations in analyses (Tabachnick & Fidell, 2007; Tinsley & Weiss, 2000). A reduction in the size of correlations will result in an underestimation of the size of the relationship between CE and outcomes, potentially obscuring a greater role for CE in the reduction of depression severity. A second effect of restricted variance is that assessing concurrent validity via correlations of the CES with other process measures is made more difficult. The pattern of differential correlations obtained for the CES with items from the CTRS may have looked different with less restricted variance.

Another limitation was the sampling of CE at three time points across the course of therapy. Although three time points is a significant improvement compared with single session sampling, rating CE at every session would enable a more fine grained analysis of the construct, for example, in understanding the proximal effect of CE on depression and on sudden gains in therapy. A related difficulty encountered in this study was the absence of other process measures at the session level. This meant that segment level data had to be aggregated to the session level in order to analyse the relationship between CE and other process constructs. For example, the data revealed differences between therapists use of CE within sessions, but this could not be related to measures of client mood or the therapeutic relationship at the segment level. Future research may benefit from more fine grain measures of process constructs.

Future research.

As the first empirical study of CE, it will be important for future research to replicate the study and extend it into wider contexts, and to address the limitation of restricted response range in the present sample. It would be useful to examine the relationship between client characteristics, such as suitability for short term CBT and comorbid anxiety, and the relationship between CE and therapeutic outcomes. Assessment of client characteristics would aid the analysis of the effectiveness of CE in more diverse clinical contexts. Future research could also usefully examine clinical trainees' capacity to segment therapy sessions into empirical events, and evaluate whether this ability differentiates skills on other relationship attributes, such as therapist competency. Previous research showing that concrete CBT skills are related to therapeutic outcomes suggests that future research could also examine the role of CE as a mediator of the use of skills in CBT.

Conclusion

This study used the CES to rate sessions of CBT for depression. The CES was found to have good to excellent reliability in this sample. Concurrent validity was demonstrated in a predicted pattern of differential correlations with the CTRS. Collaborative empiricism was found to predict reduction in depression severity post-therapy, to a statistically and clinically significant degree, after accounting for the effect of pre-therapy depression, working alliance, and therapist competence. The effect of CE on outcomes was maintained for 24 months post-therapy, and was comparable or greater in size to that reported in the literature for the working alliance and other relationship constructs. Different types of in-session activities were characteristically associated with different patterns of CE, further supporting the validity of the CES. These results suggest that our current understanding of the therapeutic relationship in CBT should be revised to include CE, and that accounting for the effect of CE may alter the results of previous alliance research. It is suggested that future research could usefully examine the effect of client characteristics on CE, clinical trainees competency in CE in relation to other competencies, and the role of CE as a change process, for example, as a mediator of the effect of CBT skills on therapeutic outcome.

Appendix A: The Collaborative Empiricism Scale

CES

Collaborative Empiricism Scale

John M. Tee
Nikolaos Kazantzis, Ph.D.
Arthur A. Stukas, Ph.D.

June, 2012
Scale Currently Under Evaluation
please do not distribute

Cognitive Behaviour Therapy
Research Unit

Appendix A: The Collaborative Empiricism Scale (cont.)

CES The Collaborative Empiricism Scale

The Collaborative Empiricism Scale (CES) is intended for use wherever independent observer, or supervisor, ratings of this core therapeutic attribute are to be assessed. It is specifically intended for use in research studies of Beck's cognitive therapy, either in terms of process studies, or studies attempting to identify processes that facilitate hypothesized mechanisms of change. The CES will also have utility in clinical training, supervision, and ongoing professional development for practitioners. A core feature of the CES is its focus on the client – therapist dyad, rather than the specific skill of the therapist in maintaining the therapeutic relationship. In this preamble, key constructs embedded in the scale are outlined, along with specific guidance on rating scale terminology, and ratings.

Key Constructs:

(a) Collaboration

Collaboration in Beck's cognitive therapy relates to a genuine *sharing of therapeutic work*. This involves the client and therapist both actively *contributing* to the therapy session in a combined effort, involving *shared decision making* and *mutually responsive interaction*. Examples include sharing in the design and implementation of agenda items, goals, and interventions. Collaboration is more than just client and therapist agreement, cooperation, or client adherence/ compliance with therapist suggestions.

(b) Empiricism

Another core principle of Beckian cognitive therapy is the focus on the client's unique experience. In the context of facilitating cognitive change, the usefulness and functionality of the client's thoughts and beliefs are explored in terms of the client's experience, rather than by analyzing the client's thinking from general principles or on logical grounds. Content and process of cognition are explored with reference to problem situations in which they characteristically occur, and with reference to the accompanying emotional experience. This involves exploration of the client's feelings, behavior, and experiences as they relate to a specific cognition being explored.

Empiricism then, guides the design, implementation, and evaluation of therapeutic work that has been specifically designed to evaluate the client's thoughts and thought processes, and is personally meaningful for the client. Importantly, collaborative empiricism involves more than evaluating whether a thought is true or false. The aim is to foster the client's exploration of the client's thoughts, feelings, and behaviors, in empirical terms, and for the client to determine the usefulness and functionality of their thinking by examining the consequences.

(c) Identifying the Focus of Discussion

The Collaborative Empiricism Scale is centered on the current *focus of discussion* in the therapy session. The focus of discussion can be a client thought, thought process, feeling, behavior, situation, or an empirical test of these.

As used here, 'thoughts' includes mental images, memories and beliefs. Thought processes include biases in learning or attribution, for example, catastrophizing, "black and white" thinking, or assuming over-responsibility for thoughts; as well as meta-cognitions – that is, beliefs about thoughts – such as "worrying is helpful" or "worrying will get out of my control".

(d) Empirical Events

Appendix A: The Collaborative Empiricism Scale (cont.)

An empirical event is a specific empirical exploration of a thought, thought process, feeling, behavior, or situation that matches one of the Items A – F below. Empirical events are centered on the current focus of discussion in the session. For example, selecting a thought or feeling to explore is an empirical event matching Item A. An empirical event may also involve other empirical events matching further items A – F. For example, in designing an empirical test of a thought (Item E) the client and therapist may discuss how strongly the client currently believes the thought, matching Item B. In this case, both these events would be rated separately. As some empirical explorations involve several stages within each session, over a significant portion of the session, it is possible that this form could be used multiple times in each session.

Instructions for use:

A rater first identifies the current focus of discussion in the therapy session (i.e., a client thought, thought process, feeling, behavior, characteristic problem situation, or an empirical test of these), and then decides whether this focus of discussion is the target of one of the actions matching the items A – F below. If so, the rater records the start time for this empirical event on the rating sheet. The identified event is then rated independently for empiricism and collaboration. Empiricism is rated using the item from the list A – F below that matches the empirical event; for example, selecting a thought for evaluation is rated using Item A. Collaboration is then rated for this specific empirical event using Item G. That is, each empirical event is rated twice, once using an item A – F and once using item G.

Note: Ratings using items A – F focus on the quality of empiricism only, irrespective of the degree of collaboration. Each rating of collaboration (using item G) relates only to the degree of collaboration involved in the empirical event under consideration, not to collaboration in other aspects of the session, or for the session as a whole.

Item A - Selecting a focus of discussion to explore
Item B - Rating the degree of emotion or belief associated with the focus of discussion
Item C - Exploring the focus of discussion
Item D - Developing an alternative interpretation of the focus of discussion
Item E - Designing an empirical test
Item F - Evaluating an empirical test
Item G - Collaboration: Sharing the work

Key to Ratings Terminology:

(a) Explicit Discussion

The use of "explicit discussion" refers to the linking of the client's experience to the current focus of discussion. That is, the client's thought, thought process, feeling, behavior, or a situation in which they typically occur, is explicitly discussed in relation to and in terms of their experience.

(b) Directly Challenged

The use of "directly challenged" on the rating scale means criticizing or objecting to a thought, feeling, behavior, or situation without clear reference to the client's unique experience. For example, either the client or therapist objects to a belief on logical grounds, including labeling the thought as an example of an irrational error in thinking; appeals to a general principle; or states that it is false/unlikely/unhelpful etc., without reference to evidence drawn from the client's experiences.

Appendix A: The Collaborative Empiricism Scale (cont.)

(c) Seeking Feedback

The use of "seeking feedback" on the rating scale means the client or therapist asked for confirmation or clarification. For example, checking whether the therapist correctly understands the client's belief, or asking whether a belief is the most important or most distressing belief in a situation.

(d) Problems with empiricism

The use of "problems with empiricism" on the rating scale means that empiricism is present, and is the basis of the empirical event; however, there are one or more problems that reduce the adequacy of the empirical basis of the event. For example, failing to specify clear criteria for the outcomes of an empirical test; ratings of belief or emotion are vague or unclear; mistaking interpretations for evidence; empirical evidence is sourced from the therapist, rather than from the client. Note: If problems reduce the empirical exploration of a thought to the extent that empiricism no longer functions as the basis for the empirical event, the event should be scored a '2', rather than a '3'.

(e) Minor problems

The use of "minor problems" on the rating scale means that one or more problems emerged, but these did not reduce the adequacy of empiricism as the basis for exploring the event. Minor problems can be problems in counseling skills, or missed opportunities to do adequate CE better. For example, not considering a related thought the client mentions; or behavior that detracts from the exploration of a cognition, but does not impair the fundamental elements of CE; such as not addressing a doubt or concern of the client's that arises during the evaluation of a thought, or some variability in the level of focus or attention on a task of evaluating a thought.

Appendix A: The Collaborative Empiricism Scale (cont.)

EMPIRICISM	Items A-F cover the range of events involving the empirical exploration of the current focus of discussion.

Item A - Selecting a focus of discussion to explore

The key concept in this item is how the CLIENT'S EXPERIENCE was USED TO SELECT the focus of discussion to explore. For example, the client's experience may be used to determine whether a thought, feeling, or behavior is central to the client's goals or distress, or typical or frequent for the client in the specific situation being discussed.

Selecting involves an explicit reference to a possible topic of discussion. A new topic that simply flows in a stream from the last topic is not a selection.

A **HOW was the client's experience used to select a focus of discussion?**

1 The CLIENT'S EXPERIENCE was NOT IDENTIFIED in SELECTING the focus of discussion (e.g., selection of a thought proceeded without consideration of the emotion or behavior associated with the thought, or was selected on logical grounds, e.g. it was 'black and white' thinking).

2 The CLIENT'S EXPERIENCE was MENTIONED but was NOT USED in selecting the focus of discussion. (E.g., Client: "I feel worried that no-one takes me seriously at work" Therapist: "Let's talk about this."

3 The CLIENT'S EXPERIENCE was EXPLICITLY USED in SELECTING the focus of discussion. (E.g., discussion of the frequency of a thought, or previous occurrences. Client: "I feel worried that no-one takes me seriously at work" Therapist: "When you say 'worried', what do you mean?" Client: "It's getting me down". Therapist: "Should we talk more about this today?"

However, there were PROBLEMS WITH EMPIRICISM that REDUCED the empiricism (e.g., failing to distinguish experience from interpretations: "My boss ignores me in the corridor because he thinks my work is no good"; experience was vague, general, ambiguous, misinterpreted, or was sourced from the therapist, rather than from the client.)

4 The CLIENT'S EXPERIENCE was EXPLICITLY USED in a SPECIFIC WAY to select a focus of discussion. (e.g., T: "So there's several examples of you putting your family before yourself. Which of those experiences could we look at to understand how this strategy might affect you, and your family?"). However, there were MINOR PROBLEMS, which DID NOT REDUCE the empiricism (e.g., some variability in focus on the task by either party; not addressing a doubt or concern that arose during the selection of a belief; not checking for other, possibly more clinically important beliefs).

5 The CLIENT'S EXPERIENCE was a FOCUS AND METHOD to select a focus of discussion to explore. Difficulties that arose during the test were discussed. Feedback was offered or sought. There were NO PROBLEMS, or any difficulties WERE RESOLVED.

Appendix A: The Collaborative Empiricism Scale (cont.)

EMPIRICISM	Items A-F cover the range of events involving the empirical exploration of the current focus of discussion.

Item B – Rating emotion or belief associated with the focus of discussion

Strongly held thoughts that elicit strong emotion (sometimes referred to as 'hot thoughts'), as well as thoughts processes (e.g., catastrophizing or minimizing positive experience), and emotionally charged behaviors or situations, are often important for CBT. Rating the degree of belief or emotion associated with these can serve as the basis for an empirical exploration of a focus of discussion, such as in the context of a 'thought record'. Rating the degree of emotion can also serve as the basis for evaluating the effectiveness of an intervention. This item is rated when a rating is made of *the degree* of emotion or belief associated with the focus of discussion. Do not rate this item when there is only a description of emotion or belief, without an attempt to quantify degree.

B **HOW was the client's experience used to rate the strength of belief or emotion associated with a focus of discussion?**

1 An attempt was made to rate the degree of EMOTION or BELIEF associated with the FOCUS OF DISCUSSION, but the rating was not completed (e.g., a request was made to rate an amount of anxiety, but client and therapist were distracted and the request was not completed.)

2 The CLIENT'S EXPERIENCE was MENTIONED but was NOT USED TO QUANTIFY the rating of belief or emotion (e.g., client responds "I was really stressed", "that was moderately anxiety provoking", "I feel severely depressed", without attempts to otherwise quantify using experience.)

3 The CLIENT'S EXPERIENCE was EXPLICITLY USED to quantify the rating the degree of belief or emotion associated with a focus of discussion, but there were PROBLEMS with EMPIRICISM in the RATING (e.g., there was a rating of 5/10, which was anchored in the client's experience, but the anchors on the scale were ambiguous, or not defined, or were suggested by the therapist.)

4 The CLIENT'S EXPERIENCE was EXPLICITLY USED in a SPECIFIC WAY to quantify the rating of belief or emotion associated with the focus of discussion. However, there were MINOR PROBLEMS, which DID NOT REDUCE the empiricism (e.g., attempt to anchor rating of 5/10, but many situations identified for the same rating, or general rather than specific situations identified, such as "5 on my anxiety scale is when I am being asked to give a talk").

5 The CLIENT'S EXPERIENCE was a FOCUS AND METHOD for rating the degree of belief or emotion associated with a focus of discussion. Difficulties that arose during the test were discussed. FEEDBACK was offered or sought. There were NO PROBLEMS, or any difficulties WERE RESOLVED.

Appendix A: The Collaborative Empiricism Scale (cont.)

EMPIRICISM	Items A-F cover the range of events involving the empirical exploration of the current focus of discussion.

Item C – Exploring the focus of discussion

Questioning and discussion, including Socratic questioning and guided discovery, are central strategies for exploring and evaluating thoughts, feelings, behaviors, and situations. This may involve identifying evidence from the client's experience for or against an interpretation of a cognition or cognitive process. Exploring means the purposeful seeking and use of information and experience, IN ORDER TO EXAMINE the functionality or helpfulness of the current focus of discussion.

C **HOW was the client's experience used to explore the focus of discussion?**

1 The CLIENT'S EXPERIENCE was NOT USED to explore the focus of discussion. Either party DIRECTLY CHALLENGED, CRITICIZED, or ASSUMED the functionality or helpfulness of the focus of discussion, such as on logical grounds, or from general principles (e.g., T: "Is it really possible that *everyone* thinks that about you?", T: "It's 'black and white thinking' to think you are a bad mother because you made one big mistake.")

2 The CLIENT'S EXPERIENCE was MENTIONED but was NOT DISCUSSED OR FOCUSED ON to explore the functionality or helpfulness of the focus of discussion. (E.g., a situation was mentioned in relation to a belief, but was not related directly to the functionality or helpfulness of the belief).

3 The CLIENT'S EXPERIENCE was EXPLICITLY USED to explore the functionality or helpfulness of the focus of discussion (e.g., Therapist: "So you felt worried writing the report?" Client: "Terrified the whole time that I couldn't do it." Therapist: "What happened?" Client: "Well, I got it done the night before." Therapist: "So the thought wasn't true then.")

However, there were PROBLEMS WITH EMPIRICISM that REDUCED the empiricism (e.g., interpretations of evidence were mistaken as facts; experience was vague, general, ambiguous, misinterpreted, or was sourced from the therapist, rather than from the client.)

4 The CLIENT'S EXPERIENCE was EXPLICITLY USED in a SPECIFIC WAY to explore a particular aspect of the focus of discussion.

There were only MINOR PROBLEMS, which DID NOT REDUCE the empiricism (e.g., some variability in focus on the task; some doubt or concern of the client's remained unaddressed; failing to offer or request feedback). (E.g.Therapist: "So you felt worried writing the report?" Client: "Terrified the whole time that I couldn't do it." Therapist: "What effect did the thought 'you couldn't do it' have on you getting it done?"

5 The CLIENT'S EXPERIENCE was a FOCUS AND METHOD FOR EXPLORING the functionality or helpfulness of the focus of discussion. Difficulties that arose during the test were discussed. Feedback was offered or sought. There were NO PROBLEMS, or any difficulties WERE RESOLVED.

Appendix A: The Collaborative Empiricism Scale (cont.)

EMPIRICISM	Items A-F cover the range of events involving the empirical exploration of the current focus of discussion.

Item D – Developing an alternative interpretation of the focus of discussion

This item measures the development of a specific alternative interpretation of the focus of discussion. This may involve, for example, looking at a situation from different perspectives, imagining what other people may think in the same situation, or brainstorming other explanations for what the client believes happened in a situation. Developing an alternative interpretation can occur during both questioning or discussion, and empirical/behavioral tests.

D **HOW was the client's experience used to develop an alternative interpretation of the focus of discussion?**

1 The CLIENT'S EXPERIENCE was NOT IDENTIFIED in developing an alternative interpretation of the focus of discussion. Discussion proceeded, for example, on logical grounds, from general principles, or by identifying an error of thinking. (E.g., Therapist: "What would an alternative be that didn't have a 'should' statement?")

2 The CLIENT'S EXPERIENCE was MENTIONED but was NOT DISCUSSED OR FOCUSED ON to develop an alternative interpretation of the focus of discussion (e.g., Therapist: "How else could you think about this situation?" Client: "I guess it's not always so bad." Therapist: "Ok")

3 The CLIENT'S EXPERIENCE was EXPLICITLY USED to develop an alternative interpretation of the focus of discussion (e.g., Therapist: "So, you didn't fail?" Client: "It wasn't very good, but I did manage to do it." Therapist: "Great. So are you a failure?" Client: "Well, I guess not.")

However, there were PROBLEMS WITH EMPIRICISM that REDUCED the empiricism (e.g., Interpretations of evidence were mistaken as facts; the client did not rate the strength of belief or emotion associated with the alternative interpretation, experience was vague, ambiguous, misinterpreted; interpretations were sourced from the therapist, rather than from the client).

4 The CLIENT'S EXPERIENCE was EXPLICITLY USED in a SPECIFIC WAY to develop an alternative interpretation of the focus of discussion. (e.g., Therapist: "So, you didn't fail?" Client: "It wasn't very good, but I did manage to do it." Therapist: "Let's look a bit more at how you managed to do it." Client: "I guess I stuck at it, even though I felt it was pretty bad. I did do that better than before." Therapist: "What might that mean for the idea 'you're a failure'?"

There were only MINOR PROBLEMS, which DID NOT REDUCE the empiricism (e.g., Some variability in focus on the task; the client expressed some doubt or reservation about the alternative interpretation, which was not explored; failing to offer or request feedback).

5 The CLIENT'S EXPERIENCE was a FOCUS AND METHOD FOR DEVELOPING an alternative interpretation of the focus of discussion. Difficulties that arose during the test were discussed. Feedback was offered or sought. There were NO PROBLEMS, or any difficulties WERE RESOLVED.

Appendix A: The Collaborative Empiricism Scale (cont.)

EMPIRICISM	Items A-F cover the range of events involving the empirical exploration of the current focus of discussion.

Item E – Designing an empirical test

A well-designed empirical test of the focus of discussion, whether it is client or therapist initiated, should be grounded in the client's experience, and requires clear *criteria* for interpreting the outcome of the test (i.e., what the outcomes will *mean*). This could include specifying criteria for what would be expected to happen if the thought were functional, helpful, or perspective shifting for the client; or criteria for evaluating the converse. A good empirical test is able to distinguish between events occurring as a function of the specific thoughts under evaluation and events occurring as a function of other factors. Anticipating difficulties, ranging from practical and environmental, to those connected to the client's other beliefs and compensatory behaviors, is also part of designing a good empirical test.

E **HOW was the client's experience used to design an empirical test?**

1 The test was developed with ONLY GENERAL OR VAGUE REFERENCE to the client's experience. Possible OUTCOMES of the test were VAGUE OR NOT DISCUSSED. (E.g., "Therapist: Let's see what happens if you start a conversation at the party and we'll discuss it next week." or, Client: "I'll try doing it differently in the meeting, maybe that will help.")

2 The CLIENT'S EXPERIENCE was USED in discussing OUTCOMES of the test, but CRITERIA for INTERPRETING OUTCOMES were NOT DISCUSSED or NOT DRAWN FROM THE CLIENT'S EXPERIENCE (e.g., Therapist: "You might find that other people are also very nervous before speaking in the meeting." / Client: "I'll check whether people talk to me at the party or not.")

3 The CLIENT'S EXPERIENCE was EXPLICITLY USED in developing the test. CRITERIA for interpreting outcomes of the test were specified FROM THE CLIENT'S EXPERIENCE.

However, there were PROBLEMS WITH EMPIRICISM that REDUCED the empiricism (e.g., the criteria were not specified in detail with reference to time, place, or amount: "Some people will talk to me at the party." vs. "At least 3 people will talk to me, after I say hello first."; criteria were ambiguous or open to alternative explanation; target outcomes were not publicly observable; target outcomes might occur by chance, e.g. the client plans to circulate around a party, waiting for people to speak to them).

4 The CLIENT'S EXPERIENCE was EXPLICITLY USED in developing the test. CRITERIA for interpreting outcomes of the test were specified FROM THE CLIENT'S EXPERIENCE. Potential difficulties and solutions WERE DISCUSSED. There were only MINOR PROBLEMS, which DID NOT REDUCE the adequacy of the criteria (e.g., some variability in focus on the task; some concern of the client's about the test remained unaddressed; solutions were developed for anticipated problems but the client remained only partially confident).

5 The CLIENT'S EXPERIENCE was COMPREHENSIVELY USED in developing the test. Difficulties that arose during the test were discussed. Feedback was offered or sought. There were NO PROBLEMS, or any difficulties WERE RESOLVED.

Appendix A: The Collaborative Empiricism Scale (cont.)

EMPIRICISM	Items A-F cover the range of events involving the empirical exploration of the current focus of discussion.

Item F – Evaluating an empirical test

Ideally, the meaning of the results of an empirical test is assessed in terms of criteria used in designing the test. This includes distinguishing the actual results of the test (e.g., thoughts, feelings, bodily sensations, behavior of the client and others) from judgments or attributions about the results of the test. A thorough evaluation of an empirical test also involves reviewing problems that arose in carrying it out, ranging from practical and environmental, to those connected to the client's other beliefs and compensatory behaviors (e.g., mistaking interpretations for evidence, distorting evidence in recall, avoiding an anxiety producing situation and thereby missing evidence).

F **HOW was the client's experience used to evaluate the results of an empirical test?**

1 The results of an empirical test were discussed WITHOUT REFERENCE TO SPECIFIC CLIENT EXPERIENCE (e.g., Therapist: "How did it go at the meeting?" Client: "It was a lot better." Therapist: "Good, so what did you learn from that?").

2 The results of an empirical test were discussed WITH REFERENCE to criteria, BUT CRITERIA WERE VAGUE OR AMBIGUOUS. There was LITTLE OR NO DISCUSSION OF POTENTIAL PROBLEMS (e.g., Therapist: "How did it go at the meeting?" Client: "I presented better. I think more people listened." Therapist: "Good, so what did you learn from that?")

3 Both the results of the empirical test and criteria for interpreting outcomes were discussed WITH REFERENCE TO SPECIFIC CLIENT EXPERIENCE (e.g., Therapist: "How did it go at the meeting?" Client: "No one was critical, and 4 people complimented the presentation." Therapist: "What did we decide that would mean about your ability to present your report?")

However, there were PROBLEMS WITH EMPIRICISM that REDUCED the empiricism (e.g., the actual results of the test were not clearly determined; compensatory "safety" behaviors were not explored; difficulties that arose during the empirical test were not discussed; or there were potential biases in reporting the results).

4 Both the results of the empirical test and criteria for interpreting outcomes were discussed WITH REFERENCE TO SPECIFIC CLIENT EXPERIENCE. There were only MINOR PROBLEMS, which DID NOT REDUCE the empiricism (e.g., some variability in focus on the task; an opportunity for learning from the experiment remained unexplored; some doubt or concern about the meaning of the results remained unaddressed; the results were evaluated well, but the practical implications were not discussed).

5 Both the results of the empirical test and criteria for interpreting outcomes were COMPREHENSIVELY discussed WITH REFERENCE TO SPECIFIC CLIENT EXPERIENCE. Criteria were sourced from the client, and the actual results were accurately determined. Difficulties that arose during the test were discussed. There were NO PROBLEMS, or any difficulties WERE RESOLVED.

Appendix A: The Collaborative Empiricism Scale (cont.)

COLLABORATION	Item G covers the extent of genuine collaboration in the empirical exploration of the focus of discussion.

Item G – Collaboration in empiricism: Sharing the work

For the focus of discussion just rated, please rate the level of collaboration (genuinely shared work) between client and therapist. Sharing the work involves *shared decisions*, *mutually responsive interactions,* and *contributions* from both parties (engagement, participation, effort). This is more than simply agreement or cooperation. *Ideal balance* means the client and therapist share the work to the extent that they are able *given the stage of therapy*. For example, early in therapy, the client may contribute less to the design of an empirical test, as this is a new activity for the client, but the client can still *fully share the work* by making contributions, involvement in shared decisions, and mutually responsive interactions.

G **For the specific empirical event just rated using one of Items A – F:**

HOW WELL did the client and therapist SHARE THE WORK of the empirical exploration?

1 There was evidence of client and therapist active shared engagement and participation in empirical exploration, but this engagement would be characterized as GENERALLY ONE SIDED (e.g., one party ignored the other's contributions, GENERALLY made UNILATERAL DECISIONS, or dominated activity in the event).

2 There was evidence of client and therapist shared active engagement in empirical exploration, but there was still MODERATE IMBALANCE of shared engagement in this event (e.g., one party was moderately controlling, discouraging, or disinterested in the other party's involvement, and made SOME UNILATERAL DECISIONS in the process).

3 There was evidence of SOME IMBALANCE in client and therapists' shared active engagement in empirical exploration in this event (e.g., one party allowed less time for the other's creative deliberation; focused on their own perspective; or rushed the other's opportunity to contribute to decision making).

4 There was evidence of client and therapist shared active engagement in empirical exploration and only MINOR IMBALANCE in this event. Client and therapist SHARED DECISIONS, gave the other time to formulate a contribution in the decision making process, were MUTUALLY RESPONSIVE, and CONTRIBUTED (asked open questions; sought/ provided suggestions, opinions, ideas, alternatives, evaluated ideas or empirical tests).

5 There was an IDEAL BALANCE in client and therapist shared active engagement in empirical evaluation. To the EXTENT THAT THE CLIENT COULD be involved in sharing the work at this stage of therapy, they were. Client and therapist SHARED DECISIONS, gave the other time to formulate a contribution in the decision making process, were MUTUALLY RESPONSIVE, and CONTRIBUTED (asked open questions; sought/ provided suggestions, opinions, ideas, alternatives, evaluated ideas or empirical tests).

Appendix B: Tee and Kazantzis (2011)

CLINICAL PSYCHOLOGY
SCIENCE AND PRACTICE

Collaborative Empiricism in Cognitive Therapy: A Definition and Theory for the Relationship Construct

John Tee and Nikolaos Kazantzis, School of Psychological Science, La Trobe University, Australia

Despite the consensus regarding the importance of collaborative empiricism (CE) in Beck's cognitive therapy, absent are operational definitions, measures, or empirical investigations of the construct. Existing research has centered on constructs such as the working alliance and has produced inconsistent findings. It is unclear whether CE is related to treatment outcomes, or whether it mediates other cognitive change processes. This article argues that the core meaning of collaboration in CE, "sharing the work," is not captured by the most frequently used construct of the therapeutic alliance in cognitive therapy research. A theory of CE, based on self-determination theory, is developed that integrates the collaborative and empirical aspects of CE and addresses the motivational aspect of the construct.

Key words: **cognitive, collaboration, collaborative empiricism, empiricism, relationship, therapy.** ***[Clin Psychol Sci Prac 18: 47–61, 2011]***

A contradiction awaits the reader who reviews the literature on collaborative empiricism (CE). The past 40 years have seen widespread agreement regarding the central role of CE in cognitive therapy (Hollon & Beck, 1979; Kuyken, Padesky, & Dudley, 2009; Madsen, 2009; Persons, 2008; Tryon & Misurell, 2008). However, there has been little theoretical analysis, and no definition or assessment of the construct. Although CE is frequently referred to as a distinctive and defining characteristic of cognitive therapy, and as a core component of the therapeutic alliance (e.g., Beck, 1989; Bishop & Fish, 1999; Cross, 2007; Dattilio, 2000; Dudley & Kuyken, 2006; Friedberg & Clark, 2006; Friedberg & Gorman, 2007; Kuyken, Padesky, & Dudley, 2008; Merali & Lynch, 1997; Overholser, 1995; Rutter & Friedberg, 1999; Segal & Shaw, 1996; Will, 1995), it remains unclear whether CE directly or indirectly facilitates hypothesized mechanisms of cognitive change and positive therapeutic outcomes. Existing discussion of CE in the literature has been limited to a brief outline of two components (collaboration and empiricism). Absent are analyses of how these components interact, or a theory of how CE may benefit clients in therapy.

This article will first review the current definition and assessment of CE in cognitive therapy. Second, the article will provide a theoretical discussion to clarify and elaborate on the nature and interaction between the elements of collaboration. Finally, the article will propose new components of CE that can form the basis of a more fully articulated definition of the construct, which can also form the basis of an operational definition, and a new measure.

EXISTING DEFINITIONS OF COLLABORATIVE EMPIRICISM

The therapeutic collaboration was first outlined as a means for identifying "raw data" and a means for encouraging client identification, observation, and

Address correspondence to Nikolaos Kazantzis, School of Psychological Science, La Trobe University, Victoria 3086, Australia. E-mail: N.Kazantzis@latrobe.edu.au.

Appendix B: Tee and Kazantzis (2011) (cont.)

evaluation of introspective beliefs in Beck, Rush, Shaw, and Emery's (1979) *Cognitive Therapy for Depression* (see also Beck, 1967). These seminal works define CE as a basic strategy or form of interaction underlying the relationship between client and therapist in cognitive therapy (Hollon & Beck, 1979). Client and therapist actively collaborate to identify problem situations, and to design, implement, and evaluate empirical tests of the client's beliefs. A key objective is for CE to enable experimentation of beliefs and behaviors outside therapy via the vehicle of homework assignments (Beck, 1970, 1995; Kazantzis, MacEwan, & Dattilio, 2005). This collaborative sharing of the work of experimental testing distinguishes CE. For example, a therapist aiming to point out errors in thought process through disputation may be attempting to foster empiricism, without collaboration:

> Client: I don't want to go to the dinner; people never talk to me—they think I'm boring.
>
> Therapist: Sounds like you're mindreading again. When you're at the dinner, notice whether people actually seem bored, or whether you're basing your feeling on what other people think, or worse yet, what you think they're thinking.

The same lead from the client could be pursued by a therapist adherent to CE, as outlined in A. T. Beck's therapy, by taking the following approach:

> Client: I don't want to go to the dinner; people never talk to me—they think I'm boring.
>
> Therapist: That sounds like a tricky situation for you Would you like to talk about that some more—and see what we can figure out here as a way to help?
>
> Therapist: Ok, well if I am hearing you correctly, a concern that you have is that people think you are boring, and this is connected to your conversations with people. It might help us to find out under which situations this happens the most.... For the moment, I wonder what might be the reasons that people (including you) might not talk much sometimes?
>
> Client: Gee, well, I don't know...
>
> Perhaps they are tired ... maybe, even they are waiting for a lead from someone else? Is it possible they think I don't make much of an effort?
>
> Therapist: Those sound like some good possibilities. I really like how you are looking around the situation from different perspectives. I wonder how could we figure out how much of the "not talking" is due to other people's views of you ... or any of these other ideas you have generated?
>
> Client: Hmmm. Maybe I could take more notice of what happens when I am socializing....
>
> Therapist: Good, let's figure out how to be specific about how you will notice that....

A true sharing of the work of identifying ideas to be tested, and then devising means of testing them out, clearly has a different presence in therapy. However, much of the literature following the early teaching in CE is less than clear. While there is agreement that CE is important for cognitive therapy (cf. Kuyken et al., 2008, 2009; Persons, 2008; Tryon & Misurell, 2008), subsequent analysis and discussion have been largely based on practical, commonsense grounds, rather than firm conceptual foundation. For instance, CE has been described as "embedded in every clinical action and decision" (Friedberg & McClure, 2002, p. 44), which comprises the "cornerstone of CBT" and is "used throughout the entire course of treatment" (Dattilio, 2000, p. 39). There is also confusion as to whether CE is a relationship attribute or technique, or both a process and an approach (e.g., Gleeson & McGorry, 2004). In some cases, CE has been referred to as a major therapeutic method and strategy in Beck's therapy (Abramson, Alloy, & Dykman, 1990; Bell, Grech, Maiden, Halligan, & Ellis, 2005; Dattilio, 2000). In other cases, CE is referred to as a "philosophy which forms the foundation of cognitive therapy" (Wilkes, 1994, p. 309), as well as a "specific technique" and an "atmosphere" that pervades the patient–therapist relationship (Turkington & Siddle, 1998, p. 237). Thus, it remains unclear whether CE can be considered a requirement for the therapeutic relationship in Beck's

Appendix B: Tee and Kazantzis (2011) (cont.)

cognitive therapy, or whether it can be better defined as a technique or a strategy (Hazlett-Stevens, 2008; van Oppen, 2004; Waddington, 2002; Westefeld et al., 2000). Importantly, as there is no measure or outline of what constitutes more or less effective CE in the literature, clarifying the definition and its connection with the therapeutic alliance is important for research, training, and practice.

COLLABORATIVE EMPIRICISM AND THE THERAPEUTIC ALLIANCE

The therapeutic alliance can be broadly defined as the alliance of a client's reasonable side with a therapist's working side (Gelso & Hayes, 1998; Horvath & Greenberg, 1994). There are reliable meta-analytic findings of a robust relationship between therapeutic alliance and treatment outcomes in cognitive therapy (Horvath & Symonds, 1991; Martin, Garske, & Davis, 2000; Shirk & Karver, 2003). This relationship holds across different therapies and disorders, despite whether outcome or alliance is rated by therapist, client, or observer (Martin et al., 2000; Shirk & Karver, 2003).

Bordin's conception of the alliance has been the most often used in cognitive therapy research. Bordin described collaboration as a "working relationship" between client and therapist (Bordin, 1979, 1994). At face value, this working relationship seems similar to the meaning of collaboration in CE as "sharing the work." However, Bordin conceived collaboration in the working alliance as an "agreed-upon contract" between client and therapist regarding the tasks and goals of therapy (Bordin, 1994, p. 254). According to Bordin, agreement is "the key process in building an initial, viable alliance" (Horvath & Greenberg, 1994, p. 1). However, collaboration in CE is not simply willing participation by the client nor agreement on tasks or goals. Rather, the cognitive therapist aims to engage the client in a genuine sharing of the work of goal setting and creative authorship of therapeutic tasks, progressively encouraging the client to take the lead role in these activities as far as is practicable (Beck, 1995; Beck, Emery, & Greenberg, 1985; Beck et al., 1979). The aim of this partnership is to actively share the work of empirically evaluating the client's beliefs, not simply to agree on a program of therapeutic tasks and goals. However, problems in the definition of CE are extended to discussions of the alliance in cognitive therapy. Incompatible definitions of CE include comment that CE is referred to as the basis or ideal cognitive therapy alliance (Beck et al., 1985; Bishop & Fish, 1999; Dobson & Shaw, 1988), whereas elsewhere it is neither the relationship, nor a component of it, but a "stylistic fulcrum that permits the helping alliance to thrive" (Stein, Kupfer, & Schatzberg, 2006, p. 359).

If the concept of collaboration in Bordin's (1979) working alliance fails to capture the meaning of collaboration in CE, it is possible that measures of the alliance based on Bordin's conception may similarly fail to capture the meaning of collaboration in CE. It also follows that research studies that have employed working alliance measures may not have accounted for CE. This possibility is demonstrated here by an analysis of the most commonly used measure of the therapeutic relationship in cognitive therapy, the Working Alliance Inventory (WAI; Horvath & Greenberg, 1989, 1994).

The Working Alliance Inventory

The WAI is a 36-item questionnaire designed specifically to measure Bordin's conception of the working alliance (Horvath & Greenberg, 1989, 1994). It has been widely used in alliance research, across many types of therapy (Horvath, 2001; Horvath & Greenberg, 1989, 1994; Horvath & Luborsky, 1993; Horvath & Symonds, 1991; Martin et al., 2000), and has good psychometric properties (Hatcher & Gillaspy, 2006). However, analysis of the items in the WAI reveals that no item attempts to measure collaboration, or indeed empiricism, as these concepts are used in cognitive therapy. In the popular WAI revised short form, three items in the 4-item Goal subscale refer to "agreement" and "good understanding," but collaboration is not covered. The fourth item does refer to "collaboration in setting goals," but this phrasing does not discriminate the meaning of collaboration as "sharing the work" that is central in CE. No items in either the Task or Bond subscales refer to collaboration. Therefore, we argue that the most frequently used construct of the therapeutic alliance in cognitive therapy outcome research fails to capture the key aspects of CE as described in Beck's cognitive therapy.

Appendix B: Tee and Kazantzis (2011) (cont.)

COLLABORATIVE EMPIRICISM AND THERAPIST COMPETENCE

The Cognitive Therapy Rating Scale

Several scales have been designed to assess a therapist's competence in cognitive therapy practice, in a specific session, with a specific client (Brosan, Reynolds, & Moore, 2007; Brosan, Reynolds, & Moore, 2008; Dobson & Christensen, 2001; Kazantzis, 2003; Trepka, Rees, Shapiro, Hardy, & Barkham, 2004). The Cognitive Therapy Scale (CTS; Young & Beck, 1980, 1988) has been designed as a supervisor/independent observer rating scale to assess the extent that cognitive therapy is conducted in a manner consistent with its undergirding theory and is being flexibly tailored to meet individual client needs (cf. Kazantzis, Reinecke, & Freeman, 2010). Included in the CTS is an item on "Collaboration." The CTS rating manual specifies three meanings of collaboration for this item: ensuring compatible goals to avoid working against each other, minimizing client resistance, and preventing misunderstandings. Although this sense of active teamwork is compatible with the genuine sharing of the work of empirical evaluation involved in CE, whether "collaboration" is interpreted in the sense meant in CE, of a creative and genuine sharing of the work, or simply agreement on tasks and goals, is a matter of interpretation by the individual rater. Further, neither the item nor the rating manual specifies empirical investigation as a focus for collaboration, but empirical investigation is a key component of CE.

Other items on the CTS also refer to collaboration (i.e., Agenda Setting, Feedback, Understanding, and Guided Discovery), but similarly do not distinguish the specific meanings of collaboration and empiricism inherent in CE or provide clear guidance on the definition of collaborative empiricism. Questions regarding the consistency of CTS ratings made by expert raters have been published (Jacobson & Gortner, 2000; Shaw et al., 1999), so it is possible that greater attention and delineation of the "empiricism" element to CE could improve in the scale's psychometric properties. A similar level of ambiguity regarding the CE construct exists in the revised CTS (e.g., Blackburn et al., 2001; Gordon, 2007).

The Cognitive Therapy Adherence and Competence Scale (Barber, Liese, & Abrams, 2003) has been designed to improve the psychometric properties of the CTS and to discriminate between therapist adherence and competence. The scale contains 25 items in five sections. Five items relate to the collaborative therapeutic relationship. Collaboration is defined as "shared responsibility for defining patient's problems and potential solutions; functioning as a team," but as with previous measures, it mainly misses the focus on empiricism. Item 23 refers to "asking for evidence/alternative views: asked for patient's evidence for maladaptive beliefs," but does not emphasize collaboration. Given that levels of inter-rater agreement were low for items measuring collaboration, guided discovery, and "asking for evidence/alternative views" (adherence ICCs range 0.43–0.55, competence ICCs range 0.48–0.53), it is possible that this lack of clarity in the CE construct contributed to the low level of reliability on these items. Thus, as with existing measures of the therapeutic relationship, current measures of therapist competence in cognitive therapy cannot discriminate the meanings of collaboration or empiricism specific to CE without idiosyncratic interpretation of these terms by raters.

Change Processes in Cognitive Therapy

Collaborative empiricism has been hypothesized to act as a change process in cognitive therapy (Dobson & Dobson, 2009; Kuyken et al., 2008, 2009). Change in cognitive therapy has often been understood via the cognitive mediation hypothesis, which proposes that improvements in therapeutic outcome result from direct change in clients' cognitions (Beck, 1970; DeRubeis, Tang, & Beck, 2001). Belief change is hypothesized to be more likely if the rationale for change comes from clients' identification of their own introspective data, and collaborative experimentation, rather than from direct illustration by the therapist (Dattilio & Padesky, 1990). Self-efficacy may be lower if the therapist is seen as the change agent, increasing the probability of relapse (Horvath & Greenberg, 1994). Self-evaluation of beliefs, using experiments that are collaboratively designed, avoids the confrontation implicit in directly challenging delusional beliefs and hallucinations (Alford & Beck, 1994). In addition, collaborative empiricism reframes the process of therapy as a shared, challenging, empirical experiment (Klosko & Sanderson, 1999). This may increase the

Appendix B: Tee and Kazantzis (2011) (cont.)

psychological activation and a sense of hope, which may be more important than specific cognitive or behavioral tasks (Cross, 2007). The atmosphere of curiosity modeled in CE may also foster flexible thinking, encourage a broader examination of the client's problems, and allow the client and therapist to get at core subjective experiences (Friedberg & McClure, 2002).

In case conceptualization, the client's unique experience guides how both the theory and empirically supported principles of cognitive therapy are employed with a particular individual in his or her unique circumstances (Beck, 1995; Persons, 2008). Kuyken et al. (2009) provided a helpful account of CE, as one of three fundamental principles informing the case conceptualization process in cognitive therapy. In their guide for practitioners, it was proposed that collaboration and empiricism function in tandem as the driving force underlying the process of case conceptualization. Accordingly, the genuine collaboration inherent in CE aids the development of an accurate shared understanding of the client's problem situation. Information central to understanding the client's concerns and goals in their unique historical, social, developmental, and psychological contexts benefits from collaboration. The empirical aspect of CE then functions as a check on the accuracy of the information generated through the collaborative data-gathering process (Kuyken et al., 2009). Aspects of the conceptualization that do not fit the client's situation can be revised or discarded after empirical testing. The process of CE first increases the range of information available to the case conceptualization and then filters out inaccurate information by testing it empirically. Therefore, CE mediates the relationship between the case conceptualization and therapeutic outcomes by increasing the relevance and accuracy of the conceptualization.

Collaborative empiricism may also increase the accuracy of the case conceptualization by overcoming heuristic biases inherent in therapists' interpretations of their clients' problem situations (Kuyken et al., 2009). In developing a case conceptualization, therapists have working hypotheses about their client's situation and interpret information arising in session in terms of these hypotheses. Therapists, like anyone else, are susceptible to overreliance or misapplication of heuristics in decision making (Kahneman, 2003). However, the structured collaboration of CE, combined with explicit empirical testing of the client's cognitions, may reduce the impact of these biases or distortions in cognitive processing (Bieling & Kuyken, 2003).

CE may also reduce erroneous or biased recall on the part of either client or therapist, by subjecting recall to ongoing empirical validation (Hollon, 1999). Reducing biased recall by the therapist may affect outcomes by increasing the accuracy of the case conceptualization, which may then affect outcomes as discussed earlier. Collaborative empiricism may increase objectivity in the session by reducing reliance on potentially biased recall, and by providing the client with a systematic method of objectively evaluating his or her beliefs (Hollon & Kriss, 1984). The merit of these approaches is that they articulate mechanisms by which case conceptualization may be related to CE.

This discussion provides some explanation of "how" empirical investigation may affect therapeutic outcomes, but it does not explain "why" a client may be motivated to engage in the arduous process of evaluating his or her beliefs. Many clients, perhaps most, come to therapy driven by an emotional pain they wish to relieve. However, as the motivational interviewing literature underscores, clients often have conflicting motivations for therapy (Miller & Rollnick, 2002). Change may offer greater freedom or less pain but comes with its own costs. A theory that addresses this motivational dimension of CE is lacking.

A THEORETICAL BASIS FOR COLLABORATIVE EMPIRICISM

A theoretical foundation for CE that addresses the motivational aspect of the construct, and further articulates the structure of its component parts, may be drawn from self-determination theory (SDT). Self-determination theory states that people have a fundamental propensity toward growth, self-determination, and the resolution of psychological discrepancy (Deci & Ryan, 1980, 1985; Ryan & Deci, 2000, 2002). The theory seeks to explain how individuals self-regulate behavior and internalize extrinsic motivation to engage in autonomous action (Markland, Ryan, Tobin, & Rollnick, 2005; Ryan & Deci, 2000; Vansteenkiste & Sheldon, 2006). It is also concerned with the effect of environmental factors, such as directives or rewards, on intrinsic motivation to engage in behaviors (Deci & Ryan, 2000).

Appendix B: Tee and Kazantzis (2011) (cont.)

Self-determination theory proposes that behaviors are regulated on a continuum of autonomy, ranging from intrinsic (autonomous) regulation to introjected regulation, and then external regulation. External regulation refers to behavior that is regulated by external factors, such as rewards or punishments. A direct attempt to persuade a client to change a belief is an example of an externally regulated motivation. Introjected regulation refers to motivation that is internalized. In this case, the pressure to engage in a behavior comes from within the person, for example, in the form of self-approval when engaging successfully in the behavior, and shame or self-criticism when failing to engage successfully in the behavior (Ryan & Deci, 2000). Empirical work suggests that introjected regulation is more likely to lead to sustained engagement in behavior than external regulation (Assor, Vansteenkiste, & Kaplan, 2009; Deci & Ryan, 2000). However, it can be accompanied by significant ambivalence, as well as negative emotional states and inner conflict (Markland et al., 2005; Ryan, Rigby, & King, 1993). At the other end of the continuum is intrinsic motivation. Intrinsically motivated behaviors are engaged in willingly and are experienced as rewarding of themselves (Deci & Ryan, 1980).

Self-determination theory proposes that three key needs underlie the innate tendency toward growth: competence, autonomy, and relatedness (Deci & Ryan, 2000). That is, humans have an innate need to experience themselves as competent, autonomous actors who are related to others in meaningful and satisfying relationships. These factors work to increase the level of intrinsic motivation associated with behavior.

There is considerable empirical support for SDT in the social psychology literature. Of the three underlying psychological needs proposed by SDT, most of the existing research has been conducted on the concept of autonomy. Autonomy in SDT involves the self-endorsement of behavior, accompanied by a perception of willingness or volition (Ryan & Deci, 2008). Increased autonomy has been associated with supporting autonomous choices (Moller, Deci, & Ryan, 2006), goals that are self-endorsed (Koestner, Ryan, Bernieri, & Holt, 1984), and promoting choice (Deci, Eghrari, Patrick, & Leone, 1994). It leads to increased perception of the self as the locus of causality in therapy, with a corresponding increase in integration of new behaviors and learning (Deci & Ryan, 2008).

Higher autonomous (intrinsic) motivation has also been associated with better session outcomes, as rated by either client or therapist, and lower psychopathology (Michalak, Klappheck, & Kosfelder, 2004), as well as lower tension, less distraction, and higher positive mood in session, and higher ratings of importance, satisfaction, and intention to continue with therapy (Pelletier, Tuson, & Haddad, 1997). It has also been found to predict symptom remission and post-treatment depression severity, independently of whether the therapeutic modality was interpersonal therapy, cognitive-behavior therapy, or pharmacotherapy with clinical management (Zuroff et al., 2007). Similarly, the self-reported willingness of clients to engage in treatment for drug dependence has been shown to predict therapist-rated engagement and post-treatment abstinence (Zeldman, Ryan, & Fiscella, 2004). Therapeutic goals that are valued as more personally important or interesting are associated with lower negative affect, higher positive affect, and higher self-esteem (Sheldon & Kasser, 1995). In addition, autonomous regulation of behavior is performed with higher quality and attention (Deci & Ryan, 1980), is more stable over time (Deci & Ryan, 1985), and is associated with more positive emotional states (Deci & Ryan, 1980; Ryan & Deci, 2002).

This relationship between intrinsic motivation and outcomes is also found in nonclinical settings. For example, higher intrinsic motivation has been shown to predict weight loss (Williams, Grow, Freedman, Ryan, & Deci, 1996), eating self-regulation (Mata et al., 2009), and greater needs satisfaction in volunteers (Weinstein & Ryan, 2010).

Self-Determination Theory and Collaborative Empiricism

According to SDT, supporting the client's sense of autonomy is required to augment competence and relatedness and promote optimal motivation for change (Deci & Ryan, 2000; Markland et al., 2005). Self-efficacy, the belief that one is competent to engage in a behavior, is not sufficient to maximize motivation. Clients may feel able to engage in behavior but not feel motivated to do so. Increasing the level of autonomous regulation of a behavior is required to maximize

Appendix B: Tee and Kazantzis (2011) (cont.)

motivation to engage in the behavior (Deci & Ryan, 1985; Reeve, 1998, 2002).

Self-determination theory can be used to explain how CE moderates therapeutic outcomes and why clients may be more motivated to engage in the process of belief change as a result of CE. The evidence reviewed above suggests that at least four environmental conditions support clients' autonomy: a meaningful rationale for behavior change, minimal external contingent reinforcers (rewards and punishments), active participation and exercising of choice, and acceptance and acknowledgment of negative feelings (Markland et al., 2005; Reeve, 1998, 2002). Collaborative empiricism directly influences the first three of these, and the fourth as part of general good cognitive therapy practice.

In CE, a meaningful rationale for changing beliefs arises from the clients' experience of the results of their own empirical tests of their beliefs. By definition, a good empirical test of the client's beliefs is one in which alternative explanations for data resulting from the experiment have been accounted for in the design of the experiment (Beck, 1995). The elimination of alternative explanations of the results of experiments increases the validity of the data. The key point suggested here is that CE also increases the degree of internally autonomous relevance of the data for the client. Reasons for changing beliefs are ideally experienced as the client's own reasons, rather than the therapist's reasons. In CE, the client experiences this data as intrinsically his or her own. The meaning of this *intrinsic data* is judged using criteria that the client decides upon before the experiment is conducted. The client is the source of these criteria, and hence, the source of the regulatory capacity of the criteria as a source of interpretation of the data is internal to the client.

Self-determination theory suggests that intrinsic data are more likely to be experienced as an internal and autonomous source of regulation, leading in turn to increased motivation for the client to change his or her belief. In addition, the process of collaboration minimizes external contingent reinforcers, further increasing internal motivation. Although CE is directive in the sense that the client is directed to engage in the process of CE, the CE process itself is genuinely collaborative. The client is encouraged to participate in all aspects of CE and to actively exercise choice and creativity in the design and evaluation of empirical tests of his or her beliefs. This genuine sharing of the work of therapy, combined with the focus on the client's intrinsic data as the source and arbiter of empirical evaluation of beliefs, increases the client's intrinsic motivation for change.

A DEFINITION OF COLLABORATIVE EMPIRICISM

The concept of CE is founded in Beck's (1967, 1976) seminal cognitive theory and system of psychotherapy. In subsequent work, CE has been described as a combination of two components, collaboration and empiricism. However, the relationship between these components remains unclear. As outlined earlier, SDT is able to clarify this relationship in terms of the key role of the client's intrinsic data (see Table 1).

When client and therapist collaborate to share the work of therapy, but do not empirically evaluate the client's cognitive content (quadrant II), the rationale for belief change is not grounded in the client's evaluation of his or her own intrinsic data. According to SDT, this lack of intrinsic data results in lower intrinsic motivation, or even external motivation if data are sourced from the therapist. Lower intrinsic motivation is associated with lower motivation to change. Conversely, when intrinsic data are used in the empirical evaluation of the client's beliefs, but the process is not genuinely collaborative (quadrant III), SDT suggests that the lack of collaboration will reduce the client's sense of autonomy, competence, and efficacy and shift the client's locus of causality externally, again resulting in lower intrinsic motivation to change. When collaboration and empiricism are combined in CE, SDT suggests that the resulting generation and evaluation of the client's intrinsic data are likely to increase the client's sense of efficacy, autonomy, internal locus of causality, competence, and therefore intrinsic motivation for change (see Figure 1).

In summary, we suggest that SDT can provide a theoretical basis for the efficacy of CE in belief change. SDT highlights the key role of the client's intrinsic data as the point of operation of the components of collaboration and empiricism. In this view, CE encourages autonomously motivated belief change by promoting the acceptance and internalization of new beliefs, and

Appendix B: Tee and Kazantzis (2011) (cont.)

Table 1. Components of collaboration, empiricism, and intrinsic data

	Low collaboration	High collaboration
Low empiricism	I. Low CE	II. Increased autonomy but without evaluation of beliefs via intrinsic data. Lower intrinsic motivation for change
High empiricism	III. External locus of causality, lower efficacy, diminished impact of intrinsic data	IV. Evaluation of client's intrinsic data. Increased autonomy, competence, intrinsic motivation

by increasing the meaningfulness of new beliefs for the client.

The application of SDT to CE raises several hypotheses for empirical examination. SDT suggests that the collaborative empirical investigation of clients' intrinsic data increases the core SDT variables of autonomy, competence, and relatedness, thereby increasing motivation to engage in behavior change. SDT also clarifies the functioning of collaboration and empiricism in CE in terms of these core variables. Before outlining a research agenda informed by these hypotheses, existing empirical work on collaboration and empiricism will be reviewed.

Empirical Evaluation of Collaborative Empiricism

There has been no empirical investigation of CE in cognitive therapy research trials. However, research in other contexts has found a relationship between aspects of collaboration in CE and therapeutic outcomes (Boardman, Catley, Grobe, Little, & Ahluwalia, 2006; Bordin, 1994; Clemence, Hilsenroth, Ackerman, Strassle, & Handler, 2005; Colson et al., 1988; Connor-Greene, 1993; Creed & Kendall, 2005; Diamond, Liddle, Hogue, & Dakof, 1999; Hatcher, 1999; Hatcher & Barends, 1996; Le Bloc'h, de Roten, Drapeau, & Despland, 2006).

One such aspect of collaboration in CE is the framing of the client as central to the process of shared decision making. Sharing decisions with the client regarding the process of therapy, and developing a clear rationale and explanation for the course of a session are associated with clients' satisfaction with that session (Eisenthal, Koopman, & Lazare, 1983). Clients are more likely to return for therapy after an intake interview in which the therapist shares the problem formulation and negotiates mutual therapeutic goals, compared to an interview where neither is done (Tracy, 1977). Similarly, therapist–client discussion of goals and expression of thoughts and feelings are associated with observer ratings of "goodness" of therapy sessions (Hoyt, 1980; Hoyt, Xenakis, Marmar, & Horowitz, 1983). Simply specifying goals is enough to increase clients' ratings of therapy helpfulness (Goldstein, Cohen, Lewis, & Struening, 1988). Client-rated agreement with therapist and client-rated experience of goal consensus at session 2 has been shown to predict the reduction in symptoms on the Symptom Check List (SCL; Derogatis, 1974), including up to six months later (Dormaar, Dijkman, & de Vries, 1989). Higher response congruence, in which therapist verbal response directly addresses the subject of the client's immediately preceding statement, is associated with higher likelihood of the client returning for therapy after the initial session (Duehn & Proctor, 1974). Finally, topic determination, the proportion of topic initiations subsequently followed by the other participant in therapy (client or therapist), predicts the continuation of therapy past session 3 (Tracey, 1986). These aspects of collaboration do not address the core meaning of collaboration in CE. Nevertheless, shared decision making, sharing the formulation, negotiation

Figure 1. Hypothetical relationships in the benefits of evaluating intrinsic data.

Appendix B: Tee and Kazantzis (2011) (cont.)

of therapeutic goals, response congruence, and topic determination are aspects of collaboration in CE for which there is empirical support.

Not all studies have found a relationship between aspects of collaboration and outcome. Beyebach and Carranza (1997) failed to replicate the relationship between topic determination and engagement reported by Tracey (1986). However, nonengaged clients in the Beyebach and Carranza study used more domineering language with therapists, interrupted more often, and engaged in more conflict. It is possible that these behaviors disrupted collaboration, resulting in lower continuation of therapy (Tryon & Winograd, 2002).

Collaborative involvement has also been used as a measure of collaboration in therapy process. Collaborative involvement is the mutual involvement of client and therapist in a helping relationship (Tryon & Winograd, 2002). The term "mutual involvement" suggests that this construct has a meaning similar to collaboration in CE. However, studies using the construct have operationalized collaborative involvement as the compliance or cooperation of the client to the strategies of the therapist. For example, Schmidt and Woolaway-Bickel (2000) operationalized collaborative involvement as homework completion. They found that therapists' ratings of compliance with homework predicted outcomes, but clients' ratings of compliance did not. This suggests a relationship between clients' compliance and outcomes (see review in Kazantzis, Whittington, & Dattilio, in press), but is a long way from "mutual involvement" in the sense of a genuinely collaborative sharing of the work indicative of CE.

In an interesting study, O'Malley, Suh, and Strupp (1983) found that client involvement at session 3 predicted outcome, whereas involvement at sessions 1 and 2 did not. They concluded that client involvement may not be simply a quality that the client brings to therapy, but may be facilitated by qualities of the therapist. Windholz and Silberschatz (1988) found that therapist ratings of client involvement were related to outcomes, but not client or observer ratings. It is suggested here that involvement may also be a function of the interaction between client and therapist, rather than a quality of one or the other in isolation. If true, the genuine collaboration inherent in CE could be expected to increase client involvement.

A MEASURE OF COLLABORATIVE EMPIRICISM

Empirical work has shown a relationship between aspects of collaboration and therapeutic outcomes. This suggests that CE, which incorporates these aspects of collaboration, may also be related to outcomes. Nevertheless, results of this research have been mixed, and much of this work has focused on collaboration as a client factor, rather than a property of the dyad. Collaboration in CE goes beyond the meaning of collaboration in these studies. Thus, while research to date provides qualified support for a relationship between aspects of CE and outcomes, research focusing specifically on CE is needed. For this, a measure is needed that incorporates the aspects of collaboration discussed earlier and operationalizes the meanings of collaboration and empiricism specific to CE.

As with collaboration, there has been no specific investigation of the empirical component of CE in cognitive therapy trials. However, there has been some evaluation of one form of empirical evaluation of clients' beliefs in cognitive therapy in the form of behavioral experiments (BEs; Beck et al., 1979; Clark, Beck, & Alford, 1999; Dobson & Dobson, 2009). Behavioral experiments are widely used and recommended in cognitive therapy as a component of standard treatment for a range of disorders (Beck, 1995; Bennett-Levy et al., 2004; Padesky & Greenberger, 1995). It has been suggested that the very ubiquity of BEs in successful therapy attests to their value (Bennett-Levy et al., 2004). However, despite their importance in cognitive therapy, only one study has directly investigated the efficacy of BEs in belief change (Bennett-Levy, 2003). In this study, 27 trainee cognitive therapists compared the effectiveness of BEs and automatic thought records in changing their own cognitions. The authors reported that BEs produced significantly higher cognitive and behavioral change compared with thought records. Several trainees credited the higher efficacy of BEs to their experiential aspect, noting that BEs led to new cognitions that were more powerfully believed than those which were understood rationally but not experienced in a BE.

There is some empirical work supporting a mediating effect of CE on therapeutic outcomes. While there has been no direct investigation of the relationship between case conceptualization and outcomes in

Appendix B: Tee and Kazantzis (2011) (cont.)

cognitive therapy, there is evidence that another type of case conceptualization, the Core Conflictual Relationship Theme (CCRT) method, used in brief psychodynamic therapy, has good reliability and validity and is related to therapeutic outcomes (Dobson & Dobson, 2009; Luborsky & Crits-Christoph, 1998). This association has been interpreted in the literature as supporting a causal link between more accurate case conceptualization and therapeutic outcomes (e.g., Bieling & Kuyken, 2003). However, while some studies have found that increased accuracy in conceptualization predicted symptom reduction (Crits-Christoph, Barber, & Kurcias, 1993), other work has not found this effect (Crits-Christoph, Cooper, & Luborsky, 1988). These findings provide qualified support for the hypothesis that the accuracy of case conceptualization may be related to outcomes in cognitive therapy. Further empirical work is needed, focusing specifically on the role of CE in case conceptualization in cognitive therapy.

The lack of empirical investigation of the role of collaboration and empiricism in CE, as well as hypotheses suggested by the application of SDT to CE above, suggests a research agenda for CE. An operational measure of CE is required to investigate the relationship between CE and the theoretically derived variables suggested in this article, as well as a component investigation of the CE construct. Development and evaluation of a measure of CE are currently being undertaken by the authors. One goal of the research program is to facilitate an assessment of therapist competence in the use of CE, as part of the training and supervision of practitioners in this important therapeutic process. Another primary goal is to enable prospective treatment outcome research to evaluate whether "sharing the work" of identifying and empirically evaluating clients' beliefs can lead to better therapeutic outcomes.

CONCLUSION

The present article reviewed definitions, provided a theoretical discussion, and presented new components of collaborative empiricism. There is support in the literature for the efficacy of aspects of collaboration and empiricism in CE. However, no theoretical or empirical work to date has defined or measured the core elements of the construct. If the criticisms regarding the capability of existing measures to capture the therapeutic relationship as conceived by Aaron T. Beck are accepted, serious questions can be asked of the empirical based on the therapeutic approach:

> Have previous studies missed an important determinant of successful cognitive therapy outcome? If training and evaluation of therapists' practices in cognitive therapy trials have missed collaborative empiricism, have the effects of the therapy been diluted?

A measure specific to CE is required to address these issues. Existing theory provides a plausible explanation for elements of the collaboration and empiricism components of CE, but remains piecemeal and ignores the motivational role of the construct. A new theory for CE has been proposed in this article, drawing on well-validated social-cognitive theory, which integrates the collaborative and empirical components of the construct and its motivational aspect.

Following from this article, the field would benefit from clinical discussion illustrating the role of CE in effective cognitive therapy, empirical work utilizing systematic case studies, and prospective evaluations of the relationship attribute as a determinant of therapeutic outcomes. Put simply, cognitive therapy researchers are encouraged to collaborate on the empirical evaluation of the CE construct.

REFERENCES

Abramson, L. Y., Alloy, L. B., & Dykman, B. M. (1990). Cognitive theory and therapy of emotional disorders: A lay epistemic reconstruction? *Psychological Inquiry, 1*, 198–200.

Alford, B. A., & Beck, A. T. (1994). Cognitive therapy of delusional beliefs. *Behaviour Research and Therapy, 32*, 369–380.

Assor, A., Vansteenkiste, M., & Kaplan, A. (2009). Identified versus introjected approach and introjected avoidance motivations in school and in sports: The limited benefits of self-worth strivings. *Journal of Educational Psychology, 101*, 482–497.

Barber, J. P., Liese, B. S., & Abrams, M. J. (2003). Development of the Cognitive Therapy Adherence and Competence Scale. *Psychotherapy Research, 13*, 205–221.

Beck, A. T. (1967). *Depression: Clinical, experimental, and theoretical aspects*. New York: Harper and Row.

Appendix B: Tee and Kazantzis (2011) (cont.)

Beck, A. T. (1970). Cognitive therapy: Nature and relation to behavior therapy. *Behavior Therapy, 1*, 184–200.

Beck, A. T. (1976). *Cognitive therapy and the emotional disorders*. New York: International Universities Press.

Beck, A. T. (1989). An interview with a depressed and suicidal patient. In D. Wedding & R. J. Corsini (Eds.), *Case studies in psychotherapy*. (pp. 125–142). Itasca, IL: F. E. Peacock.

Beck, J. S. (1995). *Cognitive therapy: Basics and beyond*. New York: Guilford.

Beck, A. T., Emery, G., & Greenberg, R. L. (1985). *Anxiety disorders and phobias: A cognitive perspective*. New York: Basic Books.

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

Bell, V., Grech, E., Maiden, C., Halligan, P. W., & Ellis, H. D. (2005). 'Internet delusions': A case series and theoretical integration. *Psychopathology, 38*, 144–150.

Bennett-Levy, J. (2003). Mechanisms of change in cognitive therapy: The case of automatic thought records and behavioral experiments. *Behavioral and Cognitive Psychotherapy, 31*, 261–277.

Bennett-Levy, J., Butler, G., Fennell, M., Hackmann, A., Mueller, M., & Westbrook, D. (Eds.). (2004). *Oxford guide to behavioral experiments in cognitive therapy*. New York: Oxford University Press.

Beyebach, M., & Carranza, V. E. (1997). Therapeutic interaction and dropout: Measuring relational communication in solution-focused therapy. *Journal of Family Therapy, 19*, 173–221.

Bieling, P. J., & Kuyken, W. (2003). Is cognitive case formulation science or science fiction? *Clinical Psychology: Science and Practice, 10*, 52–69.

Bishop, W., & Fish, J. M. (1999). Questions as interventions: Perceptions of Socratic, solution-focused, and diagnostic questioning styles. *Journal of Rational-Emotive and Cognitive Behavior Therapy, 17*, 115–140.

Blackburn, I.-M., James, I. A., Milne, D. L., Baker, C., Standart, S., Garland, A., et al. (2001). The revised Cognitive Therapy Scale (CTS-R): Psychometric properties. *Behavioural and Cognitive Psychotherapy, 29*, 431–446.

Boardman, T., Catley, D., Grobe, J., Little, T., & Ahluwalia, J. (2006). Using motivational interviewing with smokers: Do therapist behaviors relate to engagement and therapeutic alliance? *Journal of Substance Abuse Treatment, 31*, 329–339.

Bordin, E. (1979). The generalizability of the psychoanalytic concept of the working alliance. *Psychotherapy: Theory, Research & Practice, 16*, 252–260.

Bordin, E. (1994). Theory and research on the therapeutic working alliance: New directions. In A. O. Horvath & L. S. Greenberg (Eds.), *The working alliance: Theory, research, and practice* (pp. 13–37). Oxford: Wiley.

Brosan, L., Reynolds, S., & Moore, R. G. (2007). Factors associated with competence in cognitive therapists. *Behavioural and Cognitive Psychotherapy, 35*, 179–190.

Brosan, L., Reynolds, S., & Moore, R. G. (2008). Self evaluation of cognitive therapy performance: Do therapists know how competent they are?. *Behavioural and Cognitive Psychotherapy, 36*, 581–587.

Clark, D. A., Beck, A. T., & Alford, B. A. (1999). *Scientific foundations of cognitive theory and therapy of depression*. New York: Wiley.

Clemence, A., Hilsenroth, M., Ackerman, S., Strassle, C., & Handler, L. (2005). Facets of the therapeutic alliance and perceived progress in psychotherapy: Relationship between patient and therapist perspectives. *Clinical Psychology and Psychotherapy, 12*, 443–454.

Colson, D. D., Horwitz, L., Allen, J. G., Frieswyk, S. H., Gabbard, G. O., Newsom, G. E., et al. (1988). Patient collaboration as a criterion for the therapeutic alliance. *Psychoanalytic Psychology, 5*, 259–268.

Connor-Greene, P. (1993). The therapeutic context: Preconditions for change in psychotherapy. *Psychotherapy: Theory, Research, Practice, Training, 30*, 375–382.

Creed, T. A., & Kendall, P. C. (2005). Therapist alliance-building behavior within a cognitive-behavioral treatment for anxiety in youth. *Journal of Consulting and Clinical Psychology, 73*, 498–505.

Crits-Christoph, P., Barber, J. P., & Kurcias, J. S. (1993). The accuracy of therapists' interpretations and the development of the therapeutic alliance. *Psychotherapy Research, 3*, 25–35.

Crits-Christoph, P., Cooper, A., & Luborsky, L. (1988). The accuracy of therapists' interpretations and the outcome of dynamic psychotherapy. *Journal of Consulting and Clinical Psychology, 56*, 490–495.

Cross, L. W. (2007). Islands of health: Implications of academic psychology for psychotherapy technique and integration. *Journal of Psychotherapy Integration, 17*, 251–273.

Dattilio, F. M. (2000). Cognitive-behavioral strategies. In J. Carlson & L. Sperry (Eds.), *Brief therapy with individuals and couples* (pp. 33–70). Phoenix, AZ: Zeig, Tucker and Theisen.

Dattilio, F. M., & Padesky, C. A. (1990). *Cognitive therapy with couples*. Sarasota, FL: Professional Resource Exchange.

Deci, E. L., Eghrari, H., Patrick, B. C., & Leone, D. R. (1994). Facilitating internalization: The self-determination theory perspective. *Journal of Personality, 62*, 119–142.

Appendix B: Tee and Kazantzis (2011) (cont.)

Deci, E. L., & Ryan, R. M. (1980). The empirical exploration of intrinsic motivational processes. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 13, pp. 39–80). New York: Academic Press.

Deci, E. L., & Ryan, R. M. (1985). *Intrinsic motivation and self-determination in human behavior.* New York: Plenum.

Deci, E. L., & Ryan, R. M. (2000). The "what" and the "why" of goal pursuits: Human needs and the self-determination of behavior. *Psychological Inquiry, 11*, 227–268.

Deci, E. L., & Ryan, R. M. (2008). Facilitating optimal motivation and psychological well-being across life's domains. *Canadian Psychology, 49*, 14–23.

Derogatis, L. R. (1974). The Hopkins Symptom Checklist (HSCL). *Behavioral Science, 19*, 1–15.

DeRubeis, R. J., Tang, T. Z., & Beck, A. T. (2001). Cognitive therapy. In K. S. Dobson (Ed.), *Handbook of cognitive behavioral therapies* (2nd ed., pp. 349–392). New York: Guilford.

Diamond, G. M., Liddle, H. A., Hogue, A., & Dakof, G. A. (1999). Alliance-building interventions with adolescents in family therapy: A process study. *Psychotherapy: Theory, Research, Practice, Training, 36*, 355–368.

Dobson, K. S., & Christensen, A. (2001). Lessons from the conduct of clinical research trials. *Behavior Therapist, 24*, 155–156.

Dobson, D. J. G., & Dobson, K. S. (2009). *Evidence-based practice of cognitive-behavioral therapy.* New York: Guilford.

Dobson, K. S., & Shaw, B. F. (1988). The use of treatment manuals in cognitive therapy: Experience and issues. *Journal of Consulting and Clinical Psychology, 56*, 673–680.

Dormaar, J., Dijkman, C. I., & de Vries, M. W. (1989). Consensus in patient-therapist interactions: A measure of the therapeutic relationship related to outcome. *Psychotherapy and Psychosomatics, 51*, 69–76.

Dudley, R., & Kuyken, W. (2006). Formulation in cognitive-behavioral therapy: 'There is nothing either good or bad, but thinking makes it so.' In L. Johnstone & R. Dallos (Eds.), *Formulation in psychology and psychotherapy: Making sense of people's problems* (pp. 17–46). New York: Routledge.

Duehn, W. D., & Proctor, E. K. (1974). No exit: Interactional control of premature discontinuance. *American Journal of Orthopsychiatry, 44*, 201–202.

Eisenthal, S., Koopman, C., & Lazare, A. (1983). Process analysis of two dimensions of the negotiated approach in relation to satisfaction in the initial interview. *Journal of Nervous and Mental Disease, 171*, 49–54.

Friedberg, R. D., & Clark, C. C. (2006). Supervision of cognitive therapy with youth. In T. K. Neill (Ed.), *Helping others help children: Clinical supervision of child psychotherapy* (pp. 109–122). Washington, DC: American Psychological Association.

Friedberg, R. D., & Gorman, A. A. (2007). Integrating psychotherapeutic processes with cognitive behavioral procedures. *Journal of Contemporary Psychotherapy, 37*, 185–193.

Friedberg, R. D., & McClure, J. M. (2002). *Clinical practice of cognitive therapy with children and adolescents.* New York: Guilford.

Gelso, C. J., & Hayes, J. A. (1998). *The psychotherapy relationship: Theory, research and practice.* New York: Wiley.

Gleeson, J. F. M., & McGorry, P. D. (2004). *Psychological interventions in early psychosis.* Chichester, UK: Wiley.

Goldstein, J. M., Cohen, P., Lewis, S. A., & Struening, E. L. (1988). Community treatment environments: Patient vs. staff evaluations. *Journal of Nervous and Mental Disease, 176*, 227–233.

Gordon, P. (2007). A comparison of two versions of the Cognitive Therapy Scale. *Behavioural and Cognitive Psychotherapy, 35*, 343–353.

Hatcher, R. (1999). Therapists' views of treatment alliance and collaboration in therapy. *Psychotherapy Research, 9*, 405–423.

Hatcher, R., & Barends, A. (1996). Patients' view of the alliance in psychotherapy: Exploratory factor analysis of three alliance measures. *Journal of Consulting and Clinical Psychology, 64*, 1326–1336.

Hatcher, R., & Gillaspy, J. (2006). Development and validation of a revised short version of the Working Alliance Inventory. *Psychotherapy Research, 16*, 12–25.

Hazlett-Stevens, H. (2008). *Psychological approaches to generalized anxiety disorder.* New York: Springer.

Hollon, S. D. (1999). Rapid early response in cognitive behavior therapy: A commentary. *Clinical Psychology: Science and Practice, 6*, 305–309.

Hollon, S. D., & Beck, A. T. (1979). Cognitive therapy of depression. In P. C. Kendall & S. D. Hollon (Eds.), *Cognitive-behavioral interventions: Theory, research, and procedures* (pp. 153–203). New York: Academic Press.

Hollon, S. D., & Kriss, M. R. (1984). Cognitive factors in clinical research and practice. *Clinical Psychology Review, 4*, 35–76.

Horvath, A. O. (2001). The alliance. *Psychotherapy: Theory, Research, Practice, Training, 38*, 365–372.

Horvath, A. O., & Greenberg, L. S. (1989). Development and validation of the Working Alliance Inventory. *Journal of Counseling Psychology, 36*, 223–233.

Horvath, A. O., & Greenberg, L. S. (1994). *The working alliance: Theory, research, and practice.* New York: Wiley.

Appendix B: Tee and Kazantzis (2011) (cont.)

Horvath, A. O., & Luborsky, L. (1993). The role of the therapeutic alliance in psychotherapy. *Journal of Consulting and Clinical Psychology, 61*, 561–573.

Horvath, A. O., & Symonds, B. (1991). Relation between working alliance and outcome in psychotherapy: A meta-analysis. *Journal of Counseling Psychology, 38*, 139–149.

Hoyt, M. F. (1980). Therapist and patient actions in "good" psychotherapy sessions. *Archives of General Psychiatry, 37*, 159–161.

Hoyt, M. F., Xenakis, S. N., Marmar, C. R., & Horowitz, M. J. (1983). Therapists' actions that influence their perceptions of "good" psychotherapy sessions. *Journal of Nervous and Mental Disease, 171*, 400–404.

Jacobson, N. S., & Gortner, E. T. (2000). Can depression be de-medicalized in the 21st century: Scientific revolutions, counter-revolutions and the magnetic field of normal science. *Behaviour Research and Therapy, 38*, 103–117.

Kahneman, D. (2003). A perspective on judgment and choice: Mapping bounded rationality. *American Psychologist, 58*, 697–720.

Kazantzis, N. (2003). Therapist competence in cognitive-behavioural therapies: Review of the contemporary empirical evidence. *Behaviour Change, 20*, 1–12.

Kazantzis, N., MacEwan, J., & Dattilio, F. M. (2005). A guiding model for practice. In N. Kazantzis, F. P. Deane., K. R. Ronan, & L. L'Abate (Eds.), *Using homework assignments in cognitive behavior therapy* (pp. 359–407). New York: Routledge.

Kazantzis, N., Reinecke, M. A., & Freeman, A. (2010). Concluding remarks. In N. Kazantzis, M. A. Reinecke, & A. Freeman (Eds.), *Cognitive and behavioral theories in clinical practice* (pp. 363–367). New York: Guilford.

Kazantzis, N., Whittington, C. J., & Dattilio, F. M. (2010). Meta-analysis of homework effects in cognitive and behavioral therapy: A replication and extension. *Clinical Psychology: Science and Practice, 17*, 144–156.

Klosko, J. S., & Sanderson, W. C. (1999). *Depression: Clinical application of empirically supported psychotherapy*. Northvale, NJ: Jason Aronson.

Koestner, R., Ryan, R. M., Bernieri, F., & Holt, K. (1984). Setting limits on children's behavior: The differential effects of controlling versus informational styles on intrinsic motivation and creativity. *Journal of Personality, 52*, 233–248.

Kuyken, W., Padesky, C. A., & Dudley, R. (2008). The science and practice of case conceptualization. *Behavioral and Cognitive Psychotherapy, 36*, 757–768.

Kuyken, W., Padesky, C. A., & Dudley, R. (2009). *Collaborative case conceptualization: Working effectively with patients in cognitive-behavioral therapy*. New York: Guilford.

Le Bloc'h, Y., de Roten, Y., Drapeau, M., & Despland, J. (2006). New, but improved? Comparison between first and revised version of the Helping Alliance Questionnaire. *Schweizer Archiv fur Neurologie und Psychiatrie, 157*, 23–28.

Luborsky, L., & Crits-Christoph, P. (1998). *Understanding transference: The core conflictual relationship theme method* (2nd ed.). New York: Basic Books.

Madsen, W. M. (2009). Collaborative helping: A practice framework for family-centered services. *Family Process, 48*, 103–116.

Markland, D., Ryan, R. M., Tobin, V. J., & Rollnick, S. (2005). Motivational interviewing and self-determination theory. *Journal of Social and Clinical Psychology, 24*, 811–831.

Martin, D. J., Garske, J. P., & Davis, 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.

Mata, J., Silva, M. N., Vieira, P. N., Carraça, E., Andrade, A. M., Coutinho, S. R., et al. (2009). Motivational "spill-over" during weight control: Increased self-determination and exercise intrinsic motivation predict eating self-regulation. *Health Psychology, 28*, 709–716.

Merali, N., & Lynch, P. (1997). Collaboration in cognitive-behavioral counselling: A case example. *Canadian Journal of Counselling, 31*, 287–293.

Michalak, J., Klappheck, M. A., & Kosfelder, J. (2004). Personal goals of psychotherapy patients: The intensity and the "why" of goal-motivated behavior and their implications for the therapeutic process. *Psychotherapy Research, 14*, 193–209.

Miller, W. R., & Rollnick, S. (2002). *Motivational interviewing: Preparing people to change*. New York: Guilford.

Moller, A. C., Deci, E. L., & Ryan, R. M. (2006). Choice and ego-depletion: The moderating role of autonomy. *Personality and Social Psychology Bulletin, 32*, 1024–1036.

O'Malley, S. S., Suh, C. S., & Strupp, H. H. (1983). The Vanderbilt Psychotherapy Process Scale: A report on the scale development and a process-outcome study. *Journal of Consulting and Clinical Psychology, 51*, 581–586.

van Oppen, P. (2004). Cognitive therapy for obsessive-compulsive disorder. *Clinical Case Studies, 3*, 333–349.

Overholser, J. C. (1995). Elements of the Socratic method: iV. Disavowal of knowledge. *Psychotherapy: Theory, Research, Practice, Training, 32*, 283–292.

Padesky, C. A., & Greenberger, D. (1995). *Clinician's guide to mind over mood*. New York: Guilford.

Pelletier, L. G., Tuson, K. M., & Haddad, N. K. (1997). Client Motivation for Therapy Scale: A measure of

Appendix B: Tee and Kazantzis (2011) (cont.)

intrinsic motivation, extrinsic motivation, and amotivation for therapy. *Journal of Personality Assessment, 68*, 414–435.

Persons, J. B. (2008). *The case formulation approach to cognitive-behavior therapy*. New York: Guilford.

Reeve, J. (1998). Autonomy support as an interpersonal motivating style: Is it teachable? *Contemporary Educational Psychology, 23*, 312–330.

Reeve, J. (2002). Self-determination theory applied to educational settings. In E. L. Deci & R. M. Ryan (Eds.), *Handbook of self-determination research* (pp. 193–204). Rochester, NY: University of Rochester Press.

Rutter, J. G., & Friedberg, R. D. (1999). Guidelines for the effective use of Socratic dialogue in cognitive therapy. In L. VandeCreek & T. L. Jackson (Eds.), *Innovations in clinical practice: A source book* (pp. 481–490). Sarasota, FL: Professional Resource Exchange.

Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist, 55*, 68–78.

Ryan, R. M., & Deci E. L. (2002). An overview of self-determination theory: An organismic-dialectical perspective. In E. L. Deci & R. M. Ryan (Eds.), *Handbook of self-determination research* (pp. 3–36). Rochester, NY: University of Rochester Press.

Ryan, R. M., & Deci, E. L. (2008). A self-determination theory approach to psychotherapy: The motivational basis for effective change. *Canadian Psychology, 49*, 186–193.

Ryan, R. M., Rigby, S., & King, K. (1993). Two types of religious internalization and their relations to religious orientations and mental health. *Journal of Personality and Social Psychology, 65*, 586–596.

Schmidt, N. B., & Woolaway-Bickel, K. (2000). The effects of treatment compliance on outcome in cognitive-behavioral therapy for panic disorder: Quality versus quantity. *Journal of Consulting and Clinical Psychology, 68*, 13–18.

Segal, Z. V., & Shaw, B. F. (1996). Cognitive therapy. *American Psychiatric Press Review of Psychiatry, 15*, 69–90.

Shaw, B. F., Elkin, I., Yamaguchi, J., Olmsted, M., Vallis, T., Dobson, K. S., et al. (1999). Therapist competence ratings in relation to clinical outcome in cognitive therapy of depression. *Journal of Consulting and Clinical Psychology, 67*, 837–846.

Sheldon, K. M., & Kasser, T. (1995). Coherence and congruence: Two aspects of personality integration. *Journal of Personality and Social Psychology, 68*, 531–543.

Shirk, S., & Karver, M. (2003). Prediction of treatment outcome from relationship variables in child and adolescent therapy: A meta-analytic review. *Journal of Consulting and Clinical Psychology, 71*, 452–464.

Stein, D. J., Kupfer, D. J., & Schatzberg, A. F. (Eds.). (2006). *The American Psychiatric Publishing textbook of mood disorders*. Washington, DC: American Psychiatric Publishing.

Tracey, T. J. (1986). Interactional correlates of premature termination. *Journal of Consulting and Clinical Psychology, 54*, 784–788.

Tracy, J. J. (1977). Impact of intake procedures upon patient attrition in a community mental health center. *Journal of Consulting and Clinical Psychology, 45*, 192–195.

Trepka, C., Rees, A., Shapiro, D. A., Hardy, G. E., & Barkham, M. (2004). Therapist competence and outcome of cognitive therapy for depression. *Cognitive Therapy and Research, 28*, 143–157.

Tryon, W. W., & Misurell, J. R. (2008). Dissonance induction and reduction: A possible principle and connectionist mechanism for why therapies are effective. *Clinical Psychology Review, 28*, 1297–1309.

Tryon, G. S., & Winograd, G. (2002). Goal consensus and collaboration. In J. C. Norcross (Ed.), *Psychotherapy relationships that work: Therapist contributions and responsiveness to patients* (pp. 109–125). New York: Oxford University Press.

Turkington, D., & Siddle, R. (1998). Cognitive therapy for the treatment of delusions. *Advances in Psychiatric Treatment, 4*, 235–241.

Vansteenkiste, M., & Sheldon, K. M. (2006). There's nothing more practical than a good theory: Integrating motivational interviewing and self-determination theory. *British Journal of Clinical Psychology, 45*, 63–82.

Waddington, L. (2002). The therapy relationship in cognitive therapy: A review. *Behavioral and Cognitive Psychotherapy, 30*, 179–192.

Weinstein, N., & Ryan, R. M. (2010). When helping helps: Autonomous motivation for prosocial behavior and its influence on well-being for the helper and recipient. *Journal of Personality and Social Psychology, 98*, 222–244.

Westefeld, J. S., Range, L. M., Rogers, J. R., Maples, M. R., Bromley, J. L., & Alcorn, J. (2000). Suicide: An overview. *The Counseling Psychologist, 28*, 445–510.

Wilkes, T. (1994). *Cognitive therapy for depressed adolescents*. New York: Guilford.

Will, T. E. (1995). Cognitive therapy and the MCMI-III. In P. D. Retzlaff (Ed.), *Tactical psychotherapy of the personality disorders: An MCMI-III-based approach* (pp. 90–110). Needham Heights, MA: Allyn & Bacon.

Williams, G. C., Grow, V. M., Freedman, Z. R., Ryan, R. M., & Deci, E. L. (1996). Motivational predictors of weight loss and weight-loss maintenance. *Journal of Personality and Social Psychology, 70*, 115–126.

Appendix B: Tee and Kazantzis (2011) (cont.)

Windholz, M. J., & Silberschatz, G. (1988). Vanderbilt Psychotherapy Process Scale: A replication with adult outpatients. *Journal of Consulting and Clinical Psychology*, *56*, 56–60.

Young, J. E., & Beck, A. T. (1980). *Cognitive Therapy Scale: Rating manual*. Unpublished manuscript, University of Pennsylvania.

Young, J. E., & Beck, A. T. (1988). *Cognitive Therapy Scale*. Unpublished manuscript, University of Pennsylvania.

Zeldman, A., Ryan, R. M., & Fiscella, K. (2004). Motivation, autonomy support, and entity beliefs: Their role in methadone maintenance treatment. *Journal of Social and Clinical Psychology*, *23*, 675–696.

Zuroff, D. C., Koestner, R., Moskowitz, D. S., McBride, C., Marshall, M., & Bagby, M. R. (2007). Autonomous motivation for therapy: A new common factor in brief treatments for depression. *Psychotherapy Research*, *17*, 137–147.

Received December 13, 2009; revised May 16, 2010; accepted May 25, 2010.

Appendix C: Expert Feedback Questionnaire

The Collaborative Empiricism Scale: Survey Review

The Collaborative Empiricism Scale

The Collaborative Empiricism Scale – Expert Feedback

Thank you for participating in this review of the Collaborative Empiricism Scale.

As an expert, your feedback is highly important to us, particularly given the lack of an existing measure of collaborative empiricism. We are very grateful for your participation.

The aim of this review is to assess the face validity, usefulness, and psychometric structure of the scale. You will be asked to rate the seven scale items on five Likert type scales and to note any comments.

Survey Outline

The Collaborative Empiricism Scale is designed to be an assessment tool based on direct observation of therapy. We have included the instructions for those who will use the scale (i.e., trainers, supervisors, independent observers) in black type font.

Ethical Conditions for the Expert Survey

Survey responses will be automatically de-identified by the survey software and no identifying information will be placed on data records. Data records will be kept in a locked storage room, and computer files will be password protected. It will not be possible for individual survey data to be withdrawn subsequently because responses will be anonymous. At the end of the survey you may choose to receive the results of the survey by email. In this case, your email address will kept on file until the aggregated survey results are sent, after which it will be deleted.

Block 9

Empirical Events:

The Collaborative Empiricism Scale focuses on empirical events, which are identified within a session and rated separately for empiricism and collaboration. Given the central mechanism of change in CBT is evaluation of beliefs, the scale defines an empirical event as the identification and evaluation of cognition, using any or all of the items below (A-F).

Item A - Selecting a Thought or Thought Process for Evaluation
Item B - Rating Emotion or Belief in Thoughts and Thought Processes
Item C - Questioning a Thought or Thought Process
Item D - Designing an Empirical Test
Item E - Evaluating an Empirical Test
Item F - Developing an Alternative Interpretation of a Thought or Thought Process

Item G - Collaboration in Empiricism

Block 10

Key Scale Constructs:

(a) Empiricism

A core principle of Beckian cognitive therapy is the focus on the client's unique experience. In the context of facilitating the central mechanism of change, cognitive change, the usefulness and functionality of the client's thoughts and beliefs are explored in terms of the client's experience, rather than by analyzing the client's thinking from general principles or on logical grounds. As used here, 'thoughts' includes mental images, memories, and beliefs. Thoughts and thought processes are explored with reference to problem situations in which they characteristically occur, and with reference to the accompanying emotional experience. This involves exploration of the client's feelings, behavior, and experiences as they relate to a specific thought being explored.

Empiricism then, guides the design, implementation, and evaluation of therapeutic work that has been specifically designed to evaluate the client's thoughts and thought processes, and is personally meaningful for the client. Importantly, collaborative empiricism involves more than evaluating whether a thought is true or false. The aim is to foster the client's exploration of the client's thoughts, in empirical terms, and for the client to determine the usefulness and functionality of their thinking by examining the consequences.

(b) Collaboration

Collaboration in Beck's cognitive therapy relates to a genuine sharing of therapeutic work. This involves the client and therapist actively contributing to the therapy session in a combined effort, involving shared decision making and mutually responsive interaction. Examples include sharing in the design and implementation of

Appendix C: Expert Feedback Questionnaire (cont.)

agenda items, goals, and interventions. Collaboration is more than just client and therapist agreement, cooperation, or client adherence/ compliance with therapist suggestions.

(c) Thoughts and thought processes

Collaborative empiricism is focused on the client's thoughts and thought processes. As used here, 'thoughts' includes mental images, memories and beliefs. Thought processes include biases in learning or attribution, for example, catastrophizing, 'black and white' thinking, or assuming over-responsibility for thoughts; as well as meta-cognitions – that is, beliefs about thoughts – such as "worrying is helpful" or "worrying will get out of my control".

Block 11

To use the scale, a rater first identifies a point in the therapy session where there is a specific empirical event matching one of items A – F below, and records the start time for this event on the rating sheet. The identified empirical event is then rated independently for empiricism and collaboration. Empiricism is rated using the item from the list A – F below that matches the empirical event; for example, selecting a thought for evaluation is rated using Item A. Collaboration is then rated for this specific empirical event using Item G. That is, each empirical event is rated twice, once using an item A – F and once using item G.

Note: Ratings using items A – F focus on the quality of empiricism only, irrespective of the degree of collaboration. Each rating of collaboration (using item G) relates only to the degree of collaboration involved in the empirical event under consideration, not to collaboration in other aspects of the session, or for the session as a whole.

Collaborative Empiricism Scale Items

Item A - Selecting a Thought or Thought Process for Evaluation
Item B - Rating Emotion or Belief in Thoughts and Thought Processes
Item C - Questioning a Thought or Thought Process
Item D - Designing an Empirical Test
Item E - Evaluating an Empirical Test
Item F - Developing an Alternative Interpretation of a Thought or Thought Process
Item G - Collaboration in Empiricism

Block 12

Key Rating Terms:

(a) Some Problems

The use of "some problems" on the rating scale means that one or more problems emerged that reduced the adequacy of the empirical basis of the event. For example, where an alternative belief was grounded in the client's experience, but the therapist did not allow the client time to reflect further on their experience, or did not check with the client that the belief identified was in fact important to the client in that situation, or central to the client's distress.

(b) Minor Problems

The use of "minor problems" on the rating scale means that one or more problems emerged, but these did not reduce the adequacy of the empirical basis of the event. Minor problems can be missed opportunities to do adequate CE better. For example, not considering a related thought the client mentions; or behavior that detracts from the exploration of a cognition, but does not impair the fundamental elements of CE, such as not addressing a doubt or concern of the client's that arises during the evaluation of a thought, or some variability in the level of focus or attention on a task of evaluating a thought.

CES Item 1

Guidance for Expert Feedback:

For each of the seven Collaborative Empiricism Scale items, please read the scale item and then answer the Questions for Expert Feedback below it, adding comments in the text box at the end of each survey section.

Please read "Item A – Selecting a Thought or Thought Process" and answer the underlined Questions for Expert Feedback below.

Item A - Selecting a Thought or Thought Process for Evaluation

This item rates the selection of a thought or thought process for empirical exploration. The choice of a central thought to work on often impacts strongly on the efficiency and effectiveness of cognitive therapy. Typically, a thought is chosen that is central to the client's distress, important in the case formulation, and clearly related to the client's goals.

Appendix C: Expert Feedback Questionnaire (cont.)

The key concept in this item is whether the CLIENT'S EXPERIENCE IS THE BASIS for selecting the thought to explore. For example, the client's experience may be used to determine whether the thought is central to the client's goals or distress, or typical, frequent, and characteristic for the client in the specific situation being discussed.

A	HOW WELL did the client and therapist select a thought to work on?	
1	A thought was selected with NO REFERENCE TO THE CLIENT'S EXPERIENCE (e.g., discussion of a thought proceeded without consideration of the emotion or behavior associated with the thought).	
2	There was PASSING REFERENCE to the client's experience, but the client's EXPERIENCE WAS NOT THE BASIS for selecting which thought to work on (e.g., Therapist: "So you had the thought, 'I can't give the speech' and you felt very anxious?" Client: "Yes". Therapist: "Ok, let's explore this thought.")	
3	The client's EXPERIENCE WAS THE BASIS for selecting which thought to work on (e.g.,"Out of these three thoughts, which one made you feel most anxious?" / "Which of these ideas contributes most to you wanting to avoid your boss in the corridor?"). However, there were SOME PROBLEMS which reduced the quality of the empirical data collected to select which thought to work on (e.g., failing to distinguish experience from interpretations: "My boss ignores me in the corridor because he thinks my work is no good".)	
4	The client's EXPERIENCE WAS THE BASIS for selecting which thought to work on. There were only MINOR PROBLEMS, which did not reduce the use of experience to determine which thought to work on (e.g., some variability in focus on the task by either party; not addressing a doubt or concern that arose during the selection of a belief; not checking for other, possibly more clinically important beliefs).	
5	The client's EXPERIENCE WAS THE BASIS for selecting which thought to work on. There were NO PROBLEMS, or any difficulties WERE RESOLVED.	

Questions for Expert Feedback:

To what extent does the item description accurately reflect an element of collaborative empiricism?

To a very small extent ○ To a small extent ○ Somewhat ○ To a large extent ○ To a very large extent ○

To what extent does this item capture variability in collaborative empiricism?

To a very small extent ○ To a small extent ○ Somewhat ○ To a large extent ○ To a very large extent ○

To what extent is this item clear and easy to understand?

To a very small extent ○ To a small extent ○ Somewhat ○ To a large extent ○ To a very large extent ○

To what extent are the anchor points (1-5) in this item clear and easy to understand?

To a very small extent ○ To a small extent ○ Somewhat ○ To a large extent ○ To a very large extent ○

To what extent does this item describe an aspect of collaborative empiricism that you routinely use in your professional practice?

To a very small extent ○ To a small extent ○ Somewhat ○ To a large extent ○ To a very large extent ○

Comments:

CES Item 2

Appendix C: Expert Feedback Questionnaire (cont.)

Please read "Item B – Rating Emotion or Belief in Thoughts and Thought Processes" and answer the underlined Questions for Expert Feedback below.

Item B – Rating Emotion or Belief in Thoughts and Thought Processes

Strongly held thoughts that elicit strong emotion (sometimes referred to as 'hot thoughts'), as well as thoughts processes (e.g., catastrophizing or minimizing positive experience) are often important for CBT. Rating the strength of belief or associated emotion can serve as the basis for an empirical exploration of a thought or thought process, such as in the context of a 'thought record'. The emotion again serves as the basis for evaluating the utility of the intervention.

B	HOW WELL did the client and therapist rate the strength of belief or emotion associated with a thought?	
1	A thought was rated, but NO REFERENCE was made to the client's experience, and NO USE OF THE RATING was made to determine how helpful evaluating a thought had been, or whether to continue with that thought or move on (e.g., Therapist: "Which of those thoughts should we look at first?" Client: "Maybe, 'If my report is late, I'll get fired'?" Therapist: "Okay.")	
2	PASSING REFERENCE was made to the client's experience of a thought, but this was NOT THE BASIS for rating the strength of belief or emotion associated with that thought (e.g., Therapist: "Which of those thoughts should we look at first?" Client: "Well, I'm really worried about my report being late." Therapist: "Okay, let's start there.")	
3	The client's EXPERIENCE WAS THE BASIS for rating the strength of belief or emotion associated with a thought (e.g., Therapist: "Which of those thoughts do you feel most stressed about?" Client: "Getting fired." Therapist: "More than being judged by colleagues?" Client: "Yes"). However, there were SOME PROBLEMS which reduced the use of experience to rate the strength of belief or emotion associated with a thought (e.g., mistaking interpretations for evidence, ratings of strength of belief or emotion were vague, ambiguous, or were inadvertently suggested by the therapist, rather than sourced from the client).	
4	The client's EXPERIENCE WAS THE BASIS for rating the strength of belief or emotion associated with a thought. However, there were MINOR PROBLEMS, which did not reduce the use of experience as the basis for rating the strength of belief or emotion associated with a thought (e.g., some variability in focus on the task, some doubt or concern about rating remained unaddressed).	
5	The client's EXPERIENCE WAS THE BASIS for rating the strength of belief or emotion associated with a thought. There were NO PROBLEMS, or any difficulties WERE RESOLVED.	

Questions for Expert Feedback:

To what extent does the item description accurately reflect an element of collaborative empiricism?

To a very small extent ○ To a small extent ○ Somewhat ○ To a large extent ○ To a very large extent ○

To what extent does this item capture variability in collaborative empiricism?

To a very small extent ○ To a small extent ○ Somewhat ○ To a large extent ○ To a very large extent ○

To what extent is this item clear and easy to understand?

To a very small extent ○ To a small extent ○ Somewhat ○ To a large extent ○ To a very large extent ○

To what extent are the anchor points (1-5) in this item clear and easy to understand?

To a very small extent ○ To a small extent ○ Somewhat ○ To a large extent ○ To a very large extent ○

Appendix C: Expert Feedback Questionnaire (cont.)

To what extent does this item describe an aspect of collaborative empiricism that you routinely use in your professional practice?

To a very small extent ○ To a small extent ○ Somewhat ○ To a large extent ○ To a very large extent ○

Comments:

CES Item 3

Please read "Item C – Questioning a Thought or Thought Process" and answer the underlined Questions for Expert Feedback below.

Item C – Questioning a Thought or Thought Process

Questioning and discussion, including Socratic questioning and guided discovery, are central strategies for the identification and evaluation of evidence for exploring thoughts or thought processes. The client and therapist are able to explore the functionality or helpfulness of a thought based on evidence from the client's experience. Evidence can exist in the form of supporting the thought, as well as in considering evidence that does not support the thought.

C	**HOW WELL did the client and therapist question or discuss the thought?**	
1	A thought was evaluated using questioning WITHOUT REFERENCE TO THE CLIENT'S EXPERIENCE. For example, either party DIRECTLY CHALLENGED, CRITICIZED, or ASSUMED the validity/invalidity or functionality/dysfunctionality of the thought, such as on logical grounds, or from general principles (e.g., "Is it really possible that everyone thinks that about you?", "Do you think it's silly to think that?")	
2	The client's experience of a thought WAS DISCUSSED, but WAS NOT THE BASIS FOR EXPLORING the functionality or helpfulness of the thought (e.g., The client's feelings or behavior related to a belief were discussed, but there was NO EXPLICIT DISCUSSION for and against the functionality, helpfulness, or truth of the belief).	
3	The client's EXPERIENCE WAS THE BASIS FOR EXPLORING the functionality or helpfulness of a thought (e.g., Therapist: "So you felt worried writing your last report?" Client: "Terrified the whole time that I couldn't do it." Therapist: "What happened?" Client: "Well, I got it done the night before. It was 'okay', I guess." Therapist: "So how helpful was this thought, 'You couldn't do it'?"). However, there were SOME PROBLEMS that reduced the use of experience as the basis of questioning (e.g., interpretations of evidence were mistaken as facts; questioning or discussion of experience was vague, ambiguous, or was suggested by the therapist, rather than sourced from the client)	
4	The client's EXPERIENCE WAS THE BASIS FOR EXPLORING the functionality or helpfulness of a thought. There were only MINOR PROBLEMS that did not reduce the use of experience as the basis of questioning (e.g., some variability in focus on the task; some doubt or concern of the client's about rating remained unaddressed; failing to offer or request feedback).	
5	The client's EXPERIENCE WAS THE FOCUS AND METHOD FOR EXPLORING the functionality or helpfulness of a thought. There was COMPREHENSIVE DISCUSSION of evidence for and against the functionality or helpfulness of the thought. There were NO PROBLEMS, or any process difficulties WERE RESOLVED.	

Questions for Expert Feedback:

To what extent does the item description accurately reflect an element of collaborative empiricism?

To a very small extent ○ To a small extent ○ Somewhat ○ To a large extent ○ To a very large extent ○

Appendix C: Expert Feedback Questionnaire (cont.)

To what extent does this item capture variability in collaborative empiricism?

To a very small extent ○ To a small extent ○ Somewhat ○ To a large extent ○ To a very large extent ○

To what extent is this item clear and easy to understand?

To a very small extent ○ To a small extent ○ Somewhat ○ To a large extent ○ To a very large extent ○

To what extent are the anchor points (1-5) in this item clear and easy to understand?

To a very small extent ○ To a small extent ○ Somewhat ○ To a large extent ○ To a very large extent ○

To what extent does this item describe an aspect of collaborative empiricism that you routinely use in your professional practice?

To a very small extent ○ To a small extent ○ Somewhat ○ To a great extent ○ To a very great extent ○

Comments:

CES Item 4

Please read "Item D - Designing an Empirical Test" and answer the underlined Questions for Expert Feedback below.

Item D – Designing an Empirical Test

A well-designed empirical test of a thought or thought process, whether it is client or therapist initiated, requires clear criteria for interpreting the outcome of the test. This could include specifying criteria for what would be expected to happen if the thought were functional, helpful, or perspective shifting for the client; or criteria for evaluating the converse. A good empirical test is able to distinguish between events occurring as a function of the specific thoughts under evaluation and events occurring as a function of other factors. Anticipating difficulties (ranging from practical and environmental, to those connected to the client's other beliefs and compensatory behaviors) is also part of designing a good empirical test.

D	HOW WELL did the client and therapist design the empirical test?	
1	CRITERIA for possible outcomes of the test were NOT SPECIFIED (e.g., "Therapist: Let's see what happens if you start a conversation at the party and we'll discuss it next week." / Client: "I'll try doing it differently in the meeting, maybe that will help.")	
2	Possible OUTCOMES of the test WERE DISCUSSED, but CRITERIA WERE NOT SPECIFIED (e.g., Therapist: "You might find that other people are also very nervous before speaking in the meeting." / Client: "I'll check whether people talk to me at the party or not.")	
3	CRITERIA WERE SPECIFIED for interpreting outcomes of the test (i.e., the meaning of possible outcomes was discussed). However, there were SOME PROBLEMS that reduced the adequacy of the empirical basis of the criteria (e.g., the criteria were not specified in detail with reference to time, place, or amount: "Some people will talk to me at the party." vs. "At least 3 people will talk to me after I say hello."; criteria were ambiguous or open to alternative explanation; target outcomes were not publicly observable; target outcomes might occur by chance).	
4	OPERATIONAL CRITERIA WERE SPECIFIED for interpreting outcomes of the test. Potential difficulties and solutions WERE DISCUSSED. There were only MINOR PROBLEMS, which did not reduce the adequacy of the criteria (e.g., some variability in focus on the task; some concern of the client's about the test remained unaddressed; solutions were developed for anticipated problems but the client remained only partially confident).	
5	OPERATIONAL CRITERIA WERE SPECIFIED for interpreting outcomes of the test. Potential difficulties and solutions WERE DISCUSSED. There were NO PROBLEMS, or any difficulties WERE RESOLVED.	

Appendix C: Expert Feedback Questionnaire (cont.)

Questions for Expert Feedback:

To what extent does the item description accurately reflect an element of collaborative empiricism?

To a very small extent ○ To a small extent ○ Somewhat ○ To a large extent ○ To a very large extent ○

To what extent does this item capture variability in collaborative empiricism?

To a very small extent ○ To a small extent ○ Somewhat ○ To a large extent ○ To a very large extent ○

To what extent is this item clear and easy to understand?

To a very small extent ○ To a small extent ○ Somewhat ○ To a large extent ○ To a very large extent ○

To what extent are the anchor points (1-5) in this item clear and easy to understand?

To a very small extent ○ To a small extent ○ Somewhat ○ To a large extent ○ To a very large extent ○

To what extent does this item describe an aspect of collaborative empiricism that you routinely use in your professional practice?

To a very small extent ○ To a small extent ○ Somewhat ○ To a large extent ○ To a very large extent ○

Comments:

CES Item 5

Please read "Item E - Evaluating an Empirical Test" and answer the underlined Questions for Expert Feedback below.

Item E – Evaluating an Empirical Test

The meaning of the results of an empirical test are to be assessed in terms of the criteria used in designing the test. This includes distinguishing the actual results of the test (e.g., thoughts, feelings, bodily sensations, behavior of the client and others) from judgments or attributions about the results of the test. A thorough evaluation of an empirical test also involves reviewing problems that arose in carrying it out, ranging from practical and environmental, to those connected to the client's other beliefs and compensatory behaviors (e.g., mistaking interpretations for evidence, distorting evidence in recall, avoiding an anxiety producing situation and thereby missing evidence).

E	HOW WELL did the client and therapist evaluate the results of an empirical test?
1	The results of an empirical test were discussed WITHOUT REFERENCE TO CRITERIA (e.g., Therapist: "How did it go at the meeting?" Client: "It was a lot better." Therapist: "Good, so what did you learn from that?").
2	The results of an empirical test were discussed WITH REFERENCE to criteria, BUT CRITERIA WERE VAGUE, NOT SPECIFIC, OR AMBIGUOUS. There was little or no discussion of potential problems (e.g., Therapist: "How did it go at the meeting?" Client: "I presented better. I think more people listened." Therapist: "Good, so what did you learn from that?")
3	The results of an empirical test were DISCUSSED IN TERMS OF CRITERIA (e.g., Therapist: "How did it go at the meeting?" Client: "No one was critical, and 4 people complimented the presentation." Therapist: "What did we decide that would mean about your ability to present your report?")

Appendix C: Expert Feedback Questionnaire (cont.)

	However, there were SOME PROBLEMS that reduced the adequacy of the empirical basis of the evaluation (e.g., The actual results of the test were not clearly determined; compensatory "safety" behaviors were not explored; difficulties that arose during the empirical test were not discussed; or there were potential biases in reporting the results).	
4	The results of an empirical test were DISCUSSED IN TERMS OF CRITERIA. There were only MINOR PROBLEMS that did not reduce the adequacy of the empirical basis of the evaluation (e.g., some variability in focus on the task; an opportunity for learning from the experiment remained unexplored; some doubt or concern about the meaning of the results remained unaddressed; the results were evaluated well, but their practical implications remained undiscussed).	
5	The results of an empirical test were DISCUSSED IN TERMS OF CRITERIA. Criteria were sourced from the client, and the actual results were accurately determined. Difficulties that arose during the test were discussed. There were NO PROBLEMS, or any difficulties WERE RESOLVED.	

Questions for Expert Feedback:

To what extent does the item description accurately reflect an element of collaborative empiricism?

To a very small extent ○ To a small extent ○ Somewhat ○ To a large extent ○ To a very large extent ○

To what extent does this item capture variability in collaborative empiricism?

To a very small extent ○ To a small extent ○ Somewhat ○ To a large extent ○ To a very large extent ○

To what extent is this item clear and easy to understand?

To a very small extent ○ To a small extent ○ Somewhat ○ To a large extent ○ To a very large extent ○

To what extent are the anchor points (1-5) in this item clear and easy to understand?

To a very small extent ○ To a small extent ○ Somewhat ○ To a large extent ○ To a very large extent ○

To what extent does this item describe an aspect of collaborative empiricism that you routinely use in your professional practice?

To a very small extent ○ To a small extent ○ Somewhat ○ To a large extent ○ To a very large extent ○

Comments:

CES Item 6

Please read "Item F - Developing an Alternative Interpretation of a Thought or Thought Process" and answer the underlined Questions for Expert Feedback below.

Item F – Developing an Alternative Interpretation of a Thought or Thought Process

This item measures the development of a plausible alternative interpretation of a thought or thought process. This may involve, for example, looking at a situation from different perspectives, imagining what other people may think in the same situation, or brainstorming other explanations for what the client believes happened in a situation. Developing an alternative interpretation can occur both through questioning or discussion, and through empirical tests.

F	**HOW WELL did the client and therapist develop an alternative interpretation of a cognition?**	

Appendix C: Expert Feedback Questionnaire (cont.)

1	Discussion of an alternative cognition was NOT RELATED TO THE CLIENT'S EXPERIENCE (e.g., Therapist: "What would an alternative be that wasn't 'black and white' thinking?")	
2	Discussion of an alternative interpretation of a cognition was RELATED IN PASSING to the client's experience (e.g., Therapist: "Has that happened every single time?" Client: "No, not every time I guess." Therapist: "Okay, so it's not inevitable then?")	
3	Development of an alternative interpretation was RELATED IN DETAIL to the client's experience of the cognition (e.g., Therapist: "So, when that didn't happen, what went through your mind?" Client: "I thought, 'It may not be great, but I can definitely do this.'" Therapist: "So how else might you interpret being nervous, rather than it meaning you can't do it?") . However, there were SOME PROBLEMS that reduced the adequacy of the empirical basis of the alternative interpretation (e.g., Interpretations of evidence were mistaken as facts; the client did not rate the strength of belief or emotion associated with the alternative interpretation).	
4	Development of an alternative interpretation was RELATED IN DETAIL to the client's experience of the cognition. There were only MINOR PROBLEMS that did not reduce the adequacy of the empirical basis of the alternative interpretation (e.g., Some variability in focus on the task; the client expressed some doubt or reservation about the alternative interpretation, which was not explored).	
5	Development of an alternative interpretation was RELATED IN DETAIL to the client's experience of the cognition. There were NO PROBLEMS, or any difficulties WERE RESOLVED.	

Questions for Expert Feedback:

To what extent does the item description accurately reflect an element of collaborative empiricism?

To a very small extent ○ To a small extent ○ Somewhat ○ To a large extent ○ To a very large extent ○

To what extent does this item capture variability in collaborative empiricism?

To a very small extent ○ To a small extent ○ Somewhat ○ To a large extent ○ To a very large extent ○

To what extent is this item clear and easy to understand?

To a very small extent ○ To a small extent ○ Somewhat ○ To a large extent ○ To a very large extent ○

To what extent are the anchor points (1-5) in this item clear and easy to understand?

To a very small extent ○ To a small extent ○ Somewhat ○ To a large extent ○ To a very large extent ○

To what extent does this item describe an aspect of collaborative empiricism that you routinely use in your professional practice?

To a very small extent ○ To a small extent ○ Somewhat ○ To a large extent ○ To a very large extent ○

Comments:

CES Item 7

Please read "Item G - Collaboration in Empiricism" and answer the underlined Questions for Expert Feedback below.

Item G – Collaboration in Empiricism

With regards to the specific collaborative empiricism event just rated, please rate the level of collaboration (genuinely shared work) as part of the event. Focus specifically on the sharing of decisions, engagement, participation, effort, and responsiveness, not simply whether one party leads the discussion or makes more suggestions.

Appendix C: Expert Feedback Questionnaire (cont.)

G	**For the specific empirical event just rated using one of Items A – F: HOW WELL did the client and therapist collaborate on the empirical exploration?**	
1	There was evidence of client and therapist active shared engagement and participation in empirical exploration, but this engagement would be characterized as GENERALLY ONE SIDED (e.g., one party ignored the other's contributions, generally made unilateral decisions, or dominated activity in the session).	
2	There was evidence of client and therapist shared active engagement in empirical exploration, but there was still MODERATE IMBALANCE of shared engagement in this session (e.g., one party was moderately controlling, discouraging, or disinterested in the other party's involvement, and made unilateral decisions in the process).	
3	There was evidence of SOME IMBALANCE in client and therapists' shared active engagement in empirical exploration in this session (e.g., one party did not allow the other time for creative deliberation; focused on their own perspective; or rushed the other's opportunity to contribute to decision making).	
4	There was evidence of client and therapist shared active engagement in empirical exploration and only MINOR IMBALANCE in this session (e.g., asked open questions; sought/ provided suggestions, opinions, ideas, alternatives, evaluated ideas or empirical tests; gave the other time to formulate a contribution in the decision making process).	
5	There was an IDEAL BALANCE in client and therapist shared active engagement in empirical evaluation (e.g., both contributed or allowed opportunity for the other's contribution; responded to each other's contributions; shared decision making; or addressed either party's difficulty or lack of involvement).	

Questions for Expert Feedback:

To what extent does the item description accurately reflect an element of collaborative empiricism?

To a very small extent ○ To a small extent ○ Somewhat ○ To a large extent ○ To a very large extent ○

To what extent does this item capture variability in collaborative empiricism?

To a very small extent ○ To a small extent ○ Somewhat ○ To a large extent ○ To a very large extent ○

To what extent is this item clear and easy to understand?

To a very small extent ○ To a small extent ○ Somewhat ○ To a large extent ○ To a very large extent ○

To what extent are the anchor points (1-5) in this item clear and easy to understand?

To a very small extent ○ To a small extent ○ Somewhat ○ To a large extent ○ To a very large extent ○

To what extent does this item describe an aspect of collaborative empiricism that you routinely use in your professional practice?

To a very small extent ○ To a small extent ○ Somewhat ○ To a large extent ○ To a very large extent ○

Comments:

Any further comments on the Scale overall:

Survey Results Request

Appendix C: Expert Feedback Questionnaire (cont.)

We greatly appreciate your completing this survey. Thank you!

If you would like to receive the results, please tick yes and provide an email address below. Your email address will not be communicated to any other party, and will be deleted once the survey results are sent.

○ Yes, please send the survey results to the email address provided in the text box below:

Appendix D: Participant Information Sheet

LA TROBE UNIVERSITY
School of Psychological Science

The Development And Evaluation Of A Measure Of Collaborative Empiricism In Cognitive Therapy

Researcher: John Tee. email: jmtee@students.latrobe.edu.au. Tel. 9479 2543
Course of Study: Doctorate in Clinical Psychology
Project Supervisor: Dr. Nikolaos Kazantzis. email n.kazantzis@latrobe.edu.au Tel. 9479 3670

The aim of this study is to develop and evaluate a measure of collaborative empiricism. Collaborative empiricism is the basic form of interaction underlying the client-therapist relationship in cognitive behavioural therapy. It involves the client and therapist working collaboratively to understand and evaluate the client's beliefs and thoughts, and the relation between thoughts, feelings, and behaviours. Collaborative empiricism is frequently referred to as an essential component of Cognitive Behavioural Therapy. Developing a measure will allow researchers better understand how collaborative empiricism functions in cognitive therapy.

You will be asked to complete a brief email survey of proposed items to be used in the measure. The survey will take approximately 5 – 10 minutes. You will be asked to rate each test item on a 5 point scale, and note any specific comments at the end of the survey.

Participation in the survey will be taken as consent for the researchers to use aggregate and de-identified results from the survey in order to analyse the measure of collaborative empiricism being developed. All information collected in the study will remain strictly confidential. No identifying information will be placed on data records. Data records will be kept in a locked storage room, and computer files will be password protected. It will not be possible for survey data to be withdrawn subsequently because it is not identified.
Data from the research will be kept for possible future use for another project, performed by the same student or supervisor. Collated results of the study, without information that could identify any individual participant, will appear in a doctoral thesis and may be presented at conferences and published in scientific journals.

You are free to withdraw consent and discontinue participation in the study at any time, without adverse consequence or disadvantage.

You may receive a summary of the outcome of this research upon request.

Any questions regarding this study can be directed to the principle researcher, John Tee, or the research supervisor, Dr. Kazantzis, on the email or telephone numbers above.

Appendix D: Participant Information Sheet (cont.)

In the event that you have any complaint about the way you have been treated during the study, or have a query that the supervisor has been unable to satisfy, you may contact The Secretary, Human Ethics, Faculty of Science, Technology and Engineering, La Trobe University, Bundoora, 3086 (ph: 03 9479 3698, e-mail: fstehumanethics@latrobe.edu.au).

Appendix E: Recruitment Emails for Expert Feedback

INVITATION EMAIL

Subject: Collaborative Empiricism in CBT

[DATE], 2011

Dear Dr. [NAME]

We are writing to you today, as a clinical researcher who has recently published data on important processes or outcomes of Cognitive Therapy, commonly referred to as Cognitive Behaviour Therapy (CBT).

We are currently working on an innovative research project, designed to further the community's understanding of Collaborative Empiricism – the specific therapeutic relationship element in CBT. The research team comprises John Tee (Doctoral Candidate), Drs. Nikolaos Kazantzis and Art Stukas (Supervisory Team), and our chief collaborator is Dr. Keith Dobson (University of Calgary, Canada).

We have designed a new measure of Collaborative Empiricism, and we thought you would be interested to provide expert feedback on its content and structure.

Out of respect for your busy schedule, and your valued expertise, we are only seeking your response to a limited number of feedback questions on face validity. We have constructed a very brief on-line questionnaire, which can be accessed via the weblink:

http://latrobepsy.qualtrics.com/SE/?SID=SV_a4ALvUYoC1ouXoE

We thought that you would be eminently placed to contribute expert feedback on our scale. We recognize that you are extremely busy, and like the other colleagues who have published in premier clinical psychology outlet we have invited to participate in this questionnaire, have limited time to take on new tasks. However, we hope that the merits of this research make this an appealing opportunity.

In case you have not seen it already, we thought you would appreciate a PDF reprint of our 2010 article appearing in *Clinical Psychology: Science and Practice,* which conveys much of the thinking behind the scale.

This expert questionnaire has been approved by a La Trobe University Human Ethics Committee. All responses will be anonymous. Further detail on the ethical conditions for this expert feedback questionnaire is included below.

We hope that you will accept this invitation, and we look forward to your feedback on our collaborative empiricism scale.

Sincerely,

John Tee
Doctoral Candidate
School of Psychological Science
La Trobe University

Appendix E: Recruitment Emails for Expert Feedback (cont.)

Dr. Nikolaos Kazantzis and Dr. Art Stukas
Supervisory Team
School of Psychological Science
La Trobe University, Australia

Dr. Keith S. Dobson
Collaborator
Department of Psychology
University of Calgary, Canada

Attachment: CPSP Reprint

--

Ethical Conditions for the Expert Feedback Questionnaire

Responses will be automatically de-identified by the software and no identifying information will be placed on data records. Data records will be kept in a locked storage room, and computer files will be password protected. It will not be possible for individual questionnaire data to be withdrawn subsequently because responses will be anonymous. At the end of the questionnaire you may choose to receive the results of the questionnaire by email. In this case, your email address will kept on file until the aggregated results are sent, after which it will be deleted.

REMINDER #1

Subject: Collaborative Empiricism in CBT

[DATE], 2011

Dear Dr. [NAME]

If you have not done so already, we are writing to invite you to provide expert feedback on a new measure on the therapeutic relationship in Cognitive Therapy, commonly referred to as Cognitive-Behavioural Therapy (CBT). If you have already provided your feedback, we are grateful for your feedback, and we hope you found the process to be an interesting and informative one.

If you have not provided comment, we are writing again to seek your input and comment. Your participation is needed in order to provide rich data for this study. As you know, there is significant research on the therapeutic 'alliance', a concept that misses a great deal of the richness in the therapeutic encounter intended in Beckian CBT. It is hoped that the scale we have developed will provide researchers, supervisors, and practitioners with a more theoretically meaningful means of gauging therapeutic process in CBT.

Appendix E: Recruitment Emails for Expert Feedback (cont.)

Out of respect for your busy schedule, and your valued expertise, we are only seeking your response to a limited number of feedback questions on face validity. We have constructed a very brief on-line questionnaire, which can be accessed via the weblink:

http://latrobepsy.qualtrics.com/SE/?SID=SV_a4ALvUYoC1ouXoE

In case you have not seen it already, we thought you would appreciate a PDF reprint of our 2010 article appearing in *Clinical Psychology: Science and Practice,* which conveys much of the thinking behind the scale.

This expert questionnaire has been approved by a La Trobe University Human Ethics Committee. All responses will be anonymous. Further detail on the ethical conditions for this expert feedback questionnaire are included below.

We hope that you will accept this invitation, and we look forward to your feedback on our collaborative empiricism scale.

Sincerely,

John Tee
Doctoral Candidate
School of Psychological Science
La Trobe University

Dr. Nikolaos Kazantzis and Dr. Art Stukas
Supervisory Team
School of Psychological Science
La Trobe University, Australia

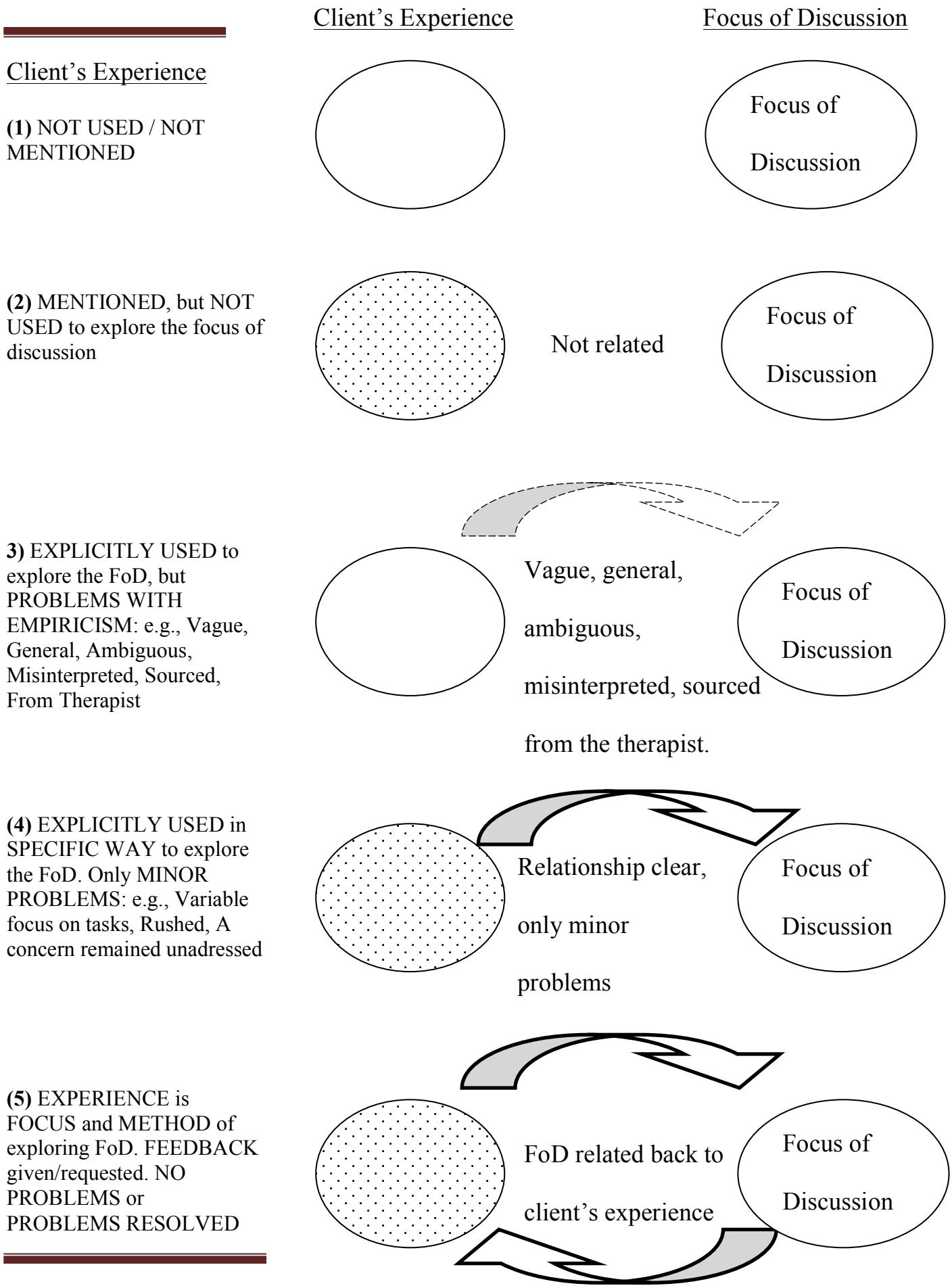
Dr. Keith S. Dobson
Collaborator
Department of Psychology
University of Calgary, Canada

Attachment: CPSP Reprint

-- **Ethical Conditions for the Expert Feedback Questionnaire**

Responses will be automatically de-identified by the software and no identifying information will be placed on data records. Data records will be kept in a locked storage room, and computer files will be password protected. It will not be possible for individual questionnaire data to be withdrawn subsequently because responses will be anonymous. At the end of the questionnaire you may choose to receive the results of the questionnaire by email. In this case, your email address will kept on file until the aggregated results are sent,

Appendix F: Relationship Between CE and the Focus of Discussion



Appendix G: Collaborative empiricism as a predictor of depression in the next session: Regression analyses for Sessions 9 and 15

At session 9, in step 1 of the hierarchical regression, depression at session 9 accounted for a statistically significant 49.5% of the variance in depression at session 10, $R^2 = .495$, adjusted $R^2 = .483$, F (1, 42) = 41.227, $p = .000$. In step 2, CE accounted for an additional, non-significant, 5.0% of the variance in depression at session 10, $\Delta R^2 = .005$, ΔF (1, 41) = 0.427, $p = .517$. The full model of two predictors explained a statistically significant 50.1% of the variance in depression at session 10, $R^2 = .501$, adjusted $R^2 = .476$, F (2, 41) = 20.546, $p = .000$. Regression coefficients and part correlations (sr^2) for each predictor in the session 9 model are presented in Table G1.

Table G1

Unstandardized (B) and Standardized (β) Regression Coefficients, and Squared Semi-Partial Correlations for the Session 9 Model Predicting Depression at Session 10

Predictor	*B* [95% CI]	*β*	sr^2	*p*
Step 1				
Session 9 depression	0.761 [0.522, 1.000]	.704	.704	.000
Step 2				
Session 9 depression	0.782 [0.533, 1.032]	.723	.698	.000
Collaborative empiricism	1.313 [-2.744, 5.371]	.075	.072	.517

As with the session 3 model, both models at session 9 were statistically significant but depression at session 9 in each model was the only significant predictor of depression at session 10.

The results at session 15 were similar. In step 1 of the hierarchical regression, depression at session 15 accounted for a statistically significant 78.8% of the variance in

depression at session 16, R^2 = .788, adjusted R^2 = .782, F (1, 37) = 137.444, p = .000. In step 2, CE accounted for an additional, non-significant, 3.0% of the variance in depression at session 16, ΔR^2 = .003, ΔF (1, 41) = 0.477, p = .494. The full model of two predictors explained a statistically significant 79.1% of the variance in depression at session 16, R^2 = .791, adjusted R^2 = .779, F (2, 41) = 67.988, p = .000. Regression coefficients and part correlations (sr^2) for each predictor in the session 15 model are presented in Table G2.

Table G2

Unstandardized (B) and Standardized (β) Regression Coefficients, and Squared Semi-Partial Correlations for the Session 15 Model Predicting Depression at Session 16

Predictor	*B* [95% CI]	*β*	sr^2	*p*
Step 1				
Session 15 depression	0.843 [0.697, 0.989]	.888	.888	.000
Step 2				
Session 15 depression	0.850 [0.701, 0.998]	.894	.887	.000
Collaborative empiricism	.818 [-1.585, 3.221]	.053	.053	.494

Both models were statistically significant but depression at session 15 was the only significant predictor in each model. These three models indicate that CE did not predict depression severity in the next session.

A similar pattern was observed when working alliance and therapist competence were added into the regression model as step 2, and CE was moved to step 3. At each time point (session 3, 9, 15) depression at the start of the current session was the only significant predictor of depression at the next session.

Appendix H: Working Alliance Inventory – Short Revised (Client Version)

Goal Scale

1. ___and I are working towards mutually agreed upon goals.

2. We agree on what is important for me to work on.

3. ___and I collaborate on setting goals for my therapy.

4. We have established a good understanding of the kind of changes that would be good for me.

Task Scale

5. What I am doing in therapy gives me new ways of looking at my problem.

6. I feel that the things I do in therapy will help me to accomplish the changes that I want.

7. As a result of these sessions I am clearer as to how I might be able to change.

8. I believe the way we are working with my problem is correct.

Bond Scale

9. I believe___likes me.

10. ___and I respect each other.

11. I feel that___appreciates me.

12. I feel___cares about me even when I do things that he/she does not approve of.

Appendix I: Data Access and Publication Agreement

Authorship and Dissemination Agreement

Between -

(i) Principal Supervisor: Dr. N Kazantzis
Co-Supervisor: Dr. J Farhall
Postgraduate Student: Alexandra M. Petrik
Research project title: Theoretical Rationales for Interventions in Cognitive Therapy

(ii) Principal Supervisor: Dr. N Kazantzis
Co-Supervisor: Dr. A Stukas
Postgraduate Student: John M. Tee
Research project title: Collaborative Empiricism in Cognitive Therapy

Date: 7 November, 2012

The purpose of this document is to assist the above named staff of the School of Psychological Science at La Trobe University and individual postgraduate student researchers to fulfill their responsibilities with respect to authorship and dissemination of research. The points outlined below aim to ensure that principles in the Australian Psychological Society (APS) Code of Ethics (2007)[1] and the Australian Code for Responsible Conduct of Research (2007) Part 5 are upheld in the research process.

Authorship

1. All researchers, defined for the purposes of this paper to mean students, supervisors and external collaborators, involved with these projects should discuss authorship on any articles or other outputs resulting from their research as early as possible.

2. All researchers involved with these projects will attribute authorship on articles/ chapters/ scales/ measures or other outputs resulting from their research in a manner that reflects the scientific work performed and ensure that the contribution made is a fair reflection of the work people have actually performed or of what they have contributed, including the substantial intellectual contribution made by a supervisor of student research, and external consultation for the conduct of that research.

3. All researchers involved with this project are aware that the student will usually be listed as principal author on an article that is substantially based on the student's dissertation or thesis. However, where a student has not prepared a suitable manuscript for publication within 12 months following the completion of their research, and/or significant work is required to produce a suitable manuscript for submission to a peer-reviewed journal, the researchers will discuss authorship and possibly arrive at a different authorship agreement.

4. All researchers involved with these projects will be aware that collection of data does not in itself merit authorship on scientific publications. This means that no research assistant paid to assist with the data collection will have rights to authorship on publications unless they have contributed made substantial intellectual contribution to the design and/or other aspects of the research. Similarly, student researchers

Appendix I: Data Access and Publication Agreement (cont.)

collecting data for other projects for other student projects will not have rights to authorship on those other student's publications, unless they have made substantial intellectual contribution to the design and/or other aspects of the research.

Dissemination of New Scales/ Measures Used in the Conduct of Research

5. Students will usually be listed as principal author on a scales/ measures that are substantially based on the student's dissertation or thesis work.

6. All researchers will also agree on the scale/questionnaires name or title.

7. All researchers involved in these projects will specify the purposes and uses of their assessment techniques and clearly indicate the limits of the assessment techniques' applicability. Each researcher consents to other researchers having a right to reproduce and use his/her scales for appropriate clinical and research purposes.

8. All researchers involved with these projects will agree to make the scales/ measures freely available to other competent professionals who seek to use the scales for appropriate clinical and/ or research purposes, provided that: (a) the scales/ measures will be used only for the purpose stated in the verification request/ approved proposal; (b) the scales/ measures will not be used for financial gain; and (c) information on scale/ measure scoring and interpretation is provided by authors to the person who has made the request. All researchers will also agree on the point of contact for the distribution of these resources to others in the professional community.

9. The researchers involved with these projects will not sell the rights to the scales / measures produced to commercial publishers. However, for the avoidance of doubt, scale/ measure authors shall be entitled to enter into agreement with publishers for the publication of scholarly works (i.e., peer reviewed articles and book chapters) reporting on the development of the scales and their properties.

Signatures

Dr. N Kazantzis **Dr. K. S. Dobson** **Dr. J Farhall**

Dr. A Stukas **Dr. D. Farchione**

Alexandra M. Petrik **John M. Tee**

1. Australian Psychological Society. (2007). *Code of ethics*. Melbourne, VIC: Author.

Appendix J: Rater Training PowerPoint

Collaborative Empiricism Scale

Rater Training

John Tee,
Dr. Nikolaos Kazantzis
September 2011

Plan for today

- Definition of Collaborative Empiricism
  - Empiricism in CE
  - Collaboration in CE
- Overview of the scale
- Identifying events to be rated
- Audio Examples
- Discussion of the scale items
- Rate a session together
- Independently rate a session
- Discuss problems and difficulties
- Adjust the rating process and rate another session

The Collaborative Empiricism Scale

- The first measure of Collaborative Empiricism
- Received extensive expert feedback on the measure
- Now: Psychometric evaluation of the measure
- Use the measure to:
  - Investigate the process of CE in cognitive therapy of depression
  - Explore the role of CE in belief change
  - Investigate whether CE relates to outcomes in cognitive therapy for depression

What is collaborative empiricism?

- Some suggestions ?!

What is collaborative empiricism?

- The basic strategy underlying interaction between client and therapist in cognitive therapy.
- **The client's world,** as it appears to them, is viewed as **data** to be **empirically investigated.**
- Work as a **team**, to formulate the **raw data** of the client's thoughts, thought processes, feelings, and behaviour as **hypotheses** that require **systematic empirical exploration**.
- Beliefs are explicitly acknowledged as possibilities that are worth evaluating.

What is collaborative empiricism?

- *Thoughts* refers to mental images, memories, and beliefs.
- *Thought processes* include:
  - Biases in learning or attribution, for example: 'catastrophizing', 'black and white' (all or nothing) thinking, assuming over-responsibility for thoughts
  - Meta-cognitions – that is, beliefs about thoughts – such as "worrying is helpful" or "worrying will get out of my control".

Appendix J: Rater Training PowerPoint (cont.)

What is **collaboration** in CE?

- A genuine **sharing of therapeutic work**.
- Three key components. Client and therapist:
  - **share decision making**
  - are **mutually responsive** to each other
  - **actively contribute** (suggestions, ideas, questions...)
- E.g., Sharing in the design and implementation of agenda items, goals, and interventions.
- **NB:** Collaboration in CE is more than just client and therapist agreement, cooperation, or client adherence/ compliance with therapist suggestions.

Overview of how to use the scale

- Firstly, a **crucial** point:
- The Collaborative Empiricism Scale rates the **quantity and quality of CE in the session** at the current moment.
- A therapist or client might do really well in fostering CE, but the other party may not respond well. In this case, there may be only a low degree of CE in the room.

Overview of Rating

1. Identify the current **focus of discussion**.
   - The focus of discussion is the central topic being discussed.
   - Typically, this focuses on an issue of importance for the client.
   - Involves the **client's experience** (Thoughts / thought processes, feelings, behaviour, physiology, or a situation).

Overview of Rating

2. Decide whether the focus of discussion **acted on** via one of these six actions?
   - **Selecting** it for evaluation / exploration
   - **Rating** the strength of emotion or belief associated with it
   - **Exploring** or evaluating it
   - **Developing** an **alternative interpretation** of it
   - **Designing** an empirical test
   - **Evaluating** an empirical test

Overview of Rating

3. Rate the scale item that matches the action:

   A. Selecting a Focus of Discussion to Explore
   B. Rating Emotion or Belief associated with the Focus of Discussion
   C. Exploring the Focus of Discussion
   D. Developing an **Alternative Interpretation** of the Focus of Discussion
   E. Designing an Empirical Test
   F. Evaluating an Empirical Test

Overview of Rating

4. Rate collaboration for the same event just rated for empiricism:

   G. Collaboration in Empiricism: Sharing the Work

Appendix J: Rater Training PowerPoint (cont.)

The Focus of Discussion

Some examples of the focus of discussion:

1. Client. "I saw my ex-girlfriend at the movies. I felt so angry. I made a real fool of myself."
2. Client. "My friend invited himself to stay. I'm sick of him sponging off me."
3. Therapist. "So it was hard to remember the homework?" Client. "I just didn't think of it…"
4. Client. "There isn't really much point in applying for jobs. I'm too old now anyway."

Item A: Selecting a focus to explore

- Measures how well the CLIENT'S EXPERIENCE was USED TO SELECT the focus of discussion to explore.
- Selecting requires an **explicit decision or deliberate intention** to choose a possible topic of discussion.
- A new topic that simply **flows in a stream** from the last topic is **not rated** as a selection.

Item A: Selecting a Focus

Item A: Selecting a Focus of Discussion to Explore

1) "Should we discuss that today?" A direct **question**.
2) C: "I've been stressing about work." T: "Ok, let's put that on the Agenda." A direct **statement**. (Agenda setting is a selection.)
3) [After discussing work] "So, how are things at home?". An **invitation** to a new topic which changes the current focus of discussion. Note: Do not rate item A if they start talking about home without any request/invitation to a new topic.
4) T: "The other thing you mentioned was your sick friend." [They go on to discuss this]. Referring to an earlier concern of the client's. The therapist has **provoked** a change of topic. Still deliberate.

Item A: Selecting a Focus

- If they select something at 10 mins, and then discuss it again at 15 mins, don't rate it as a new selection if they merely refer to a selection decision they already made.
  - E.g., don't rate a new selection if you hear "Ok, next one".
- Do rate it as a different a selection if they make *another decision* at 15 minutes.
  - If you hear: "Should we do another one of these thoughts?" then this is a *new selection decision*, so rate a new Selection (A).

Item A: Selecting a Focus

- E.g. Working through a list of things to discuss.
  - T: "Should we look at these thoughts (on the thought record) and see how you did with them?"
  - This is a selection to work through several thoughts.
  - As the client and therapist work through the list item after item, a statement like "OK, next one." **IS NOT** a separate selection. The selection of a focus to explore **has already been made**, to work through several thoughts. Now they are working through them.
- Working through several thoughts in a thought record is a single 'Selection' if they make a single decision to work through the list.

Item B: Rating Emotion or Belief associated with the Focus of Discussion

- Client and therapist **deliberately** rate how much emotion is associated with the focus of discussion, or how strongly a thought is believed.
- They may use a verbal description, e.g. "I felt really stressed out"
- They may use a 'subjective scale', i.e. "From 0 – 10", "From 0 – 100", "From the most anxious you've ever felt, to feeling no anxiety at all"
- Degree of belief is typically rated from 0 – 100, e.g. T: "How much do you believe that things will get out of control if you delay worrying?" C: "90%"

Appendix J: Rater Training PowerPoint (cont.)

Item B. Examples

Item B: Rating Emotion or Belief associated with the Focus of Discussion

Measuring the amount or degree of emotion or belief

1. T: How anxious do you feel when you think you won't get the job? C: Pretty anxious. T: Out of 10?
2. T: Which of those thoughts do you feel most stressed about? C: Getting fired. T: More than being judged by colleagues?
3. T: How much do you believe this thought now? C: 90%
4. C: After the homework, the thought shifted a bit. T: How much would you say? C: Maybe from 80 to 50%

Item B. Examples

Item B: Rating Emotion or Belief associated with the Focus of Discussion

- Yes:
  - T: How anxious did you feel? C: Really anxious
- No:
  - I went to a party on the weekend. I felt really anxious.

In this case there is no deliberate intention or decision to rate the degree of emotion (or strength of belief).

Item C: Exploring the Focus of Discussion

- Exploring means the purposeful seeking and use of information and experience, IN ORDER TO EXAMINE the functionality or helpfulness of the current focus of discussion
- If it doesn't do this, it isn't an 'Explore'
- HOW was the client's experience used to explore the focus of discussion?

Item C. Examples

Item C. Exploring the Focus of Discussion

1. T: What might other people do in that situation?
2. C: I still felt shy at the party. I was doing ok until this cute guy came into the kitchen (sigh). T: What happened? C: I felt stupid, couldn't say anything. I left the room. T: Did you notice any particular automatic thoughts?
3. How helpful was this thought, 'You couldn't do it'?"
4. Therapist: "So you felt worried writing your last report?" Client: "Terrified the whole time that I couldn't do it." Therapist: "What happened?" Client: "Well, I got it done the night before. It was 'ok', I guess."

Item C: Exploring the Focus of Discussion

- What **IS NOT** an explore?
  - Listing the week's events
  - Recapping or summarising a discussion
  - Psychoeducation isn't an explore, if no examination is made. For example, teaching how to do a thought record
  - Making an appointment for next week

Item C: Exploring the Focus of Discussion

- Sometimes the focus of discussion changes, without deliberately selecting a new focus of discussion.
- The discussion flows from the current focus of discussion, *without a decision* to select the new focus.
- There is no selection (Item A).

Appendix J: Rater Training PowerPoint (cont.)

Item C: Exploring the Focus of Discussion

- Rules for when to rate a new Item C:
- A part of a discussion is the **same** Explore (C) if it is:
  - Another **perspective** on the current 'Explore'
    - Other people's views
    - Past experiences with the same focus
    - Thoughts, feelings about the current explore)
  - An **example** of the current 'Explore'
  - Further **detail** of the current 'Explore'

Selecting (A) and Exploring (C)

- Selections (A) are **independent** of Explores (C).
- (A) without (C)
  - If they make a Selection (A), but then don't Explore it (C), it is still a selection.
- (C) without (A)
  - If they start exploring something (C) *without making a decision* to select that as a focus to explore, then you would rate (C) but not (A).

Item D: Developing an Alternative Interpretation of the Focus of Discussion

- Involves the **deliberate development** of a plausible alternative interpretation of the focus of discussion.
- Can occur through both questioning / discussion, and empirical / behavioral tests.
  - Looking at a situation from different perspectives
  - Imagining what other people may think in the same situation
  - Brainstorming other explanations for what the client believes happened in a situation.

Item D. Examples

Item D. Developing an Alternative Interpretation of the Focus of Discussion

1. T: What might finishing the essay mean about the idea that "you're going to fail uni"?
2. T: So, you didn't fail? C: No. It wasn't very good, but at least I got it done. T: What might this mean about the idea, "I can't do it?"
3. An alternative belief arising out of a thought record.

Item D. Examples

- These *are not* item D.
  - "Is it rational or irrational?" "Irrational."
  - "Maybe your mother was trying to be helpful?" "Yeah, I guess."
- They are in passing comments, *without deliberate consideration* of evidence, information, perspectives, or client experience.

Exploring (C) and Developing an Alternative (D)

- (C) then (D)
- When they start exploring something without mentioning the development an alternative interpretation, it is Exploring (C). If they then move to developing an alternative interpretation, it is now a new Item (D).
- Only (D)
- If they go straight away into developing an alternative interpretation (D), then only rate (D). There is no Exploring Item (C) rated in this case because all of the 'exploring' is directed at developing an alternative interpretation.

Appendix J: Rater Training PowerPoint (cont.)

Item E: Designing an Empirical Test

- An empirical test is an experiment to test out a thought or belief, typically contrasted with an alternative.
- A good test has:
  - A clear purpose / rationale
  - A clear target idea to be tested
  - An alternative belief
  - A prediction of the outcome of the test
  - A rating of the degree of belief in the target idea
  - Specific details about when/where/how
  - A plan for problems that might get in the way

Item E. Examples

Item E. Designing an Empirical Test

- Commonly a behavioural experiment
- Can be set up for homework, but also occurs in session
- Examples:
  - Client will write down her worries during the day, and plan to delay worrying about them until 'worry time' after dinner (evaluating, "I need to respond to worries straight away").
  - Client will ask for help filling in a simple form at the post office (testing, "Other people will judge me harshly if I don't know what I'm doing").

Item F: Evaluating an Empirical Test

- The meaning of the results of an empirical test are assessed in terms of CRITERIA used in designing the test.
- Distinguishing the actual results of the test (e.g., thoughts, feelings, bodily sensations, behaviour of the client and others) from judgments or attributions about the results of the test.
- Reviewing problems that arose, ranging from practical and environmental, to those connected to the client's other beliefs and compensatory behaviours (e.g., mistaking interpretations for evidence, distorting evidence in recall, avoiding an anxiety producing situation and thereby missing evidence).

Item F. Examples

Item F. Evaluating an Empirical Test

- Working out what happened, and deciding what it means.
- Examples:
  - T: How did you go with the 'worry list'? C: I wrote down my worries during the day, and then looked at them at 8pm. T: What did you observe? C: Well, actually, I felt a bit less worried because I knew I wasn't going to forget anything important.
  - C: Most people I asked said they felt nervous when the boss reviewed their work. But when I watched them, they didn't look nervous to me.

Rating Collaboration: Item G

- This item is used to rate the level of collaboration (**sharing the work**) in the empirical event.
- Focus specifically on the *sharing* of decisions, engagement, participation, effort, and *responsiveness*, not simply whether one party leads the discussion or makes more suggestions.
- Key term: BALANCE / IMBALANCE
  - These terms relate to the LEVEL OF ENGAGEMENT. This is not necessarily the same thing as who did the most talking.

Walkthrough

- See DocCam

Appendix J: Rater Training PowerPoint (cont.)

Practice 1: Rating an audio clip

- During this audio clip:
- Identify the focus of discussion
- The start times
- The **type** of item to be rated (A – F)

- We will then discuss!

Practice 2: Rating an audio clip

- During this audio clip:
- Identify the focus of discussion
- The start times
- **Rate** each item (A – F)

- We will then discuss!

Practice 3: Rating the whole scale

- During this audio clip:
- Identify the focus of discussion
- The start times
- Rate each item (A – F)
- For each event, rate Item G

- We will then discuss!

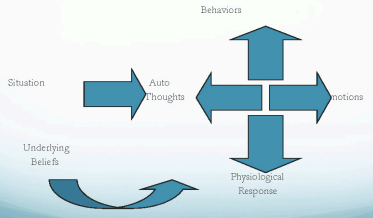
Practice 4: Rating a session

- During this audio clip:
- Identify the focus of discussion
- The start times
- Rate each item (A – F)
- For each event, rate Item G

- We will then discuss!

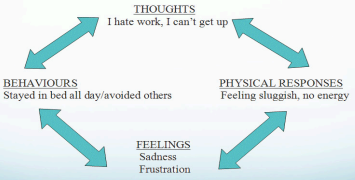
The Cognitive Model

Thoughts, feelings, behaviour, physiology, and situations are **mutually determined and interactive**.



The Cognitive Model

- "I hate work, I can't get up"



Appendix J: Rater Training PowerPoint (cont.)

The Cognitive Model

- Our thoughts about events; our beliefs about self, others, and the world; our perception of control and self-efficacy, all influence our behaviours and emotions when responding to a situation.
- The depression "negative cognitive triad":
- Negative view of self (inadequate, unworthy)
- Negative view of the others/world (Defeating or depriving)
- Negative view of the future (Expecting the worst, catastrophizing).
- Cognitive therapy **explores this relationship**

Appendix K: Regression of CES on BDI, controlling for PreBDI, WAI, and CTRS

This appendix reports the results of a hierarchical linear regression analysis of CE as a predictor of depression at 0 months post-therapy, controlling for the effect of pre-therapy depression, the working alliance, and therapist competence. The linear regression analysis is presented as a more frequently used alternative to the linear mixed model presented in Study 3. The linear mixed model has the advantage that it accounts in a single analysis for CE as a predictor of depression at the five time points post-therapy, whereas the linear regression is restricted to an analysis of CE as a predictor of depression at 0 months post-therapy only. The linear regression has the advantage that it provides an effect size in a format (ΔR^2) that is readily compared with effect sizes for other processes variables in the literature, such as the working alliance (WAI) and therapist competence (CTRS). The results of the linear regression analysis presented here are in close agreement with the results of the linear mixed model reported in Chapter 8.

The assumptions of hierarchical regression were tested before interpreting the output of analyses. Univariate outliers were treated during preliminary data analysis. Inspection of the normal probability plot of standardized residuals and the scatterplot of standardized residuals against standardized predicted values indicated that the assumptions of normality, linearity, and homoscedasticity of residuals were met. One case had a Mahalanobis distance marginally exceeding the critical χ^2 for $df = 4$ (at $\alpha = .001$) of 18.467, indicating that this case was a multivariate outlier. Cook's distance for this case was 0.062, indicating that this case had only a very small influence on the regression model as a whole. Accordingly, this case was retained unchanged. As an additional check, re-running the regression analysis with this case deleted did not change the statistical significance of any predictor. Finally, all predictors in the model had tolerances greater than .927, indicating that multicollinearity would not affect the interpretability of the regression analysis.

To investigate the individual contribution of each therapy process variable as a predictor of depression severity post-therapy, the hierarchical regression was conducted in four steps. Step 1 regressed BDI onto pre-therapy BDI, to control for the effect of initial depression. Step 2 added working alliance (WAI) to the regression, step 3 added therapist competence (CTRS), and step 4 added collaborative empiricism (CES). The results showed that in step 1, pre-therapy depression accounted for a non-significant 4.7% of the variance in post-therapy depression, $R^2 = .047$, adjusted $R^2 = .024$, $F\ (1, 42) = 2.051$, $p = .159$. In step 2, working alliance accounted for an additional, statistically significant, 9.3% of the variance in post-therapy depression, $\Delta R^2 = .093$, $\Delta F\ (1, 41) = 4.415$, $p = .042$. In step 3, therapist competence accounted for an additional, but not statistically significant, 0.8% of the variance in post-therapy depression, $\Delta R^2 = .008$, $\Delta F\ (1, 40) = 0.358$, $p = .553$. In step 4, the CES accounted for an additional, and statistically significant, 8.8% of the variance in post-therapy depression, $\Delta R^2 = .088$, $\Delta F\ (1, 39) = 4.499$, $p = .030$. This indicates that CE explained 8.8% of unique variance in post-therapy depression, beyond that explained by other predictors in the model. The full (step 4) model of four predictors explained a statistically significant 23.5% of the variance in BDI post-therapy, $R^2 = .235$, adjusted $R^2 = .157$, $F\ (4, 39) = 2.940$, $p = .030$. This shows that the full model was a statistically significant predictor of depression severity post-therapy. Regression coefficients and squared semi-partial correlations (sr^2) for each predictor in the model are presented in Table K1.

Table K1

Unstandardized (B) and Standardized (β) Regression Coefficients, and Squared Semi-Partial Correlations for Each Predictor in a Regression Model Predicting Depression Severity

Predictor	*B* [95% CI]	*β*	sr^2	*p*
Step 1				
Pre-therapy depression	0.253 [-0.103, 0.609]	.216	.047	.159
Step 2				
Pre-therapy depression	0.222 [-0.122, 0.566]	.189	.036	.200
Working alliance	0.619 [0.024, 1.213]*	.306	.092	.042
Step 3				
Pre-therapy depression	0.242 [-0.112, 0.595]	.206	.041	.175
Working alliance	0.736 [0.017, 1.454]*	.363	.091	.045
Therapist competence	0.177 [-0.421, 0.775]	.107	.008	.553
Step 4				
Pre-therapy depression	0.272 [-0.068, 0.613]	.219	.051	.114
Working alliance	0.628 [-0.069, 1.325]	.250	.065	.076
Therapist competence	0.227 [-0.348, 0.803]	.097	.013	.429
Collaborative empiricism	-5.141 [-10.044, -.238]*	-.306	.088	.040

Note. N = 44. * = $p < .05$.

As seen in Table K1, in the step 2 model, working alliance was a significant predictor of post-therapy depression. Collaborative empiricism was the only significant predictor in the full (step 4) model. The results indicate that an increase of 1 point on the CES was associated with a decrease of 5.141 points post-therapy on the BDI, after controlling for the effect of working alliance and therapist competence.

References

Abramson, L. Y., Alloy, L. B., & Dykman, B. M. (1990). Cognitive theory and therapy of emotional disorders: A lay epistemic reconstruction? *Psychological Inquiry, 1*, 198-200.

Ackerman, S. J., & Hilsenroth, M. J. (2001). A review of therapist characteristics and techniques negatively impacting the therapeutic alliance. *Psychotherapy: Theory, Research, Practice, Training, 38*, 171-185.

Ackerman, S. J., & Hilsenroth, M. J. (2003). A review of therapist characteristics and techniques positively impacting the therapeutic alliance. *Clinical Psychology Review, 23*, 1-33.

Adams, J. R., & Drake, R. E. (2006). Shared Decision-Making and Evidence-Based Practice. *Community Mental Health Journal, 42*, 87-105. doi: http://dx.doi.org/10.1007/s10597-005-9005-8

Adler, M., & Ziglio, E. (1996). *Gazing into the Oracle: The Delphi method and its application to social policy and public health*. London: Kingsley.

Agresti, A. (2002). *Categorical data analysis* (2nd ed.). New York: Wiley.

Alford, B. A., & Beck, A. T. (1994). Cognitive therapy of delusional beliefs. *Behaviour Research and Therapy, 32*, 369-380.

Allen, J. G., Newsom, G. E., Gabbard, G. O., & Coyne, L. (1984). Scales to assess the therapeutic alliance from a psychoanalytic perspective. *Bulletin of the Menninger Clinic, 48*, 383-400.

Arnkoff, D. B. (2000). Two examples of strains in the therapeutic alliance in an integrative cognitive therapy. *Journal of Clinical Psychology, 56*, 187-200. doi: http://dx.doi.org/10.1002/%28SICI%291097-4679%28200002%2956:2%3C187::AID-JCLP5%3E3.0.CO;2-Y

Aspland, H., Llewelyn, S., Hardy, G. E., Barkham, M., & Stiles, W. (2008). Alliance ruptures and rupture resolution in cognitive-behavior therapy: A preliminary task analysis. *Psychotherapy Research, 18*, 699-710. doi: http://dx.doi.org/10.1080/10503300802291463

Ayers, S. (Ed.). (2007). *Cambridge handbook of psychology, health and medicine*. New York: Cambridge University Press.

Bachelor, A. (1991). Comparison and relationship to outcome of diverse dimensions of the helping alliance as seen by client and therapist. *Psychotherapy: Theory, Research, Practice, Training, 28*, 534-539.

Bachelor, A., Laverdière, O., Gamache, D., & Bordeleau, V. (2007). Clients' collaboration in therapy: Self-perceptions and relationships with client psychological functioning, interpersonal relations, and motivation. *Psychotherapy: Theory, Research, Practice, Training, 44*, 175-192.

Baer, L. (2010). Handbook of Clinical Rating Scales and Assessment in Psychiatry and Mental Health Retrieved from http://latrobe.eblib.com.au/patron/FullRecord.aspx?p=510459

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.

Barber, J. P., Crits-Christoph, P., & Luborsky, L. (1996). Effects of therapist adherence and competence on patient outcome in brief dynamic therapy. *Journal of Consulting and Clinical Psychology, 64*, 619-622. doi: http://dx.doi.org/10.1037/0022-006X.64.3.619

Barber, J. P., & DeRubeis, R. J. (1992). The ways of responding: A scale to assess compensatory skills taught in cognitive therapy. *Behavioral Assessment, 14*, 93-115.

Barber, J. P., & DeRubeis, R. J. (2001). Change in compensatory skills in cognitive therapy for depression. *Journal of Psychotherapy Practice & Research, 10*, 8-13.

Barber, J. P., Liese, B. S., & Abrams, M. J. (2003). Development of the Cognitive Therapy Adherence and Competence scale. *Psychotherapy Research, 13*, 205-221. doi: http://dx.doi.org/10.1093/ptr/kpg019

Barkham, M., Mullin, T., Leach, C., Stiles, W. B., & Lucock, M. (2007). Stability of the CORE-OM and the BDI-I prior to therapy: Evidence from routine practice. *Psychology and Psychotherapy: Theory, Research and Practice, 80*, 269-278. doi: http://dx.doi.org/10.1348/147608306X148048

Baron, J. (2007). *Thinking and deciding*. New York: Cambridge University Press.

Beck, A. T. (1967). *Depression: Clinical, experimental, and theoretical aspects*. New York: Harper and Row.

Beck, A. T. (1970). Cognitive therapy: Nature and relation to behavior therapy. *Behavior Therapy, 1*, 184-200.

Beck, A. T. (1976). *Cognitive therapy and the emotional disorders*. New York: International Universities Press.

Beck, A. T. (1989). An interview with a depressed and suicidal patient. In D. Wedding & R. J. Corsini (Eds.), *Case studies in psychotherapy* (pp. 125-142). Itasca, IL: F E Peacock.

Beck, A. T., & Dozois, D. J. A. (2011). Cognitive Therapy: Current Status and Future Directions. *Annual Review of Medicine, 62*, 397-409. doi: 10.1146/annurev-med-052209-100032

Beck, A. T., Emery, G., & Greenberg, R. L. (1985). *Anxiety disorders and phobias: A cognitive perspective*. New York: Basic Books.

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

Beck, A. T., Steer, R. A., & Garbin, M. G. (1988). Psychometric properties of the Beck Depression Inventory: Twenty-five years of evaluation. *Clinical Psychology Review, 8*, 77-100. doi: http://dx.doi.org/10.1016/0272-7358%2888%2990050-5

Beck, A. T., Ward, C. H., Mendelson, M., Mock, J., & Erbaugh, J. (1961). An Inventory for Measuring Depression. *Archives of General Psychiatry, 4*, 561-571.

Beck, C. T., & Gable, R. K. (2001). Ensuring content validity: An illustration of the process. *Journal of Nursing Measurement, 9*, 201-215.

Beck, J. S. (1995). *Cognitive therapy: Basics and beyond*. New York: Guilford.

Beck, J. S. (2005). *Cognitive therapy for challenging problems: What to do when the basics don't work*. New York: Guilford.

Beck, J. S. (2011). *Cognitive therapy: Basics and Beyond* (2nd ed.). New York: Guilford.

Bell, V., Grech, E., Maiden, C., Halligan, P. W., & Ellis, H. D. (2005). 'Internet delusions': A case series and theoretical integration. *Psychopathology, 38*, 144-150. doi: PSP2005038003144 [pii]

10.1159/000085845

Bennett-Levy, J. (2003). Mechanisms of change in cognitive therapy: The case of automatic thought records and behavioural experiments. *Behavioural and Cognitive Psychotherapy, 31*, 261-277.

Bennett-Levy, J., Butler, G., Fennell, M., Hackmann, A., Mueller, M., & Westbrook, D. (Eds.). (2004). *Oxford guide to behavioural experiments in cognitive therapy*. New York: Oxford University Press.

Beyebach, M., & Carranza, V. E. (1997). Therapeutic interaction and dropout: Measuring relational communication in solution-focused therapy. *Journal of Family Therapy, 19*, 173-221.

Bieling, P. J., & Kuyken, W. (2003). Is cognitive case formulation science or science fiction? *Clinical Psychology: Science and Practice, 10*, 52-69.

Bishop, W., & Fish, J. M. (1999). Questions as interventions: Perceptions of Socratic, solution-focused, and diagnostic questioning styles. *Journal of Rational-Emotive and Cognitive Behavior Therapy, 17*, 115-140.

Blackburn, I.-M., James, I. A., Milne, D. L., Baker, C., Standart, S., Garland, A., & Reichelt, F. (2001). The revised Cognitive Therapy Scale (CTS-R): Psychometric properties. *Behavioural and Cognitive Psychotherapy, 29*, 431-446.

Bordin, E. (1979). The generalizability of the psychoanalytic concept of the working alliance. *Psychotherapy: Theory, Research and Practice, 16*, 252-260.

Bordin, E. (1980). *Of human bonds that bind or free*. Paper presented at the Annual meeting of the Society for Psychotherapy Research, Pacific Grove, CA.

Bordin, E. (1994). Theory and research on the therapeutic working alliance: New directions. In A. O. Horvath & L. S. Greenberg (Eds.), *The working alliance: Theory, research, and practice* (pp. 13-37). Oxford, UK: Wiley.

Boswell, J. F., Gallagher, M. W., Sauer-Zavala, S. E., Bullis, J., Gorman, J. M., Shear, M., . . . Barlow, D. H. (in press). Patient characteristics and variability in adherence and competence in cognitive-behavioral therapy for panic disorder. *Journal of Consulting and Clinical Psychology*. doi: http://dx.doi.org/10.1037/a0031437

Boulkedid, R., Abdoul, H., Loustau, M., Sibony, O., & Alberti, C. (2011). Using and reporting the Delphi Method for selecting healthcare quality indicators: A systematic review. *PLoS ONE, 6*, e20476. doi: 10.1371/journal.pone.0020476

Bradley, M. M., Greenwald, M. K., Petry, M. C., & Lang, P. J. (1992). Remembering pictures: Pleasure and arousal in memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18*, 379-390.

Brotman, M. A. (2004). Therapeutic alliance and adherence in cognitive therapy for depression. *Dissertation Abstracts International: Section B: The Sciences and Engineering, 65*(6-B), 3146.

Burlingame, G. M., McClendon, D. T., & Alonso, J. (2011). Cohesion in group therapy. *Psychotherapy, 48*, 34-42. doi: 10.1037/a0022063

Butler, G., Fennell, M., & Hackmann, A. (2008). *Cognitive-Behavioral Therapy for Anxiety Disorders: Mastering Clinical Challenges*. New York: Guilford.

Castonguay, L. G., Constantino, M. J., & Holtforth, M. G. (2006). The working alliance: Where are we and where should we go? *Psychotherapy: Theory, Research, Practice, Training, 43*, 271-279.

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. doi: http://dx.doi.org/10.1037/0022-006X.64.3.497

Center for Clinical Research. (1996). *Cognitive and behavioral treatment of depression. A research treatment manual*. University of Washington. Seattle, WA.

Chadwick, P., Birchwood, M. J., & Trower, P. (1996). *Cognitive therapy for delusions, voices and paranoia*. Chichester, New York: Wiley.

Charlesworth, G., & Greenfield, S. (2004). Overcoming barriers to collaborative conceptualization in cognitive therapy with older adults. *Behavioural and Cognitive Psychotherapy, 32*, 411-422. doi: doi:10.1017/S1352465804001699

Christ, T. J., & Boice, C. (2009). Rating scale items: A brief review of nomenclature, components, and formatting to inform the development of Direct Behavior Rating (DBR). *Assessment for Effective Intervention, 34*(4), 242-250. doi: http://dx.doi.org/10.1177/1534508409336182

Christianson, S.-A. (1992). Emotional stress and eyewitness memory: A critical review. *Psychological Bulletin, 112*, 284-309.

Clark, D. A. (in press). Collaborative empiricism: A cognitive response to exposure reluctance and low distress tolerance. *Cognitive and Behavioral Practice*. doi: http://dx.doi.org/10.1016/j.cbpra.2012.06.001

Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment, 7*, 309-319.

Clemence, A., Hilsenroth, M., Ackerman, S., Strassle, C., & Handler, L. (2005). Facets of the therapeutic alliance and perceived progress in psychotherapy: Relationship between patient and therapist perspectives. *Clinical Psychology and Psychotherapy, 12*, 443-454.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*, 37-46. doi: http://dx.doi.org/10.1177/001316446002000104

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin, 70*, 213-220. doi: http://dx.doi.org/10.1037/h0026256

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Erlbaum.

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

Cohen, L. L., La Greca, A. M., Blount, R. L., Kazak, A. E., Holmbeck, G. N., & Lemanek, K. L. (2008). Introduction to Special Issue: Evidence-based Assessment in Pediatric Psychology. *Journal of Pediatric Psychology, 33*(9), 911-915. doi: 10.1093/jpepsy/jsj115

Colson, D. D., Horwitz, L., Allen, J. G., Frieswyk, S. H., Gabbard, G. O., Newsom, G. E., & Coyne, L. (1988). Patient collaboration as a criterion for the therapeutic alliance. *Psychoanalytic Psychology, 5*, 259-268.

Constantino, M. J., Castonguay, L. G., & Schut, A. J. (2002). The working alliance: A flagship for the "Scientist-Practitioner" model in psychotherapy. In G. S. Tryon (Ed.), *Counseling based on process research*. Boston: Allyn and Bacon.

Cook, J. M., Biyanova, T., Elhai, J., Schnurr, P. P., & Coyne, J. C. (2010). What do psychotherapists really do in practice? An Internet study of over 2,000 practitioners. *Psychotherapy (Chicago, Ill.), 47*, 260-267. doi: 10.1037/a0019788

Couch, R., & Childers, J. H. (1991). Changes in the theory and practice of individual psychotherapy by the year 2000. *TACD Journal, 19*, 43-55.

Crits-Christoph, P., Connolly Gibbons, M. B., & Hearon, B. (2006). Does the alliance cause good outcome? Recommendations for future research on the alliance. *Psychotherapy: Theory, Research, Practice, Training, 43*, 280-285.

Crits-Christoph, P., Gibbons, M. B. C., & Hearon, B. (2006). Does the alliance cause good outcome? Recommendations for future research on the alliance. *Psychotherapy: Theory, Research, Practice, Training Fall, 43*, 280-285.

Cross, L. W. (2007). Islands of health: Implications of academic psychology for psychotherapy technique and integration. *Journal of Psychotherapy Integration, 17*, 251-273.

Cuijpers, P., Hollon, S. D., van Straten, A., Bockting, C., Berking, M., & Andersson, G. (2013). Does cognitive behaviour therapy have an enduring effect that is superior to keeping patients on continuation pharmacotherapy? A meta-analysis. *BMJ Open, 3*. Retrieved from doi:10.1136/bmjopen-2012-002542

Dalkey, N. C. (1972). The Delphi method: An experimental study of group opinion. In N. C. Dalkey, D. L. Rourke, R. Lewis & D. Snyder (Eds.), *Studies in the quality of life: Delphi and decision-making* (pp. 13-54). Lexington, MA: Lexington.

Dattilio, F. M. (2000). Cognitive-behavioral strategies. In J. Carlson & L. Sperry (Eds.), *Brief therapy with individuals and couples* (pp. 33-70). Phoenix, AZ: Zeig, Tucker and Theisen.

Dattilio, F. M. (2010). *Cognitive-behavioral therapy with couples and families: A comprehensive guide for clinicians*. New York: Guilford.

Dattilio, F. M., & Hanna, M. A. (2012). Collaboration in Cognitive-Behavioral Therapy. *Journal of Clinical Psychology, 68*, 146-158. doi: 10.1002/jclp.21831

Dattilio, F. M., & Padesky, C. A. (1990). *Cognitive therapy with couples*. Sarasota, FL: Professional Resource Exchange.

Davis, D., & Perkowitz, W. T. (1979). Consequences of responsiveness in dyadic interactions: Effects of probability of responses and proportion of content related responses. *Journal of Personality and Social Psychology, 37*, 534-550. doi: 10.1037/0022-3514.37.4.534

Davis, L. L. (1992). Instrument review: Getting the most from a panel of experts. *Applied Nursing Research, 5*, 194-197. doi: http://dx.doi.org/10.1016/S0897-1897(05)80008-4

Davison, G. C. (1998). Being bolder with the Boulder model: The challenge of education and training in empirically supported treatments. *Journal of Consulting and Clinical Psychology, 66*, 163-167. doi: http://dx.doi.org/10.1037/0022-006X.66.1.163

Dawis, R. V. (1987). Scale construction. *Journal of Counseling Psychology, 34*, 481-489.

Dawson, R. (2011). How Significant Is A Boxplot Outlier? *Journal of Statistics Education, 19*, 1-12.

Deci, E. L., & Ryan, R. M. (1980). The empirical exploration of intrinsic motivational processes. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 13, pp. 39-80). New York: Academic Press.

Deci, E. L., & Ryan, R. M. (1985). *Intrinsic motivation and self-determination in human behavior*. New York: Plenum.

Deci, E. L., & Ryan, R. M. (2000). The "what" and the "why" of goal pursuits: Human needs and the self-determination of behavior. *Psychological Inquiry, 11*, 227-268.

Deci, E. L., & Ryan, R. M. (2008). Self-determination theory: A macrotheory of human motivation, development, and health. *Canadian Psychology, 49*, 182-185.

Deegan, P. E., & Drake, R. E. (2006). Shared decision making and medication management in the recovery process. *Psychiatric Services, 57*, 1636-1639. doi: http://dx.doi.org/10.1176/appi.ps.57.11.1636

Del Re, A. C., Flückiger, C., Horvath, A. O., Symonds, D., & Wampold, B. E. (2012). Therapist effects in the therapeutic alliance–outcome relationship: A restricted-maximum likelihood meta-analysis. *Clinical Psychology Review, 32*, 642-649. doi: 10.1016/j.cpr.2012.07.002

Delbecq, A. L., Van de Ven, A. H., & Gustafson, D. H. (1975). *Group techniques for program planning*. Glenview, IL: Scott, Foresman, and Co.

Delgado-Rico, E., Carretero-Dios, H., & Ruch, W. (2012). Content validity evidences in test development: An applied perspective. *International Journal of Clinical and Health Psychology, 12*, 449-460.

Denton, W. H., Johnson, S. M., & Burleson, B. R. (2009). Emotion-Focused Therapy–Therapist Fidelity Scale (EFT-TFS): Conceptual Development and Content Validity. *Journal of Couple and Relationship Therapy: Innovations in Clinical and Educational Interventions, 8*, 226 - 246.

Derisley, J., & Reynolds, S. (2000). The transtheoretical stages of change as a predictor of premature termination, attendance and alliance in psychotherapy. *British Journal of Clinical Psychology, 39*(4), 371-382. doi: http://dx.doi.org/10.1348/014466500163374

Derlega, V. J., Hendrick, S. S., Winstead, B. A., & Berg, J. H. (1992). Psychotherapy as a personal relationship: A social psychological perspective. *Psychotherapy: Theory, Research, Practice, Training, 29*, 331-335. doi: http://dx.doi.org/10.1037/h0088534

Derogatis, L. R. (1974). The Hopkins Symptom Checklist (HSCL). *Behavioral Science, 19*, 1-15.

DeRubeis, R. J., Brotman, M. A., & Gibbons, C. (2005). A conceptual and methodological analysis of the nonspecifics argument. *Clinical Psychology: Science and Practice, 12*, 174-183.

DeRubeis, R. J., & Feeley, M. (1990). Determinants of change in cognitive therapy for depression. *Cognitive Therapy and Research, 14*, 469-482.

DeRubeis, R. J., Tang, T. Z., & Beck, A. T. (2001). Cognitive therapy. In K. S. Dobson (Ed.), *Handbook of cognitive behavioral therapies* (2nd ed., pp. 349-392). New York: Guilford.

DeVellis, R. F. (2003). *Scale development: Theory and applications*. Thousand Oaks, CA: Sage.

DeVon, H. A., Block, M. E., Moyle-Wright, P., Ernst, D. M., & et al. (2007). A Psychometric Toolbox for Testing Validity and Reliability. *Journal of Nursing Scholarship, 39*(2), 155-164.

Dillman, D. A. (2007). *Mail and internet surveys: The tailored design method* Hoboken, N.J: Wiley.

Dimidjian, S., Hollon, S. D., Dobson, K. S., Schmaling, K., Kohlenberg, R., Addis, M., . . . Jacobson, N. (2006). Randomized trial of behavioral activation, cognitive therapy, and antidepressant medication in the acute treatment of adults with major depression. *Journal of Consulting and Clinical Psychology, 74*, 658-670.

Dobson, D. J. G., & Dobson, K. S. (2009). *Evidence-based practice of cognitive-behavioral therapy*. New York: Guilford.

Dobson, D. J. G., & Dobson, K. S. (in press). In-session structure and collaborative empiricism. *Cognitive and Behavioral Practice*. doi: http://dx.doi.org/10.1016/j.cbpra.2012.11.002

Dobson, K. S. (2012). *Cognitive therapy*. Washington, DC: American Psychological Association.

Dobson, K. S., Hollon, S. D., Dimidjian, S., Schmaling, K. B., Kohlenberg, R. J., Gallop, R. J., . . . Jacobson, N. S. (2008). Randomized trial of behavioral activation, cognitive therapy, and antidepressant medication in the prevention of relapse and recurrence in major depression. *Journal of Consulting and Clinical Psychology, 76*, 468-477. doi: http://dx.doi.org/10.1037/0022-006X.76.3.468

Dobson, K. S., & Shaw, B. F. (1988). The use of treatment manuals in cognitive therapy: Experience and issues. *Journal of Consulting and Clinical Psychology, 56*, 673-680.

Dobson, K. S., Shaw, B. F., & Vallis, T. (1985). Reliability of a measure of the quality of cognitive therapy. *British Journal of Clinical Psychology, 24*, 295-300.

Dormaar, J., Dijkman, C. I., & de Vries, M. W. (1989). Consensus in patient-therapist interactions: A measure of the therapeutic relationship related to outcome. *Psychotherapy and Psychosomatics, 51*, 69-76.

Dryden, W. (1989). The therapeutic alliance as an integrating framework. In W. Dryden (Ed.), *Key issues for counselling in action* (pp. 1-15). Thousand Oaks, CA: Sage.

Dudley, R., & Kuyken, W. (2006). Formulation in cognitive-behavioural therapy: 'There is nothing either good or bad, but thinking makes it so'. In L. Johnstone & R. Dallos (Eds.), *Formulation in psychology and psychotherapy: Making sense of people's problems* (pp. 17-46). New York: Routledge.

Duehn, W. D., & Proctor, E. K. (1977). Initial clinical interaction and premature discontinuance in treatment. *American Journal of Orthopsychiatry, 47*(284-290).

Eisenthal, S., Koopman, C., & Lazare, A. (1983). Process analysis of two dimensions of the negotiated approach in relation to satisfaction in the initial interview. *Journal of Nervous and Mental Disease, 171*, 49-54.

Elliott, R., Bohart, A. C., Watson, J. C., & Greenberg, L. S. (2011). Empathy. *Psychotherapy, 48*, 43-49. doi: 10.1037/a0022187

Ellis, A. (1980). Rational-emotive therapy and cognitive behavior therapy: Similarities and differences. *Cognitive Therapy and Research, 4*, 325-340.

Elvins, R., & Green, J. (2008). The conceptualization and measurement of therapeutic alliance: An empirical review. *Clinical Psychology Review, 28*, 1167-1187.

En-Hong, D., & Yong, B. (2012). Reliability and validity of the Chinese Version of Wake Forest Physician Trust Scale. *Chinese Mental Health Journal, 26*, 171-175.

Engelkamp, J. (1988). *Memory for actions*. Hove: Psychology Press.

Evans, M. D., Hollon, S. D., DeRubeis, R. J., Piasecki, J. M., Grove, W. M., Garvey, M. J., & Tuason, V. B. (1992). Differential relapse following cognitive therapy and pharmacotherapy for depression. *Archives of General Psychiatry, 49*, 802-808.

Fagot, R. F. (1991). Reliability of ratings for multiple judges: Intraclass correlation and metric scales. *Applied Psychological Measurement, 15*, 1-11. doi: http://dx.doi.org/10.1177/014662169101500101

Farber, B. A., & Doolin, E. M. (2011). Positive regard. *Psychotherapy, 48*, 58-64. doi: 10.1037/a0022141

Feeley, M., DeRubeis, R., & Gelfand, L. (1999). The temporal relation of adherence and alliance to symptom change in cognitive therapy for depression. *Journal of Consulting and Clinical Psychology, 67*, 578-582.

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

Finger, M. S., & Rand, K. L. (2008). Addressing validity concerns in clinical psychology research *Handbook of Research Methods in Clinical Psychology* (pp. 13-30). Oxford, UK: Blackwell.

Finn, R. H. (1970). A note on estimating the reliability of categorical data. *Educational and Psychological Measurement, 30*, 70-76. doi: http://dx.doi.org/10.1177/001316447003000106

Fish, L. S., & Busby, D. M. (1996). The Delphi method. In D. H. Sprenkle & S. M. Moon (Eds.), *Research methods in family therapy* (pp. 469-482). New York: Guilford.

Fitzpatrick, M. R., Iwakabe, S., & Stalikas, A. (2005). Perspective divergence in the working alliance. *Psychotherapy Research, 15*, 69-79. doi: http://dx.doi.org/10.1080/10503300512331327056

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin, 5*, 378-382.

Fleiss, J. L. (1986). *The design and analysis of clinical experiments*. New York: Wiley.

Freud, S. (1912). The dynamics of transference (J. Strachey, Trans.). In J. Strachey (Ed.), *The standard edition of the complete psychological works of Sigmund Freud* (Vol. 12, pp. 99-108). London: Hogarth Press.

Friedberg, R. D., & Brelsford, G. M. (2011). Core principles in cognitive therapy with youth. *Child and Adolescent Psychiatric Clinics of North America, 20*, 369-378. doi: http://dx.doi.org/10.1016/j.chc.2011.01.009

Friedberg, R. D., & Clark, C. C. (2006). Supervision of cognitive therapy with youth. In T. K. Neill (Ed.), *Helping others help children: Clinical supervision of child psychotherapy* (pp. 109-122). Washington, DC: American Psychological Association.

Friedberg, R. D., & Gorman, A. A. (2007). Integrating psychotherapeutic processes with cognitive behavioral procedures. *Journal of Contemporary Psychotherapy, 37*, 185-193.

Friedberg, R. D., & McClure, J. M. (2002). *Clinical practice of cognitive therapy with children and adolescents*. New York: Guilford.

Friedman, E. S., & Thase, M. E. (2006). Cognitive-Behavioral Therapy for depression and dysthymia. In D. J. Stein, D. J. Kupfer & A. F. Schatzberg (Eds.), *The American Psychiatric Publishing textbook of mood disorders* (pp. 353-371). Arlington, VA: American Psychiatric Publishing.

Frieswyk, S. H., Allen, J. G., Colson, D. B., Coyne, L., Gabbard, G. O., Horwitz, L., & Newsom, G. (1986). Therapeutic alliance: Its place as a process and outcome variable in dynamic psychotherapy research. *Journal of Consulting and Clinical Psychology, 54*, 32-38.

Gajewski, B. J., Coffland, V., Boyle, D. K., Bott, M., Price, L. R., Leopold, J., & Dunton, N. (2012). Assessing content validity through correlation and relevance tools: A Bayesian randomized equivalence experiment. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 8*(3), 81-96. doi: http://dx.doi.org/10.1027/1614-2241/a000040

Garratt, G., Ingram, R. E., Rand, K. L., & Sawalani, G. (2007). Cognitive Processes in Cognitive Therapy: Evaluation of the Mechanisms of Change in the Treatment of Depression. *Clinical Psychology: Science and Practice, 14*, 224-239. doi: 10.1111/j.1468-2850.2007.00081.x

Gaston, L., Thompson, L., Gallagher, D., Cournoyer, L.-G., & Gagnon, R. (1998). Alliance, technique, and their interactions in predicting outcome of behavioral, cognitive, and brief dynamic therapy. *Psychotherapy Research, 8*, 190-209.

Gelso, C. J., & Carter, J. A. (1994). Components of the psychotherapy relationship: Their interaction and unfolding during treatment. *Journal of Counseling Psychology, 41*, 296-306. doi: http://dx.doi.org/10.1037/0022-0167.41.3.296

Gelso, C. J., & Hayes, J. A. (1998). *The psychotherapy relationship: Theory, research and practice*. New York: Wiley.

Giordano, F. L., & Briones, D. F. (2003). Assessing residents' competence in psychotherapy. *Academic Psychiatry, 27*, 145-147. doi: http://dx.doi.org/10.1176/appi.ap.27.3.145

Gleeson, J. F. M., & McGorry, P. D. (2004). *Psychological interventions in early psychosis*. Chichester, UK: Wiley.

Goldstein, J. M., Cohen, P., Lewis, S. A., & Struening, E. L. (1988). Community treatment environments: Patient vs. staff evaluations. *Journal of Nervous and Mental Disease, 176*, 227-233.

Grant, J. S., & Davis, L. L. (1997). Selection and use of content experts for instrument development. *Research in Nursing & Health, 20*, 269-274.

Greenson, R. R. (1965). The working alliance and the transference neurosis. *The Psychoanalytic Quarterly, 34*, 155-181.

Haaga, D. A. (2007). Could we speed this up? Accelerating progress in research on mechanisms of change in cognitive therapy of depression. *Clinical Psychology: Science and Practice, 14*, 240-243.

Haaga, D. A., Dyck, M. J., & Ernst, D. (1991). Empirical status of cognitive theory of depression. *Psychological Bulletin, 110*, 215-236. doi: http://dx.doi.org/10.1037/0033-2909.110.2.215

Hamann, J., Cohen, R., Leucht, S., Busch, R., & Kissling, W. (2005). Do patients with schizophrenia wish to be involved in decisions about their medical treatment? *The American Journal of Psychiatry, 162*(12), 2382-2384. doi: http://dx.doi.org/10.1176/appi.ajp.162.12.2382

Hamann, S. B. (2001). Cognitive and neural mechanisms of emotional memory. *Trends in Cognitive Sciences, 5*, 394-400.

Hamilton, M. (1960). A rating scale for depression. *Journal of Neurology, Neurosurgery & Psychiatry, 23*. doi: http://dx.doi.org/10.1136/jnnp.23.1.56

Hamilton, M. (1967). Development of a Rating Scale for Primary Depressive Illness. *British Journal of Social & Clinical Psychology, 6*, 278-296.

Hartmann, D. P. (1977). Considerations in the choice of interobserver reliability measures. *Journal of Applied Behavioral Analysis, 10*, 103-116. doi: http://dx.doi.org/10.1901%2Fjaba.1977.10-103

Hatcher, R. (2010). Alliance Theory and Measurement. In J. Muran & J. P. Barber (Eds.), *The Therapeutic Alliance: An Evidence-Based Guide to Practice* (pp. 7-28). New York: Guilford.

Hatcher, R., 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., & Barends, A. W. (2006). How a return to theory could help alliance research. *Psychotherapy: Theory, Research, Practice, Training, 43*, 292-299.

Hatcher, R., & Gillaspy, J. (2006). Development and validation of a revised short version of the Working Alliance Inventory. *Psychotherapy Research, 16*, 12-25.

Haubert, L. C., & Dobson, K. S. (2007). Treatment of depression and mechanisms of change: Strengthening the links among theory, research, and practice. *Clinical Psychology: Science and Practice, 14*, 247-251. doi: http://dx.doi.org/10.1111/j.1468-2850.2007.00084.x

Hayes, J. R., & Hatch, J. A. (1999). Issues in measuring reliability: Correlation versus percentage of agreement. *Written Communication, 16*, 354-367.

Haynes, S. N., Richard, D. C., & Kubany, E. S. (1995). Content validity in psychological assessment: A functional approach to concepts and methods. *Psychological Assessment, 7*, 238-247. doi: http://dx.doi.org/10.1037/1040-3590.7.3.238

Hazlett-Stevens, H. (2008). *Psychological approaches to generalized anxiety disorder*. New York: Springer.

Helstrup, T. (2004). The enactment effect is due to more than guesses and beliefs. *Scandinavian Journal of Psychology, 45*, 259-263.

Hill, C., & Lambert, M. J. (2004). Methodological issues in studying psychotherapy processes and outcomes. In M. J. Lambert (Ed.), *Bergin and Garfield's Handbook of Psychotherapy and Behavior Change* (5th ed., pp. 84-135). New York: Wiley.

Hinkin, T. R. (1995). A review of scale development practices in the study of organizations. *Journal of Management, 21*(5), 967-988. doi: http://dx.doi.org/10.1177/014920639502100509

Hollanders, H., & McLeod, J. (1999). Theoretical orientation and reported practice: A survey of eclecticism among counsellors in Britain. *British Journal of Guidance and Counselling, 27*, 405-414.

Hollon, S. D. (1999). Rapid early response in cognitive behavior therapy: A commentary. *Clinical Psychology: Science and Practice, 6*, 305-309.

Hollon, S. D., & Beck, A. T. (1979). Cognitive therapy of depression. In P. C. Kendall & S. D. Hollon (Eds.), *Cognitive-behavioral interventions: Theory, research, and procedures* (pp. 153-203). New York: Academic Press.

Hollon, S. D., DeRubeis, R. J., Shelton, R. C., Amsterdam, J. D., Salomon, R. M., O'Reardon, J. P., . . . Gallop, R. (2005). Prevention of relapse following cognitive therapy vs medications in moderate to severe depression. *Archives of General Psychiatry, 62*, 417-422. doi: 10.1001/archpsyc.62.4.417

Hollon, S. D., Evans, M. D., Auerbach, A., DeRubeis, R., Elkin, I., Lowery, A., . . . Piasecki, J. (1988). *Development of a system for rating therapies of depression: Differentiating cognitive therapy, interpersonal psychotherapy, and clinical management pharmacotherapy*. Unpublished Work.

Hollon, S. D., Evans, M. D., & DeRubeis, R. J. (1990). Cognitive mediation of relapse prevention following treatment for depression: Implications of differential risk. In R.

E. Ingram (Ed.), *Contemporary psychological approaches to depression: Theory, research, and treatment* (pp. 117-136). New York: Plenum.

Hollon, S. D., Evans, M. D., Elkin, I., & Lowery, A. (1984). *System for rating therapies for depression*. Paper presented at the 92nd Annual Convention of the American Psychological Association, Toronto, Ontario, Canada.

Hollon, S. D., & Kriss, M. R. (1984). Cognitive factors in clinical research and practice. *Clinical Psychology Review, 4*, 35-76.

Horvath, A. O. (1994). Research on the alliance. In A. O. Horvath & L. S. Greenberg (Eds.), *The working alliance: Theory, research, and practice* (pp. 259-286). New York: Wiley.

Horvath, A. O. (2000). The therapeutic relationship: From transference to alliance. *Journal of Clinical Psychology, 56*, 163-173.

Horvath, A. O. (2001). The alliance. *Psychotherapy: Theory, Research, Practice, Training, 38*, 365-372.

Horvath, A. O. (2005). The therapeutic relationship: Research and theory. *Psychotherapy Research, 15*, 3-7. doi: http://dx.doi.org/10.1080/10503300512331339143

Horvath, A. O. (2006). The alliance in context: Accomplishments, challenges, and future directions. *Psychotherapy: Theory, Research, Practice, Training, 43*, 258-263.

Horvath, A. O. (2011). Alliance in common factor land: A view through the research lens. *Research in Psychotherapy: Psychopathology, Process and Outcome, 14*, 121-135.

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

Horvath, A. O., Del Re, A., Fluckiger, C., & Symonds, D. (2011). Alliance in individual psychotherapy. *Psychotherapy, 48*, 9-16. doi: http://dx.doi.org/10.1037/a0022186

Horvath, A. O., & Greenberg, L. S. (1989). Development and validation of the Working Alliance Inventory. *Journal of Counseling Psychology, 36*, 223-233.

Horvath, A. O., & Greenberg, L. S. (1994). *The working alliance: Theory, research, and practice*. New York: Wiley.

Horvath, A. O., & Luborsky, L. (1993). The role of the therapeutic alliance in psychotherapy. *Journal of Consulting and Clinical Psychology, 61*, 561-573.

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

Hougaard, E. (1994). The therapeutic alliance: A conceptual analysis. *Journal of Psychology, 35*, 67-85.

Hoyt, M. F., Xenakis, S. N., Marmar, C. R., & Horowitz, M. J. (1983). Therapists' actions that influence their perceptions of "good" psychotherapy sessions. *Journal of Nervous and Mental Disease, 171*, 400-404.

Hsu, C.-C., & Sandford, B. A. (2007). The Delphi Technique: Making sense of consensus. *Practical Assessment, Research and Evaluation, 12*. Retrieved from http://pareonline.net/

Hutton, P., & Morrison, A. P. (in press). Collaborative empiricism in cognitive therapy for psychosis: A practice guide. *Cognitive and Behavioral Practice*. doi: http://dx.doi.org/10.1016/j.cbpra.2012.08.003

Imel, Z. E., & Wampold, B. E. (2008). The importance of treatment and the science of common factors in psychotherapy *Handbook of counseling psychology* (4th ed., pp. 249-266). Hoboken, NJ: Wiley.

Ingram, R. E., & Hollon, S. D. (1986). Cognitive therapy for depression from an information processing perspective. In R. E. Ingram (Ed.), *Information processing approaches to clinical psychology*. San Diego, CA: Academic Press.

Ingram, R. E., Miranda, J., & Segal, Z. V. (1998). *Cognitive vulnerability to depression*. New York: Guilford.

Jacobson, N., Dobson, K. S., Truax, P., Addis, M., Koerner, K., Gollan, J., . . . Prince, S. (1996). A component analysis of cognitive-behavioral treatment for depression. *Journal of Consulting and Clinical Psychology, 64*, 295-304.

Jinks, M., McMurran, M., & Huband, N. (2012). Engaging clients with personality disorder in treatment. *Mental Health Review Journal, 17*, 139-144. doi: http://dx.doi.org/10.1108/13619321211287229

Jones, A. P., Johnson, L. A., Butler, M. C., & Main, D. S. (1983). Apples and oranges: An empirical comparison of commonly used indices of interrater agreement. *Academy of Management Journal, 26*, 507-519. doi: http://dx.doi.org/10.2307/256262

Jones, J., & Hunter, D. (1995). Consensus methods for medical and health services research. *British Medical Journal, 311*, 376-380. doi: 10.2307/29728325

Kahneman, D. (2003). A perspective on judgment and choice: Mapping bounded rationality. *American Psychologist, 58*, 697-720.

Kanzer, M. (1981). Freud's "analytic pact": The standard therapeutic alliance. *Journal of the American Psychoanalytic Association, 29*, 69-87. doi: http://dx.doi.org/10.1177/000306518102900104

Kaplan, R. M., & Saccuzzo, D. P. (2013). *Psychological testing: Principles, applications, and issues*. Belmont, CA: Cengage Learning.

Karver, M., Handelsman, J., Fields, S., & Bickman, L. (2005). A theoretical model of common process factors in youth and family therapy. *Mental Health Services Research, 7*, 35-51.

Karver, M., Handelsman, J. B., Fields, S., & Bickman, L. (2006). Meta-analysis of therapeutic relationship variables in youth and family therapy: The evidence for

different relationship variables in the child and adolescent treatment outcome literature. *Clinical Psychology Review, 26*, 50-65.

Kaufman, K. L., Holden, E., & Walker, C. (1989). Future directions in pediatric and clinical child psychology. *Professional Psychology: Research and Practice, 20*, 148-152. doi: http://dx.doi.org/10.1037/0735-7028.20.3.148

Kazantzis, N., Beck, J. S., Dattilio, F. M., Dobson, K. S., & Rapee, R. (in press). Collaborative empiricism as the central therapeutic relationship element in cognitive behavior therapy: An expert panel discussion at the 7th International Congress of Cognitive Psychotherapy. *International Journal of Cognitive Psychotherapy*.

Kazantzis, N., Busch, R., Ronan, K. R., & Merrick, P. L. (2007). Using homework assignments in psychotherapy: Differences by theoretical orientation and professional training. *Behavioural and Cognitive Psychotherapy, 35*, 121-128. doi: http://dx.doi.org/10.1017/S1352465806003328

Kazantzis, N., Cronin, T. J., Dattilio, F. M., & Dobson, K. S. (in press). Introduction: Using techniques via the therapeutic relationship. *Cognitive and Behavioral Practice*. doi: http://dx.doi.org/10.1016/j.cbpra.2013.03.001

Kazantzis, N., Deane, F., & Ronan, K. (2000). Homework assignments in cognitive and behavioral therapy: A meta-analysis. *Clinical Psychology: Science and Practice, 7*, 189-202.

Kazantzis, N., Deane, F. P., Ronan, K. R., & L'Abate, L. (2005). *Using homework assignments in cognitive behavior therapy*. New York: Routledge/Taylor & Francis.

Kazantzis, N., Deane, F. R., & Ronan, K. R. (2004). Assessing compliance with homework assignments: Review and recommendations for clinical practice. *Journal of Clinical Psychology, 60*, 627-641.

Kazantzis, N., Freeman, A., Fruzzetti, A. E., Persons, J. B., & Smucker, M. (2013). Unresolved issues regarding collaborative empiricism in cognitive and behavioural therapies: An expert panel discussion at AACBT. *Behaviour Change, 30*, 1-11. doi: http://dx.doi.org/10.1017/bec.2013.1

Kazantzis, N., & Kellis, E. (2012). A Special Feature on Collaboration in Psychotherapy. *Journal of Clinical Psychology, 68*, 133-135. doi: 10.1002/jclp.21837

Kazantzis, N., Osborne, C., & Cronin, T. J. (2012). *The Working Alliance Inventory – Short Revised – Observer version*. Unpublished Instrument.

Kazantzis, N., Tee, J. M., Dattilio, F. M., & Dobson, K. S. (in press). How to develop collaborative empiricism in cognitive behavior therapy: Conclusions from the Cognitive and Behavioral Practice special series. *Cognitive and Behavioral Practice*. doi: http://dx.doi.org/10.1016/j.cbpra.2013.03.005

Keeney, S., Hasson, F., & McKenna, H. P. (2001). A critical review of the Delphi technique as a research methodology for nursing. *International Journal of Nursing Studies, 38*, 195-200.

Kennedy, P. (2008). *A guide to econometrics*. Malden, MA: Blackwell.

Kerlinger, F. N. (1973). *Foundations of behavioral research*. New York: Holt, Rinehart, and Winston.

Kim, S., Boren, D., & Solem, S. (2001). The Kim Alliance Scale: Development and preliminary testing. *Clinical Nursing Research, 10*, 314-331.

Kline, P. (1999). *The Handbook of Psychological Testing* (2nd ed.). London, UK: Routledge.

Klosko, J. S., & Sanderson, W. C. (1999). *Depression: Clinical application of empirically supported psychotherapy*. Northvale, NJ: Jason Aronson.

Kolb, D. (1984). *Experiential learning: Experience as the source of learning and development*. Englewood Cliffs, NJ: Prentice Hall.

Kolbe, R. H., & Burnett, M. S. (1991). Content-analysis research: An examination of applications with directives for improving research reliability and objectivity. *Journal of Consumer Research, 18*, 243-250. doi: http://dx.doi.org/10.1086/209256

Krupnick, J., Sotsky, S., Simmens, S., Moyer, J., Elkin, I., Watkins, J., & Pilkonis, P. (1996). The role of the therapeutic alliance in psychotherapy and pharmacotherapy outcome: Findings in the National Institute of Mental Health Treatment of Depression Collaborative Research Program. *Journal of Consulting and Clinical Psychology, 64*, 532-539.

Kuyken, W., Padesky, C. A., & Dudley, R. (2008). The science and practice of case conceptualization. *Behavioural and Cognitive Psychotherapy, 36*, 757-768.

Kuyken, W., Padesky, C. A., & Dudley, R. (2009). *Collaborative case conceptualisation: Working effectively with clients in cognitive-behavioral therapy*. New York: Guilford.

Lahey, M. A., Downey, R. G., & Saal, F. E. (1983). Intraclass correlations: There's more there than meets the eye. *Psychological Bulletin, 93*, 586-595.

Lambert, M. J. (Ed.). (2004). *Bergin and Garfield's Handbook of psychotherapy and behavior change* (5th ed.). New York: Wiley.

Lambert, M. J., & Barley, D. E. (2001). Research summary on the therapeutic relationship and psychotherapy outcome. *Psychotherapy: Theory, Research, Practice, Training, 38*, 357-361. doi: http://dx.doi.org/10.1037/0033-3204.38.4.357

Lambert, M. J., & Cattani, K. (2012). Practice-Friendly Research Review: Collaboration in Routine Care. *Journal of Clinical Psychology, 68*, 209-220. doi: 10.1002/jclp.21835

Lambert, M. J., & Ogles, B. M. (2004). The efficacy and effectiveness of psychotherapy. In M. J. Lambert (Ed.), *Bergin and Garfield's Handbook of psychotherapy and behavior change* (pp. 139-193). New York: Wiley.

Lambert, M. J., & Shimokawa, K. (2011). Collecting client feedback. *Psychotherapy, 48*, 72-79. doi: http://dx.doi.org/10.1037/a0022238

Lawshe, C. (1975). A quantitative approach to content validity. *Personnel Psychology, 28*, 563-575. doi: http://dx.doi.org/10.1111/j.1744-6570.1975.tb01393.x

Leahy, R. L. (2003). *Cognitive therapy techniques: A practitioner's guide*. New York: Guilford.

Leahy, R. L. (2008). The therapeutic relationship in cognitive-behavioral therapy. *Behavioural and Cognitive Psychotherapy, 36*, 769-777. doi: http://dx.doi.org/10.1017/S1352465808004852

LeBreton, J. M., Burgess, J. R., Kaiser, R. B., Atchley, E., & James, L. R. (2003). The restriction of variance hypothesis and interrater reliability and agreement: Are ratings from multiple sources really dissimilar? *Organizational Research Methods, 6*, 80-128. doi: http://dx.doi.org/10.1177/1094428102239427

Lejuez, C., Hopko, D., Levine, S., Gholkar, R., & Collins, L. (2005). The therapeutic alliance in behavior therapy. *Psychotherapy: Theory, Research, Practice, Training, 42*, 456-468.

Lepper, G., & Mergenthaler, E. (2007). Therapeutic collaboration: How does it work? *Psychotherapy Research, 17*, 576-587. doi: http://dx.doi.org/10.1080/10503300601140002

Levin, L., Henderson, H. A., & Ehrenreich-May, J. (2012). Interpersonal predictors of early therapeutic alliance in a transdiagnostic cognitive-behavioral treatment for adolescents with anxiety and depression. *Psychotherapy, 49*, 218-230. doi: 10.1037/a0028265

Lewin, K. (1946). Action research and minority problems. *Journal of Social Issues, 2*, 34-46.

Linstone, H. A., & Turoff, M. (1975). *The Delphi method: Techniques and applications*. Reading, MA: Addison-Wesley.

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. doi: http://dx.doi.org/10.1037/a0022426

Lombard, M., Snyder-Duch, J., & Bracken, C. C. (2002). Content analysis in mass communication: Assessment and reporting of intercoder reliability. *Human Communication Research, 28*, 587-604. doi: http://dx.doi.org/10.1111/j.1468-2958.2002.tb00826.x

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

Luborsky, L., & Auerbach, A. H. (1985). The therapeutic relationship in psychodynamic psychotherapy: The research evidence and its meaning for practice. In R. E. Hales & A. J. Frances (Eds.), *Annual review* (Vol. 4, pp. 550-561). Washington, DC: American Psychiatric Press.

Ludwig, B. (1997). Predicting the future: Have you considered using the Delphi methodology? *Journal of Extension, 35*, 1-4. Retrieved from http://www.joe.org/joe/1997october/tt2.php

Lynn, M. R. (1986). Determination and quantification of content validity. *Nursing Research, 35*, 382-385. doi: http://dx.doi.org/10.1097/00006199-198611000-00017

Madsen, W. M. (2009). Collaborative helping: A practice framework for family-centered services. *Family Process, 48*, 103-116.

Markland, D., Ryan, R. M., Tobin, V. J., & Rollnick, S. (2005). Motivational interviewing and self-determination theory. *Journal of Social and Clinical Psychology, 24*, 811-831.

Martin, D. J., Garske, J. P., & Davis, 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: http://dx.doi.org/10.1037/0022-006X.68.3.438

Mastaglia, B., Toye, C., & Kristjanson, L. J. (2003). Ensuring content validity in instrument development: Challenges and innovative approaches. *Contemporary Nurse, 14*, 281-291.

Masumoto, K., Yamaguchi, M., Sutani, K., Tsuneto, S., Fujita, A., & Tonoike, M. (2006). Reactivation of physical motor information in the memory of action events. *Brain Research, 1101*, 102-109.

McCabe, R., & Priebe, S. (2004). The therapeutic relationship in the treatment of severe mental illness: A review of methods and findings. *International Journal of Social Psychiatry, 50*, 115-128.

McCormack, B., McCarthy, G., Wright, J., & Coffey, A. (2009). Development and Testing of the Context Assessment Index (CAI). *Worldviews on Evidence-Based Nursing, 6*, 27-35.

McGraw, K. O., & Wong, S. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods, 1*, 30-46. doi: http://dx.doi.org/10.1037/1082-989X.1.1.30

Meissner, W. W. (2006). The therapeutic alliance: A proteus in disguise. *Psychotherapy: Theory, Research, Practice, Training, 43*, 264-270.

Merali, N., & Lynch, P. (1997). Collaboration in cognitive-behavioural counselling: A case example. *Canadian Journal of Counselling, 31*, 287-293.

Messer, S. B. (1991). The case formulaton approach: Issues of reliability and validity. *American Psychologist, 46*, 1348-1350.

Miller, L. C., & Berg, J. H. (1984). Selectivity and urgency in interpersonal exchange. In V. J. Derlega (Ed.), *Communication, intimacy, and close relationships* (pp. 161-205). Orlando, FL: Academic Press.

Miller, W. R., & Rollnick, S. (2002). *Motivational interviewing: Preparing people to change*. New York: Guilford.

Mintz, J., & Luborsky, L. (1971). Segments versus whole sessions: Which is the better unit for psychotherapy process research? *Journal of Abnormal Psychology, 78*, 180-191. doi: http://dx.doi.org/10.1037/h0031969

Morrison, A. P., & Barratt, S. (2010). What are the components of CBT for psychosis? A Delphi study. *Schizophrenia Bulletin, 36*, 136-142. doi: http://dx.doi.org/10.1093/schbul/sbp118

Morrison, K. H., Bradley, R., & Westen, D. (2003). The external validity of controlled clinical trials of psychotherapy for depression and anxiety: A naturalistic study. *Psychology and Psychotherapy: Theory, Research and Practice, 76*, 109-132. doi: http://dx.doi.org/10.1348/147608303765951168

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

Murphy, M. K., Black, N., Lamping, D. L., McKee, C. M., Sanderson, C. F. B., Askham, J., & Marteau, T. (1998). Consensus development methods, and their use in clinical guideline development. *Health Technology Assessment, 2*, 1-88.

Muse, K., & McManus, F. (2013). A systematic review of methods for assessing competence in cognitive-behavioural therapy. *Clinical Psychology Review, 33*, 484-499.

Needleman, L. D. (1999). *Cognitive case conceptualisation: A guidebook for practitioners*. Mahwah, NJ: Erlbaum.

Nelson, H. (1997). *Cognitive behavioural therapy with schizophrenia: A practice manual*. Cheltenham, UK: Stanley Thornes Ltd.

Norcross, J. C. (2001). Purposes, processes and products of the Task Force on Empirically Supported Therapy Relationships. *Psychotherapy: Theory, Research, Practice, Training, 38*, 345-356. doi: http://dx.doi.org/10.1037/0033-3204.38.4.345

Norcross, J. C. (2002). *Psychotherapy relationships that work: Therapist contributions and responsiveness to patients*. New York: Oxford University Press.

Norcross, J. C. (Ed.). (2011). *Psychotherapy relationships that work: Evidence-based responsiveness* (2nd ed.). New York: Oxford University Press.

Norcross, J. C., Hedges, M., & Prochaska, J. O. (2002). The face of 2010: A Delphi poll on the future of psychotherapy. *Professional Psychology: Research and Practice, 33*, 316-322. doi: http://dx.doi.org/10.1037/0735-7028.33.3.316

Norcross, J. C., Koocher, G. P., Fala, N. C., & Wexler, H. K. (2010). What does not work? Expert consensus on discredited treatments in the addictions. *Journal of Addiction Medicine, 4*, 174-180. doi: http://dx.doi.org/10.1097/ADM.0b013e3181c5f9db

Norcross, J. C., Koocher, G. P., & Garofalo, A. (2006). Discredited psychological treatments and tests: A Delphi poll. *Professional Psychology: Research and Practice, 37*, 515-522. doi: http://dx.doi.org/10.1037/0735-7028.37.5.515

Norcross, J. C., & Lambert, M. J. (2011). Psychotherapy relationships that work II. *Psychotherapy, 48*(1), 4-8. doi: http://dx.doi.org/10.1037/a0022180

Norcross, J. C., & Wampold, B. E. (2011). Evidence-based therapy relationships: Research conclusions and clinical practices *Psychotherapy relationships that work: Evidence-based responsiveness* (2nd ed., pp. 423-430). New York: Oxford University Press.

Nunnally, J. C., & Bernstein, I. H. (1994). *Pyschometric Theory* (3rd ed.). New York: McGraw-Hill.

O'Brien, W. H. (2010). Evaluating case formulation decision-making and therapist responsiveness: A perspective from the area of behavioral assessment and case formulation. *Pragmatic Case Studies in Psychotherapy, 6*, 293-306.

O'Donohue, W. T., & Fisher, J. E. (2008). *Cognitive behavior therapy: Applying empirically supported techniques in your practice* (2nd ed.). New Jersey: Wiley.

O'Malley, S. S., Suh, C. S., & Strupp, H. H. (1983). The Vanderbilt Psychotherapy Process Scale: A report on the scale development and a process-outcome study. *Journal of Consulting and Clinical Psychology, 51*, 581-586.

Orlinsky, D. E., Grawe, K., & Parks, B. K. (1994). Process and outcome in psychotherapy: Noch einmal. In A. E. Bergin & S. L. Garfield (Eds.), *Handbook of psychotherapy and behavior change* (4th ed., pp. 270-376). New York: Wiley.

Orlinsky, D. E., Ronnestad, M. H., & Willutzki, U. (2004). Fifty years of psychotherapy process-outcome research: Continuity and change. In M. J. Lambert (Ed.), *Bergin and Garfield's Handbook of Psychotherapy and Behavior Change* (pp. 307-389). New York: Wiley.

Overholser, J. C. (1995). Elements of the Socratic method: IV. Disavowal of knowledge. *Psychotherapy: Theory, Research, Practice, Training, 32*, 283-292.

Overholser, J. C. (2011). Collaborative Empiricism, Guided Discovery, and the Socratic Method: Core Processes for Effective Cognitive Therapy. *Clinical Psychology: Science and Practice, 18*, 62-66. doi: 10.1111/j.1468-2850.2011.01235.x

Padesky, C. A., & Greenberger, D. (1995). *Clinician's guide to mind over mood*. New York: Guilford.

Perepletchikova, F., Treat, T. A., & Kazdin, A. E. (2007). Treatment integrity in psychotherapy research: Analysis of the studies and examination of the associated

factors. *Journal of Consulting and Clinical Psychology, 75*, 829-841. doi: http://dx.doi.org/10.1037/0022-006X.75.6.829

Persons, J. B. (2008). *The case formulation approach to cognitive-behavior therapy*. New York: Guilford.

Piner, K. E., Berg, J. H., & Miller, L. C. (1986). *The influence of timing on responsiveness*. Paper presented at the Southeastern Psychological Association Convention, Orlando, FL.

Polit, D. F., & Beck, C. T. (2006). The content validity index: Are you sure you know what's being reported? Critique and recommendations. *Research in Nursing & Health, 29*, 489-497. doi: 10.1002/nur.20147

Polit, D. F., & Hungler, B. (1991). *Nursing research: Principles and methods*. Philidelphia, PA: Davis.

Powell, C. (2003). The Delphi technique: Myths and realities. *Journal of Advanced Nursing, 41*, 376-382.

Prochaska, J. O., & Norcross, J. C. (1983). Contemporary psychotherapists: A national survey of characteristics, practices, orientations, and attitudes. *Psychotherapy: Theory, Research & Practice, 20*, 161-173. doi: http://dx.doi.org/10.1037/h0088487

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

Reeve, J. (1998). Autonomy support as an interpersonal motivating style: Is it teachable? *Contemporary Educational Psychology, 23*, 312-330.

Reeve, J. (2002). Self-determination theory applied to educational settings. In E. L. Deci & R. M. Ryan (Eds.), *Handbook of self-determination research* (pp. 193-204). Rochester, NY: University of Rochester Press.

Reilly, L. C., Ciesla, J. A., Felton, J. W., Weitlauf, A. S., & Anderson, N. L. (2012). Cognitive vulnerability to depression: A comparison of the weakest link, keystone and additive models. *Cognition and Emotion, 26*, 521-533. doi: http://dx.doi.org/10.1080/02699931.2011.595776

Ribeiro, E., Ribeiro, A. P., Gonçalves, M. M., Horvath, A. O., & Stiles, W. B. (2012). How collaboration in therapy becomes therapeutic: The therapeutic collaboration coding system. *Psychology and Psychotherapy: Theory, Research and Practice, Early View (Online Version of Record published before inclusion in an issue)*, no-no. doi: 10.1111/j.2044-8341.2012.02066.x

Rogers, C. R. (1957). The necessary and sufficient conditions of therapeutic personality change. *Journal of Consulting Psychology, 21*, 95-103.

Rogers, C. R. (1961). *On becoming a person: A therapist's view of psychotheraphy*. Boston: Houghton Mifflin.

Romero, A. A., Agnew, C. R., & Insko, C. A. (1996). The cognitive mediation hypothesis revisited: An empirical response to methodological and theoretical criticism. *Personality and Social Psychology Bulletin, 22*, 651-665. doi: http://dx.doi.org/10.1177/0146167296227001

Roth, A. D., & Pilling, S. (2007). *The competences required to deliver effective cognititve and behavioural therapy for people with depression and with anxiety disorders*. London: Department of Health.

Roth, A. D., & Pilling, S. (2008). Using an evidence-based methodology to identify the competences required to deliver effective cognitive and behavioural therapy for depression and anxiety disorders. *Behavioural and Cognitive Psychotherapy, 36*, 129-147. doi: http://dx.doi.org/10.1017/S1352465808004141

Rowe, G., Wright, G., & Bolger, F. (1991). Delphi: A re-evaluation of research and theory. *Technical Forecasting Social Change, 39*, 235-251.

Rutter, J. G., & Friedberg, R. D. (1999). Guidelines for the effective use of Socratic dialogue in cognitive therapy. In L. VandeCreek & T. L. Jackson (Eds.), *Innovations in clinical practice: A source book* (pp. 481-490). Sarasota, FL: Professional Resource Exchange.

Ryan, A., Safran, J. D., Doran, J. M., & Moran, J. (2012). Therapist mindfulness, alliance and treatment outcome. *Psychotherapy Research, 22*, 289-297. doi: http://dx.doi.org/10.1080/10503307.2011.650653

Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist, 55*, 68-78.

Ryan, R. M., & Deci, E. L. (2002). An overview of self-determination theory: An organismic-dialectical perspective. In E. L. Deci & R. M. Ryan (Eds.), *Handbook of self-determination research* (pp. 3-36). Rochester, NY: University of Rochester Press.

Ryan, R. M., & Deci, E. L. (2008). A self-determination theory approach to psychotherapy: The motivational basis for effective change. *Canadian Psychology, 49*, 186-193.

Ryan, R. M., Rigby, S., & King, K. (1993). Two types of religious internalization and their relations to religious orientations and mental health. *Journal of Personality and Social Psychology, 65*, 586-596.

Sackett, P. R., Lievens, F., Berry, C. M., & Landers, R. N. (2007). A cautionary note on the effects of range restriction on predictor intercorrelations. *Journal of Applied Psychology, 92*, 538-544. doi: http://dx.doi.org/10.1037/0021-9010.92.2.538

Sackett, P. R., & Yang, H. (2000). Correction for range restriction: An expanded typology. *Journal of Applied Psychology, 85*, 112-118. doi: http://dx.doi.org/10.1037/0021-9010.85.1.112

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

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

Safran, J. D., & Muran, J. C. (2006). Has the concept of the therapeutic alliance outlived its usefulness? *Psychotherapy: Theory, Research, Practice, Training, 43*, 286-291.

Safran, J. D., Muran, J. C., Samstag, L. W., & Stevens, C. (2001). Repairing alliance ruptures. *Psychotherapy: Theory, Research, Practice, Training, 38*, 406-412.

Saketopoulou, A. (1999). The therapeutic alliance in psychodynamic psychotherapy: Theoretical conceptualizations and research findings. *Psychotherapy: Theory, Research, Practice, Training, 36*, 329-342.

Samstag, L. W. (2006). The working alliance in psychotherapy: An overview of the invited papers in the special section. *Psychotherapy: Theory, Research, Practice, Training, 43*, 300-307.

Sburlati, E. S., Lyneham, H. J., Mufson, L. H., & Schniering, C. A. (2012). A model of therapist competencies for the empirically supported interpersonal psychotherapy for adolescent depression. *Clinical Child and Family Psychology Review, 15*, 93-112. doi: http://dx.doi.org/10.1007/s10567-012-0111-1

Schauer, C., Everett, A., del Vecchio, P., & Anderson, L. (2007). Promoting the value and practice of shared decision-making in mental health care. *Psychiatric Rehabilitation Journal, 31*, 54-61. doi: http://dx.doi.org/10.2975/31.1.2007.54.61

Schmidt, N. B., & Woolaway-Bickel, K. (2000). The effects of treatment compliance on outcome in cognitive-behavioral therapy for panic disorder: Quality versus quantity. *Journal of Consulting and Clinical Psychology, 68*, 13-18.

Scott, W. A. (1955). Reliability of content analysis: the case of nominal scale coding. *Public Opinion Quarterly, 19*, 321-325. doi: http://dx.doi.org/10.1086/266577

Segal, Z. V., & Shaw, B. F. (1996). Cognitive therapy. *American Psychiatric Press Review of Psychiatry, 15*, 69-90.

Selvage, R. (1976). Comments on the analysis of variance strategy for the computation of intraclass reliability. *Educational and Psychological Measurement, 36*, 605-609. doi: http://dx.doi.org/10.1177/001316447603600303

Sheehan, K. B. (2001). E-mail Survey Response Rates: A Review. *Journal of Computer-Mediated Communication, 6*, 0-0. doi: 10.1111/j.1083-6101.2001.tb00117.x

Shelef, K., & Diamond, G. M. (2008). Short form of the revised vanderbilt therapeutic alliance scale: Development, reliability, and validity. *Psychotherapy Research, 18*, 433 - 443.

Shirk, S., & Karver, M. (2003). Prediction of treatment outcome from relationship variables in child and adolescent therapy: A meta-analytic review. *Journal of Consulting and Clinical Psychology, 71*, 452-464.

Shrout, T. H., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*, 420-428.

Simons, A. D., Rozek, D. C., & Serrano, J. L. (2013). Wanted: Reliable and valid measures for the science of Cognitive Behavioral Therapy dissemination and implementation. *Clinical Psychology: Science and Practice, 20*, 181-194. doi: 10.1111/cpsp.12033

Singer, J., & Willet, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. New York: Oxford University Press.

Sireci, S. G. (1998). The construct of content validity. *Social Indicators Research, 45*, 83-117. doi: 10.1023/a:1006985528729

Sireci, S. G., & Geisinger, K. F. (1995). Using Subject-Matter Experts to Assess Content Representation: An MDS Analysis. *Applied Psychological Measurement, 19*, 241-255. doi: 10.1177/014662169501900303

Sori, C. F., & Sprenkle, D. H. (2004). Training family therapists to work with children and families: A modified Delphi study. *Journal of Marital and Family Therapy, 30*, 479-495. doi: http://dx.doi.org/10.1111/j.1752-0606.2004.tb01257.x

Spitzer, R. L., Williams, J. B., & Gibbon, M. (1987). *Instruction manual for the Structured Clinical Interview for DSM-III-R*. Biometrics Research Department, New York State Psychiatric Institute: New York.

Spring, B. (2007). Evidence-based practice in clinical psychology: What it is, why it matters; What you need to know. *Journal of Clinical Psychology, 63*, 611-631. doi: http://dx.doi.org/10.1002/jclp.20373

Stark, S., Chernyshenko, O. S., & Guenole, N. (2011). Can Subject Matter Experts' Ratings of Statement Extremity Be Used to Streamline the Development of Unidimensional Pairwise Preference Scales? *Organizational Research Methods, 14*, 256-278. doi: 10.1177/1094428109356712

Stein, D. J., Kupfer, D. J., & Schatzberg, A. F. (Eds.). (2006). *The American Psychiatric Publishing textbook of mood disorders*. Washington, DC: American Psychiatric Publishing.

Stemler, S. R. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research and Evaluation, 9*. Retrieved from http://PAREonline.net/getvn.asp?v=9&n=4

Stiles, W. B., Glick, M. J., Osatuke, K., Hardy, G. E., Shapiro, D. A., Agnew-Davies, R., . . . Barkham, M. (2004). Patterns of Alliance Development and the Rupture-Repair

Hypothesis: Are Productive Relationships U-Shaped or V-Shaped? *Journal of Counseling Psychology, 51*, 81-92. doi: http://dx.doi.org/10.1037/0022-0167.51.1.81

Stiles, W. B., Honos-Webb, L., & Surko, M. (1998). Responsiveness in psychotherapy. *Clinical Psychology: Science and Practice, 5*, 439-458.

Stiles, W. B., Leiman, M., Shapiro, D. A., Hardy, G. E., Barkham, M., Detert, N. B., & Llewelyn, S. P. (2006). What does the first exchange tell? Dialogical sequence analysis and assimilation in very brief therapy. *Psychotherapy Research, 16*, 408-421. doi: http://dx.doi.org/10.1080/10503300500288829

Streiner, D. L., & Norman, G. R. (1995). *Health measurement scales: A practical guide to their development and use*. New York: Oxford University Press.

Strong, S. R. (1968). Counseling: An interpersonal influence process. *Journal of Counseling Psychology, 15*, 215-224.

Strunk, D. R., Brotman, M. A., DeRubeis, R. J., & Hollon, S. D. (2010). Therapist competence in cognitive therapy for depression: Predicting subsequent symptom change. *Journal of Consulting and Clinical Psychology, 78*, 429-437.

Strunk, D. R., DeRubeis, R., Chiu, A., & Alvarez, J. (2007). Patients' competence in and performance of cognitive therapy skills: Relation to the reduction of relapse risk following treatment for depression. *Journal of Consulting and Clinical Psychology, 75*, 523-530.

Summers, R. F., & Barber, J. P. (2010). *Psychodynamic therapy: A guide to evidence- based practice*. New York: Guilford.

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate Statistics*. Boston, MA: Pearson/Allyn & Bacon.

Tang, T., & DeRubeis, R. (1999). Sudden gains and critical sessions in cognitive-behavioral therapy for depression. *Journal of Consulting and Clinical Psychology, 67*, 894-904.

Tang, T., DeRubeis, R., Beberman, R., & Pham, T. (2005). Cognitive changes, critical sessions, and sudden gains in Cognitive-Behavioral Therapy for depression. *Journal of Consulting and Clinical Psychology, 73*, 168-172.

Teasdale, J. D. (1997a). The relationship between cognition and emotion: The mind-in-place in mood disorders. In D. M. Clark & C. G. Fairburn (Eds.), *The science and practice of cognitive behaviour therapy* (pp. 67-93): Oxford University Press.

Teasdale, J. D. (1997b). The transformation of meaning: The interacting cognitive subsystems approach. In M. Power & C. R. Brewin (Eds.), *Meaning in psychological therapies: Integrating theory and practice* (pp. 141-156). New York: Wiley.

Tee, J., & Kazantzis, N. (2011). Collaborative empiricism in cognitive therapy: A definition and theory for the relationship construct. *Clinical Psychology: Science and Practice, 18*, 47-61. doi: 10.1111/j.1468-2850.2010.01234.x

Tee, J., Kazantzis, N., & Stukas, A. (2012). *The Collaborative Empiricism Scale*. Unpublished instrument.

Tetzlaff, J. M., Moher, D., & Chan, A.-W. (2012). Developing a guideline for clinical trial protocol content: Delphi consensus survey. *Trials, 13*, 43-54. doi: 10.1186/1745-6215-13-176

Thoma, N. C., & Cecero, J. J. (2009). Is integrative use of techniques in psychotherapy the exception or the rule? Results of a national survey of doctoral-level practitioners. *Psychotherapy (Chicago, Ill.), 46*, 405-417. doi: 10.1037/a0017900

Thomas, C., & Persons, J. B. (2013). Sudden gains can occur in psychotherapy even when the pattern of change is gradual. *Clinical Psychology: Science and Practice, 20*, 127-142. doi: 10.1111/cpsp.12029

Thompson, B. (2002). "Statistical," "practical", and "clinical": How many kinds of significance do counselors need to consider? *Journal of Counseling & Development, 80*, 64-71. doi: http://dx.doi.org/10.1002/j.1556-6678.2002.tb00167.x

Thomson, B. R. (1990). Appropriate and inappropriate uses of humor in psychotherapy as perceived by certified reality therapists: A delphi study. *Journal of Reality Therapy, 10*, 59-65.

Thwaites, R., & Bennett-Levy, J. (2007). Conceptualizing empathy in cognitive behaviour therapy: Making the implicit explicit. *Behavioural and Cognitive Psychotherapy, 35*, 591-612. doi: http://dx.doi.org/10.1017/S1352465807003785

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

Tinsley, H. E. A., & Weiss, D. J. (2000). Interrater reliability and agreement. In H. E. A. Tinsley & S. D. Brown (Eds.), *Handbook of applied multivariate statistics and mathematical modeling* (pp. 95-124). San Diego, CA: Academic Press.

Tracey, T. J. (1986). Interactional correlates of premature termination. *Journal of Consulting and Clinical Psychology, 54*, 784-788.

Tracy, J. J. (1977). Impact of intake procedures upon client attrition in a community mental health center. *Journal of Consulting and Clinical Psychology, 45*, 192-195.

Tryon, G. S., Blackwell, S. C., & Hammel, E. F. (2007). A meta-analytic examination of client–therapist perspectives of the working alliance. *Psychotherapy Research, 17*, 629-642.

Tryon, G. S., Blackwell, S. C., & Hammel, E. F. (2008). The magnitude of client and therapist working alliance ratings. *Psychotherapy: Theory, Research, Practice, Training, 45*, 546-551.

Tryon, G. S., & Winograd, G. (2001). Goal consensus and collaboration. *Psychotherapy: Theory, Research, Practice, Training, 38*, 385-389. doi: http://dx.doi.org/10.1037/0033-3204.38.4.385

Tryon, G. S., & Winograd, G. (2002). Goal consensus and collaboration. In J. C. Norcross (Ed.), *Psychotherapy relationships that work: Therapist contributions and responsiveness to patients* (pp. 109-125). New York: Oxford University Press.

Tryon, G. S., & Winograd, G. (2011). Goal consensus and collaboration. *Psychotherapy: Theory, Research, Practice, Training, 48*, 50-57.

Tryon, W. W., & Misurell, J. R. (2008). Dissonance induction and reduction: A possible principle and connectionist mechanism for why therapies are effective. *Clinical Psychology Review, 28*, 1297-1309.

Tukey, J. W. (1977). *Exploratory Data Analysis*. Reading, Mass.: Addison-Wesley.

Turkington, D., & Siddle, R. (1998). Cognitive therapy for the treatment of delusions. *Advances in Psychiatric Treatment, 4*, 235-241.

Ulschak, F. L. (1983). *Human resource development: The theory and practice of need assessment*. Reston, VA: Reston Publishing.

Ulvenes, P. G., Berggraf, L., Hoffart, A., Stiles, T. C., Svartberg, M., McCullough, L., & Wampold, B. E. (2012). Different processes for different therapies: Therapist actions, therapeutic bond, and outcome. *Psychotherapy, 49*, 291-302. doi: http://dx.doi.org/10.1037/a0027895

Vallis, T. M., Shaw, B. F., & Dobson, K. S. (1986). The Cognitive Therapy Scale: Psychometric properties. *Journal of Consulting and Clinical Psychology, 54*, 331-385.

van Oppen, P. (2004). Cognitive therapy for obsessive-compulsive disorder. *Clinical Case Studies, 3*, 333-349.

Vansteenkiste, M., & Sheldon, K. M. (2006). There's nothing more practical than a good theory: Integrating motivational interviewing and self-determination theory. *British Journal of Clinical Psychology, 45*, 63-82.

Waddington, L. (2002). The therapy relationship in cognitive therapy: A review. *Behavioural and Cognitive Psychotherapy, 30*, 179-192.

Waltz, C. W., & Bausell, R. B. (1981). *Nursing research: Design, statistics and computer analysis*. Philadelphia, PA: Davis.

Waltz, J., Addis, M. E., Koerner, K., & Jacobson, N. S. (1993). Testing the integrity of a psychotherapy protocol: Assessment of adherence and competence. *Journal of Consulting and Clinical Psychology, 61*, 620-630. doi: http://dx.doi.org/10.1037/0022-006X.61.4.620

Webb, C. A., DeRubeis, R. J., & Barber, J. P. (2010). Therapist adherence/competence and treatment outcome: A meta-analytic review. *Journal of Consulting and Clinical Psychology, 78*, 200-211.

Webb, C. A., DeRubeis, R. J., Dimidjian, S., Hollon, S. D., Amsterdam, J. D., & Shelton, R. C. (2012). Predictors of patient cognitive therapy skills and symptom change in two randomized clinical trials: The role of therapist adherence and the therapeutic alliance. *Journal of Consulting and Clinical Psychology, 80*, 373-381. doi: http://dx.doi.org/10.1037/a0027663

Weck, F., Bohn, C., Ginzburg, D. M., & Stangier, U. (2011). Assessment of adherence and competence in cognitive therapy: Comparing session segments with entire sessions. *Psychotherapy Research, 21*, 658-669. doi: http://dx.doi.org/10.1080/10503307.2011.602751

West, B. T., Welch, K. B., & Galecki, A. T. (2007). *Linear mixed models: A practical guide using statistical software*. Boca Raton, FL: Chapman Hall / CRC.

Westefeld, J. S., Range, L. M., Rogers, J. R., Maples, M. R., Bromley, J. L., & Alcorn, J. (2000). Suicide: An overview. *The Counseling Psychologist, 28*, 445-510.

Whitehurst, G. J. (1984). Interrater agreement for journal manuscript reviews. *American Psychologist, 39*, 22-28. doi: http://dx.doi.org/10.1037/0003-066X.39.1.22

Wilkes, T. (1994). *Cognitive therapy for depressed adolescents*. New York: Guilford.

Will, T. E. (1995). Cognitive therapy and the MCMI-III. In P. D. Retzlaff (Ed.), *Tactical psychotherapy of the personality disorders: An MCMI-III-based approach* (pp. 90-110). Needham Heights, MA: Allyn & Bacon.

Williams, M., & Haverkamp, B. E. (2010). Identifying critical competencies for psychotherapeutic practice with eating disordered clients: A delphi study. *Eating Disorders: The Journal of Treatment & Prevention, 18*(2), 91-109. doi: http://dx.doi.org/10.1080/10640260903585524

Willis, G. B. S., Schechter, S., & Whitaker, K. (2000). A comparison of cognitive interviewing, expert review, and behavior coding: What do they tell us? *American Statistical Association*.

Windholz, M. J., & Silberschatz, G. (1988). Vanderbilt Psychotherapy Process Scale: A replication with adult outpatients. *Journal of Consulting and Clinical Psychology, 56*, 56-60.

Wiseman, H., Tishby, O., & Barber, J. P. (2012). Collaboration in Psychodynamic Psychotherapy. *Journal of Clinical Psychology, 68*, 136-145. doi: 10.1002/jclp.21834

Wong, C. W. (in press). Collaborative empiricism in culturally sensitive cognitive behavior therapy. *Cognitive and Behavioral Practice*. doi: http://dx.doi.org/10.1016/j.cbpra.2012.08.005

Worthington, R. L., & Whittaker, T. A. (2006). Scale development research: A content analysis and recommendations for best practices. *The Counseling Psychologist, 34*, 806-838.

Wynd, C. A., Schmidt, B., & Schaefer, M. A. (2003). Two quantitative approaches for estimating content validity. *Western Journal of Nursing Research, 25*, 508-518.

Yates, S. L., Morley, S., Eccleston, C., de, C., & Williams, A. C. (2005). A scale for rating the quality of psychological trials for pain. *Pain, 117*(3), 314-325. doi: http://dx.doi.org/10.1016/j.pain.2005.06.018

Young, J. E., & Beck, A. T. (1980a). *Cognitive Therapy Scale*. Unpublished Manuscript. University of Pennsylvania. Philadelphia, PA.

Young, J. E., & Beck, A. T. (1980b). *Cognitive Therapy Scale: Rating manual*. Unpublished Manuscript. University of Pennsylvania. Philadelphia, PA.

Young, J. E., & Beck, A. T. (1990). *Manual of the Cognitive Therapy Rating Scale*. Philadelphia: University of Pennsylvania, Center for Cognitive Therapy.

Young, J. E., Klosko, J. S., & Weishaar, M. E. (2006). *Schema therapy: A practitioner's guide*. New York: Guilford.

Young, S. J., & Jamieson, L. M. (2001). Delivery methodology of the Delphi: A comparison of two approaches. *Journal of Park and Recreation Administration, 19*, 42-58.

Yousuf, M. I. (2007). Using experts' opinions through delphi technique. *Practical Assessment, Research and Evaluation, 12*. Retrieved from http://pareonline.net/