The Compassion Focused Therapy Therapist Rating Scale: Psychometric properties and expert opinion

Thesis submitted in partial fulfilment of the degree of

Doctorate in Clinical Psychology

University of Leicester

by

Grace Thorne

Department of Clinical Psychology

University of Leicester

May 2020

Declaration

I hereby declare that the following literature review and empirical project are original pieces of work, submitted in partial fulfilment of the degree of Doctorate in Clinical Psychology (DClinPsy) and not for any other academic award or degree. Prior to submission this work was checked to ensure that it was complete.

The Compassion Focused Therapy Therapist Rating Scale: Psychometric properties and expert opinion

Grace Thorne

Thesis Abstract

Treatment fidelity concerns whether therapy is implemented as intended and an important aspect of treatment fidelity is therapist competence. Therapist competence scales are a popular method of assessing whether therapy is competently delivered, however these scales vary in their development and evaluation, and in their resulting reliability and validity. This thesis presents a review of therapist competence scales and an empirical study that investigated a new therapist competence scale for Compassion Focused Therapy.

Systematic Literature Review

The systematic literature review explored the development and psychometric evaluation of therapist competence scales for cognitive and cognitive-behavioural therapies. Four databases were searched, and thirteen papers were included in the review. The standard of papers was assessed using a bespoke quality appraisal tool and results were narratively synthesised. There was little consensus about the methods that should be used to develop and evaluate competence scales and it was concluded that researchers might use multiple methods of assessing competence rather than rely on competence scales alone.

Empirical Research Project

A mixed-methods approach was used to explore the psychometric properties and expert feedback of the Compassion-Focused Therapy Therapist Rating Scale (CFT-TCRS; Horwood *et al.,* 2019). CFT experts watched clips of simulated CFT and used the CFT-TCRS to assign competence ratings before engaging in a semi-structured interview. Inter-rater reliability between participants was 'good'. Content analysis of the expert feedback provided a number of useful suggestions to improve the scale. Once amended the CFT-TCRS may become a useful tool for clinical practice, therapist training and continued research into aspects of treatment fidelity in CFT.

Acknowledgments

Firstly, I would like to say thank you to the participants who took part in my study and contended with time differences and (occasionally) patchy video calls. This research would not have been possible without you.

I would also like to thank my research supervisor, Dr Steve Allan, for guiding me through the mysterious and unknown territory of doctoral research. Your input, support and reassurance gave me confidence and kept me on track during times of anxiety (of which there were several). To my field supervisor, Dr Ken Goss, thank you for helping me design and set up this research. Thank you also to Professor Paul Gilbert OBE and the Compassionate Mind Foundation for their involvement in this research.

To my incredible family, I literally would not be here without you, and I cannot thank you enough for supporting me through this journey and for telling me, “You can do this!” at times I felt I couldn’t.

To my friends, thank you for your unwavering belief in me, lending your ears, dragging me out of the house, feeding me and making me laugh when I needed it most; I honestly could not have done this without you. Special thanks go to my cohort pals for making the process hilarious and memorable, and for the endless supportive phone calls.

Finally, I would like to thank my cats, Dwight and Eddie, for always being non-judgmental and providing much needed comfort during times of stress. But next time I’m writing a thesis please stop being so cute, it’s distracting.

Word Counts

Thesis Abstract:	258
Part 1: Systematic literature review	
Abstract	274
Full text	6806
References	1332
Part 2: Empirical research project	
Abstract	300
Full text	7900
References	1246
Appendices	
Mandatory appendices	7123
Non-mandatory appendices	5976
Total Thesis Word Count	20940

NB: All word counts exclude diagrams and tabulated numeric data. The total thesis word count also excludes contents pages, references and the mandatory appendices.

Table of Contents

Appendices

*Denotes mandatory appendices

Addendum 1: Interview transcripts (submitted electronically)

Addendum 2: Content analysis summary tables (available on request)

List of Tables

Part 1: Systematic literature review

Part 2: Empirical research project

List of Figures

Part 1: Systematic literature review

Part 1: Systematic Literature Review

The development and psychometric evaluation of therapist competence scales for cognitive and cognitive-behavioural therapies: A systematic review

Abstract

Objectives

The National Institute for Health and Clinical Excellence (NICE) guidelines recommend cognitive behavioural therapy (CBT) for treating a wide variety of psychological difficulties. Research has highlighted the importance of measuring therapist competence to ensure therapy quality and to improve clinical practice, research, training and audits. Several scales have been developed to measure therapist competence in CBT. The aim of the current review was to systematically evaluate and synthesise evidence on the methods used to develop and psychometrically evaluate these scales.

Methods

Four databases (PsychInfo, Scopus, Medline and Web of Science) were systematically searched using terms related to measuring therapist competence. Thirteen articles describing and examining eleven CBT competence scales met the inclusion criteria and were evaluated using a bespoke quality appraisal tool. Evidence was narratively synthesised to address the aims of the review.

Results

There was little consensus in the methods used to develop and evaluate CBT competence scales. Some studies used comprehensive development procedures and evaluated several psychometric properties, whereas others focused on just one or two aspects of reliability and validity. The findings highlighted variation in the psychometric properties of the scales. However, it was not always the case that comprehensive development procedures increased the validity and reliability found.

Conclusions

Some scales are used more widely than others, however it was difficult to conclude which development and evaluation methods should be favoured when developing competence scales. Future research should focus on employing more comprehensive development strategies and psychometric evaluations. Although some scales can be used to reliably assess therapist competence in CBT, scale users should be aware of their limitations and might consider using multiple methods of assessing therapist competence.

1 Introduction

1.1 Treatment fidelity and Cognitive Behavioural Therapy

Cognitive Behavioural Therapy (CBT) was developed in the 1960s and is an effective treatment for psychological difficulties (Hazell *et al.,* 2016; Stewart & Chambless, 2009). In the UK it is recommended by the National Institute for Health Care Excellence (NICE) for the treatment and management of experiences such as depression (NICE, 2018), anxiety (NICE, 2019) and psychosis (NICE, 2014). CBT has continued to evolve since its inception (Boyle *et al.,* 2019) and has influenced the development of more recent 'third-wave' psychotherapies such as Compassion-Focused Therapy (CFT; Gilbert, 2009) and Cognitive Analytic Therapy (CAT; Ryle, 2005).

1.2 Treatment fidelity: Competence and adherence

An important issue for all psychological therapies is 'treatment fidelity' or 'therapy fidelity'. This has been defined as "the degree to which treatment is delivered as intended" (Yeaton & Sechrest, 1981, p.160). One aspect of treatment fidelity is 'competence', understood as "the extent to which a therapist has the knowledge and skill required to deliver a treatment" (Fairburn & Cooper, 2011, p.374). When considering competence, it is also important to consider 'adherence' (another aspect of treatment fidelity) due to an overlap in the concepts. Therapy adherence is defined as the extent to which therapists apply the methods and techniques of a manualised intervention (Webb *et al.,* 2010).

Effective delivery of psychological therapy is dependent on therapist competence and adherence. Whilst measures have been developed for assessing adherence (e.g., Barber & Crits-Christoph, 1996), adhering to a therapy does not necessarily mean it is being competently executed (Bennett & Parry, 2004). Therefore, it is also critical that treatments are skilfully and competently implemented. Unfortunately, the evidence base around therapist competence is limited by a lack of psychological research into variables confounding treatment fidelity (Perepletchikova, 2011), and difficulties with defining and assessing competence (Barber *et al.,* 2007). Despite these challenges, it remains important that adherence and competence is assessed in order to clarify

therapeutic change mechanisms (Perepletchikova & Kadzin, 2005) and improve implementation of psychological interventions (McLeod *et al.,* 2013).

1.3 Therapist competence scales

Several methods have been developed for assessing therapeutic competence, such as essays and questionnaires (Muse & McManus, 2013). One important method has been to use therapist competence scales to evaluate treatment sessions. Competence scales have been developed for several therapeutic modalities, such as dynamic therapy (Barber *et al.,* 1997) and Compassion-Focused Therapy (CFT; Horwood *et al.,* 2019). Competence scales also commonly allow for the measurement of therapy adherence given that is a prerequisite for therapeutic competence (Barber *et al.,* 2003). Competence scales usually assess knowledge of the underlying theory and techniques specific to the intervention in question, as well as the more general therapeutic skills (Barber *et al.,* 2007).

Most competence scales have been designed for use by highly trained or expert clinicians as they often assess trainee therapists (e.g., CAT; Bennett & Parry, 2004). Furthermore, they can be used for therapy research, clinical audits and service evaluation (Bennett & Parry, 2004). However, the literature indicates variability in the methods used to develop and psychometrically evaluate competence scales (Barlow & Brown, 2019) and there are no recommendations about which methods are best. Some developers (e.g., Horwood *et al.,* 2019) used structured approaches such as the Delphi method (Linstone & Turoff, 1975) and others (e.g., Barber *et al.,* 1997) have used existing scales and treatment manuals to develop competence scales. Scale developers have also used different approaches when evaluating the psychometric properties of competence scales. For example, some have used videotapes of therapy sessions and others have used audio tapes, some have recruited experts to use the scale to rate tapes for competence and others have recruited students.

1.4 Competence in CBT

Competence in CBT is defined as "the degree to which a therapist demonstrates the general therapeutic and treatment specific knowledge and skills required to appropriately deliver CBT", and should be based in evidence and on the specific problem

of the patient (Muse & McManus, 2013, p.485). CBT competence measures have been used since the 1980s (Young & Beck, 1980), and a specific CBT competence framework has been developed (Roth & Pilling, 2007). Since the publication of this framework, several measures of CBT competence have been published, but not all have been included in a systematic review.

1.5 Previous literature reviews

Two previous reviews have been conducted into the development, properties and use of measures of therapist competence. Barlow and Brown (2019) explored competence scales for interpersonal, dynamic and relational models and found no consensus around which methods should be used to develop them. Furthermore, some studies did not provide adequate information on participants, recruitment and scale developers. They found that the quality and depth of psychometric testing was not adequate for the scales to be considered reliable and valid. For example, some studies inappropriately measured inter-rater reliability. They suggested that variability in the development and evaluation of competence scales might be contributing to a lack of consistently used measures and concluded that more research is required into the development and psychometric evaluation of competence scales.

Muse and McManus (2013) explored the psychometric properties of competence assessments. They noted four types of CBT competence assessment: 'knowledge-based' (e.g., questionnaires), 'practical' (e.g., case reports), 'practical application' (e.g., role-plays) and 'clinical practice' (e.g., competence scales). Seven competence scales were identified, where some were transdiagnostic and measured general CBT competencies, and some were disorder-specific and covered specific treatment protocols. Limitations included a lack of psychometric exploration outside of the controlled trials in which the scales were developed and difficulty identifying the best methods of psychometrically evaluating them. Despite this, they concluded that competence scales were the most comprehensive method of assessing therapists' skills. It was suggested that new and existing competence measures should have implementation protocols, clear benchmarking and clarity around who should complete the assessment. Finally, they posited that a 'multi-method' approach to assessing CBT competence might be more appropriate given that no one method suitably assessed all aspects of CBT.

1.6 Summary and rationale

No previous reviews have focused on the methods of developing and psychometrically evaluating therapist competence scales for cognitive therapy (CT) and CBT. Several new scales have been published since the recommendation by Muse and McManus (2013) but have not been included in a systematic review. The present review aimed to synthesise current quantitative literature on therapist competence scales for CT/CBT. Cognitive therapies were prioritised due to their widespread implementation and influence on recent 'third wave' psychotherapies. The review aimed to focus on the methods used to develop and psychometrically evaluate these scales. This was in line with the review by Barlow and Brown (2019) but differed from that of Muse and McManus (2013) in that the present review focused on just the competence scale approach to measuring therapist competence. Given that adherence is required for competent intervention delivery (Barber *et al.,* 2003), scales measuring both adherence and competence were considered for the present review.

Therefore, the main questions addressed in this review were:

- What methods have been used to develop therapist competence scales for CT and CBT?
- What methods have been used to psychometrically evaluate therapist competence scales for CT and CBT?

Based on the findings of the research questions, the present review aimed to provide recommendations for clinical practice and research in relation to the development and evaluation of published competence scales for CT and CBT.

2 Method

2.1 Inclusion criteria

To meet the objectives of this review, studies had to meet the following inclusion criteria: (a) studies described the development and validation of a measure of therapist competence or competence and adherence, (b) studies described measures specifically designed for use with CT/CBT, (c) studies described measures designed for use by experts or trainers in CT/CBT (assessor-rated scales are the most commonly used

methods of assessing competence according to previous reviews), (d) studies described competence measures designed to assess individual face-to-face therapy, rather than group, family or online therapy, (e) studies were written in English, (f) the methodology was quantitative (due to the focus on psychometric evaluation methods), (g) studies were published in a peer reviewed journal. Studies included for review were not restricted by the year of their publication.

2.2 Search strategy

A search strategy was designed to identify all papers for possible inclusion in this review. Scoping searches were carried out in August 2019 which enabled the author to ascertain the breadth of existing literature. Some relevant studies were identified, and their keywords explored to develop inclusive search terms. Furthermore, the Cochrane Database and PROSPERO (International Prospective Register of Systematic Reviews) were searched to identify previous, current or proposed systematic reviews on the current topic.

Database searches of PsychInfo, Scopus, Medline and Web of Science were conducted in September 2019 and January 2020. The broad search terms, chosen to ensure that no key papers were missed, were related to the development and validation of scales or measures of therapeutic competence, for example; 'therap*', 'competen*', 'scale' and 'psychometric*' (see Appendix A for full terms and filters). Articles were exported into referencing software and duplicates were removed before the titles and abstracts were screened for relevance. Following this, the remaining articles were read in full and checked against the inclusion criteria. Final papers had their reference lists screened to identify further papers which might have been missed during searches, and these papers were also checked against the inclusion criteria.

2.3 Study selection

For an overview of study selection see Figure 1. Following searches, 12098 articles were exported into Mendeley where duplicates were removed. The titles of 10642 articles were screened, of which 10113 were removed leaving 529 to have their abstracts screened for relevance. Following this, 117 potentially relevant articles were obtained in full and assessed against the inclusion criteria which led to 104 being excluded.

Following this process 13 articles remained. Their reference lists were screened for further relevant papers, but none were identified. Therefore, these 13 articles were subjected to quality appraisal and data extraction. It became apparent that the data was largely descriptive, and methods used in scale development and evaluation were heterogeneous. For example, different methods of statistical analysis were used and difficulties with synthesis meant that a meta-analytic approach to the review was not appropriate. For these reasons, and due to the nature of the aims and research questions proposed in the review, a narrative synthesis of the data was undertaken.

2.4 Quality appraisal

No published quality appraisal tools for assessing studies exploring psychometric properties were found. Therefore, in line with Barlow and Brown's (2019) review, the current review used a bespoke tool for assessing study quality (see Appendix B). This tool included four items from the 11-item Quality Appraisal of Diagnostic Reliability (QAREL; Lucas *et al.*, 2010), designed to assess the quality of diagnostic tests. It also included three items from the 27-item Downs and Black (1998) checklist for measuring study quality which was designed for assessing the quality of healthcare interventions. Items from these two tools that were not relevant to investigating psychometric properties were excluded from the bespoke quality tool designed for the current review.

Finally, in order to fully assess papers, nine items were generated to cover aspects of quality related to development and validation of therapist competence scales (e.g., methods used and what therapy the scale was designed for use with). A further four items were included to obtain data on types of analysis, conclusions, limitations and clinical implications. In total, the final bespoke tool had 20 quality appraisal questions (see Appendix C for information on each item included in the quality appraisal tool).

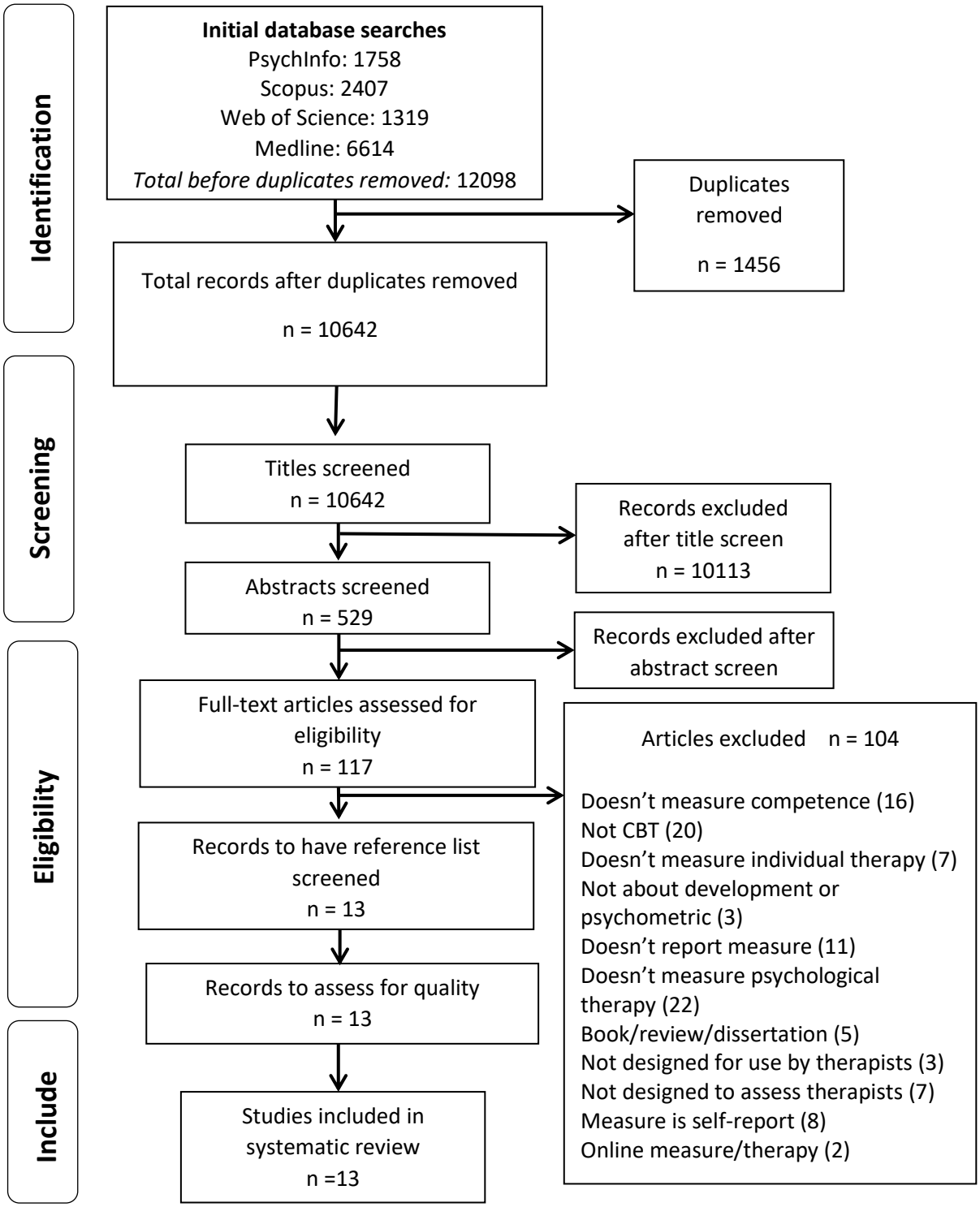


Figure 1. *Flow chart detailing the systematic search process*

2.5 Data Extraction

The data extraction was built into the quality appraisal tool (see Appendix B) and focused on scale development, characteristics of developers, measure characteristics (e.g., items and scoring), the methods and samples used in psychometric evaluation (e.g., audio/videotapes and raters, clients and therapists), aspects of reliability and validity measures and the statistical analyses used to examine these, the results and the main discussions points (conclusions, limitations and implications). A summary of extracted data can be found in Appendix D (aims, samples and design) and Appendix E (methods of scale development and testing reliability/validity, analyses and results of interest).

3 Results

A total of 13 articles were included following the systematic search process (summarised in Appendix D and E). Study characteristics, quality and the main findings are summarised and reported in relation to scale development and psychometric evaluation. Due to similarities to the review by Barlow and Brown (2019), brief summary tables in this review will use the same headings.

3.1 Description of CBT competence scales

The included studies covered the development and validation of 11 CBT therapist competence scales (see Table 1 for basic scale characteristics). The psychometric properties of nine measures were described in one paper each. The properties of the Cognitive Therapy Scale (CTS; Young & Beck, 1980) were reported in two papers (Dobson *et al.*, 1985; Vallis *et al.,* 1986). The development of the UCL Scale was described by Roth (2016), and the psychometric evaluation by Roth *et al.* (2019). Eight studies reported the scale development and psychometric evaluation. However, information on the development of the CTS was reported in an unavailable paper (Young & Beck, 1980), as was the CTCS-SP (Clark *et al.,* 2007).

The type of therapy and clients for whom the scales were designed differed across studies (see Table 1). Five studies measured competence of therapists in cognitive therapy (CT), four described CBT competence scales, three measured competence in

CBT for children and young people and one was for CBT for psychosis. The YACS (Carroll *et al.,* 2000) has one subscale for measuring CBT but also measure clinical management (CM) and twelve-step facilitation (TSF), and the UCL scales (Roth, 2016) measure CBT and generic therapeutic competence. Given the extensive conceptual and therapeutic overlap of CT and CBT this distinction is not raised again in the remainder of the review.

Eight studies presented scales measuring therapist competence alone, and five presented measures of both competence and adherence. The number of items in the scales varied, but all were designed to be used by an observer with expertise in the intervention being applied. Furthermore, the scales used 7-point Likert scales to measure competence and adherence to the intervention except for the ITIS (Boyle *et al.,* 2019) which used a 3-point scale for adherence and a 7-point scale for competence, and the YACS (Carroll *et al.,* 2000) and the UCL scale (Roth, 2019) which used a 5-point Likert scale.

3.2 Quality

The quality appraisal tool enabled assessment of overall study quality. It provided a score out of 20 and based on a grading scale published by O'Connor *et al.* (2015) a score of less than half (>10) was 'poor'. Based on this, six papers were 'excellent' (17-20), five were 'good' (15-17), one was 'fair' (11-14) and one was 'poor'. Based on the quality appraisal summary (see Appendix F), there did not appear to be a common element of quality which separated the 'good' and 'excellent' papers from the 'poor' and 'fair'. However, the factors most commonly impacting the quality of papers were; scale developer characteristics not being adequately described and lack of clarity around whether developers represented a specific population.

The paper by Vallis *et al.* (1986) was assessed as being 'fair' due to a lack of clarity around scale development, which was described in an unpublished paper (Young & Beck, 1980). The paper by Roth (2016) was assessed as 'poor' as it only described scale development. The psychometric evaluation was published by Roth *et al.* (2019), and therefore the two papers describing the UCL scales were considered in conjunction, and the 'poor' paper was included.

3.3 The development of CBT competence scales

Nine of the included papers presented information on the scale development, including who was involved in the process and the methods used. See Table 2 and Appendix E for full details of scale development methods, scale characteristics, psychometric properties, results and conclusions.

Table 1. Study and measure characteristics

Study	Measure Name	Model	Number of items	Method of measurement
Barber *et al.,* (2003)	The Cognitive Therapy Adherence and Competence Scale (CTACS)	Cognitive therapy (CT)	21 (competence and adherence)	Observer rating scale (7-pt)
Bjaastad *et al.* (2016)	The Competence and Adherence Scale for Cognitive Behavioural Therapy (CAS-CBT)	Cognitive behavioural therapy (CBT) for youth anxiety	11 (competence and adherence)	Observer rating scale (7-pt)
Blackburn *et al.* (2001)	The Revised Cognitive Therapy Scale (CTS-R)	Cognitive therapy (CT)	14 (competence and adherence)	Observer rating scale (7-pt)
Boyle *et al.* (2019)	The Inventory of Therapeutic Interventions and Skills (ITIS)	Modern, personalised cognitive behaviour therapy (CBT)	19 measuring adherence, 11 measuring competence	Observer rating scale (3-pt adherence, 7-pt competence)
Carroll *et al.* (2000)	The Yale Adherence and Competence Scale (YACS)	Cognitive behavioural therapy (CBT)	55 in total (competence and adherence) CBT subscale = 6	Observer rating scale (5-pt)
Dobson *et al.,* (1985)	The Cognitive Therapy Scale (CTS)	Cognitive therapy (CT)	11 (competence only)	Observer rating scale (7-pt)
Haddock *et al.* (2001)	The Cognitive Therapy Scale for Psychosis (CTS-Psy)	Cognitive behavioural therapy (CBT) for psychosis	10 (competence only)	Observer rating scale (7-pt)
McLeod *et al.* (2018)	The Cognitive-Behavioural Treatment for Anxiety in Youth Competence Scale (CBAY-C)	Cognitive behavioural therapy (CBT) for youth anxiety	25 (competence only)	Observer rating scale (7-pt)
Stallard *et al.,* (2014)	The Cognitive Behaviour Therapy Scale for Children and Young People (CBTS-CYP)	Cognitive behavioural therapy (CBT) for children and young people	14 (competence only)	Observer rating scale (7-pt)
Vallis *et al.,* (1986)	The Cognitive Therapy Scale (CTS)	Cognitive therapy (CT)	11 (competence only)	Observer rating scale (7-pt)
von Consbruch *et al.,* (2012)	The Cognitive Therapy Competence Scale for Social Phobia (CTCS-SP)	Cognitive therapy (CT) for social phobia	16 (competence only)	Observer rating scale (7-pt)

3.3.1 Developers of competence scales

Six scales were developed by the study's authors (Boyle *et al.,* 2019; Carroll *et al.,* 2000; Haddock *et al.,* 2001; McLeod *et al.,* 2018; Roth, 2016; 2018; Stallard *et al.*, 2014). Three studies employed trained CBT therapists or mental health professionals (Barber *et al.*, 2003; Bjaastad et *al.,* 2016; Stallard *et al.*, 2014) who were asked to review and pilot the scales before suggesting further developments. Three had input from 'clinicians' or 'experts' in CBT (Bjaastad et *al.,* 2016; Blackburn *et al.,* 2001; Roth, 2016) who generated scale items and fed back on early versions of scales. One study employed trainee therapists and trainee clinical psychologists (Stallard *et al.,* 2014) who were included in discussions about the CBTS-CYP. One study used graduate psychology students (Barber *et al.,* 2003) who used the scale to rate therapy videos before providing feedback, and one study included other 'therapists' (Boyle *et al.,* 2019) during discussions about the ITIS.

3.3.2 Methods used to develop CBT therapist competence scales

Scale development methods varied across studies (see Table 2). Most were based on published measures of therapist competence, with some including the items from an existing scale (e.g., the CBTS-CTP; Stallard *et al.,* 2014) (see Appendix E). Four studies used existing therapy manuals and/or competence frameworks to aid scale development and two conducted literature reviews to determine required competencies. Five studies relied on the opinions and expertise of the scale authors and six used discussions with other experts to gain consensus about which items to include. Six scales were piloted during development to assess the suitability of items and to inform further revisions.

Based on authors' descriptions, and through comparison of development methods, the CTACS (Barber *et al.,* 2003) went through a comprehensive development process including reviews of existing scales, consensus building with experts and pilot studies. Similarly, the CTCS-CYP (Stallard *et al.,* 2014) was comprehensively developed through consensus building, existing scales, a literature review and piloting. Based on comparisons with other scales, the YACS (Carroll *et al.,* 2000) appeared to have been less comprehensively developed through reviews of therapy manuals and opinions of

the scale authors. See Table 2 for a brief summary of the development process of CBT competence scales.

Table 2. Methods of scale development, analysis of psychometric properties and results.

Measure	Development	Internal Consistency	Inter-rater reliability	Convergent validity	Other construct validity	Factor analysis	Content validity
CTACS	Existing measure, manuals, pilot studies, expert opinion	Cronbach's alpha: α = .92 (adher) α = .93 (comp)	ICC: .67 (adher) .73 (comp)	Not reported	Sensitivity to therapy type: difference between CT and other therapies p < .0005	Not reported	Not reported
CAS-CBT	Existing measure, expert opinion, pilot studies.	Cronbach's alpha: α = .87	ICC: .83 (adher) .64 (comp)	Not reported	Rater stability, ICC: .89 (adher) .92 (competence)	2 factors explaining 66.6% of variance	Not reported
CTS-R	Existing measure, expert opinion.	Cronbach's alpha: α = .92 - .97	ICC: .63 Pearson correlation (one rater removed): r = 0.77 p = <.0001	Not reported	Sensitivity to trainee improvement: t = 2.68 p < .02	Not reported	Not reported
ITIS	Existing measures, author consensus building	Not reported	Kendall's W: .832 (intervention items) .700 (skill items)	Not reported	Not reported	1 factor explaining 50.9% of variance	Not reported
YACS	Manuals, treatment sessions, author consensus/ opinion.	Not reported	ICC: .80 – .95 (adher) .71 – .97 (comp)	Correlations: subscales differ (p = <0.001)	Sensitivity to therapy type: CBT items higher when CBT rated p = <.05	Not reported	Not reported

Measure	Development	Internal Consistency	Inter-rater reliability	Convergent validity	Other construct validity	Factor analysis	Content validity
CTS (Dobson *et al.,* 1985)	Developed by Young & Beck (1980)	Cronbach's alpha: α = .95	ICC: .96 Pearson correlation: r = 0.94	Not reported	Item-total correlations, Cronbach's alpha: α = .72 (apart from homework item)	Not reported	Not reported
CTS (Vallis *et al.,* 1986)	Developed by Young & Beck (1980)	Pearson correlation-items vs total score: r = 0.59 – 0.91	ICC: .59 for single rater	Not reported	Sensitivity to therapy quality using Rao: correct quality classification in 84.91% of cases	2 factors explaining 73.7% of variance	Not reported
CTS-Psy	Existing scale, author opinions, pilot studies.	Not reported	ICC: Total score: .94	Not reported	Sensitivity to trainee improvement/skill acquisition: F = 10.5, p = .004	Not reported	Good content validity decided by mental health professionals.
CBAY-C	Existing scales, scale developers/ authors opinion, scoring consensus, pilot studies.	Not reported	ICC: Items: 0.69	CBAY-C scores significantly differed from scores related measures (p < .01)	Not reported	Not reported	Not reported

Measure	Development	Internal Consistency	Inter-rater reliability	Convergent validity	Other construct validity	Factor analysis	Content validity
CBTS-CYP	Existing scale, literature review, consensus building of authors/experts, pilot studies	Not reported	ICC: Total score: .96	Correlation with CTS-R: r = .98, p < .0001	Sensitivity to therapy quality: agreed with CTS-R in 77% of cases. Sensitivity to trainee improvement: similar scores on CBTS-CYP and CTS-R	Not reported	Not reported
UCL scales (Roth, 2016; Roth *et al.,* 2019)	Existing framework, framework author, other clinicians, pilot studies.	Not reported	ICC (Roth, 2019) CBT scales: .39 Generic scale: .27	Not reported	Not reported	Not reported	Not reported
CTCS-SP	Existing scale, expert opinions.	Cronbach's alpha: α = .89	ICC: Total score: .81 Items: .62 - .92	Not reported	Retest reliability, ICC: .86	Not reported	Not reported

3.4 Participants and methods used to psychometrically evaluate CBT competence scales

All included papers, aside from Roth (2016), described the process of psychometrically evaluating the scales, although the information provided was varied. Psychometric evaluation is considered in terms of participants (patient, therapists and raters of therapists) and methods.

3.4.1 Participants used in psychometric evaluation

Most studies adequately described the patients, however three papers (Carroll *et al.,* 2000; Vallis *et al.,* 1986; von Consbruch *et al.,* 2012) provided limited details (see Appendix F). For eight of the studies, patient data came from previous RCTs, and the remaining four collected therapy tapes from trainee therapists. Patient samples varied, which was to be expected as the scales were developed for different types of CBT and clinical populations. The therapists were adequately described by most of the studies. However, two (Carroll *et al.,* 2000; von Consbruch *et al.,* 2012) provided insufficient information (see Appendix F).

All but two papers (Blackburn *et al.,* 2001; Dobson *et al.*, 1985) sufficiently described the raters using the scale and all detailed the number recruited and their professions. An average of 5.41 raters were used across studies (range 2-12), drawn from varying populations. Seven studies used CBT experts (e.g., therapists and supervisors), four used psychology graduates or clinical psychology trainees, two used qualified clinical psychologists/psychotherapists and three used masters/doctoral level clinicians. Haddock *et al.* (2001) also used a mental-health nurse and a research fellow. Bjaastad *et al.* (2016) used two of the authors as 'expert raters' and von Consbruch *et al.* (2012) used one of the authors as a rater. Full participant details are presented in Appendix D.

3.4.2 Methods used in psychometric evaluation

Of the 12 papers exploring psychometric properties of competence scales, ten used videotaped therapy sessions. Of these, seven used videos from an RCT and three used videos from trainee CBT therapists. One study used audio tapes from an RCT and one used audio tapes from trainee therapists (see Appendix D). Recordings were randomly selected from larger samples, apart from those used by Haddock *et al.* (2001), who

selected videos to reflect a range of competence levels, and Roth *et al.* (2019) who purposively selected videos from a larger sample to represent different clinical presentations.

In six studies, raters watched all available material and used the scale to rate therapist competence, and in four studies the raters viewed some of the material, either in assigned pairs or as part of a balanced design (see Appendix E). In the study by Bjaastad *et al.* (2016) videos were rated by students in a balanced design, and a percentage of these videotapes (20%) were also rated by experts. In the study by von Consbruch *et al. (*2012) one of the authors rated all the videotapes as a "standard rater", and others rated only some of the material. Their ratings were paired with the ratings of the "standard rater" to determine inter-rater reliability.

3.5 Psychometric characteristics of CBT competence scales

All studies, aside from Roth (2016), reported psychometric evaluation and properties of therapist competence scales; however, not all studies assessed the same aspects of reliability and validity (see Table 2). Furthermore, none provided details on how decisions were made about which psychometric properties to evaluate and how.

However, despite a lack of consistency across studies, scales all had at least two aspects of validity and reliability tested, apart from Roth *et al.,* (2019). Inter-rater reliability was measured for all 11 scales, with the CTS being tested twice (Dobson *et al.*, 1985; Vallis *et al.,* 1986). Six studies measured internal consistency, three assessed convergent validity, one reported content validity, and several reported other aspects of construct validity.

Of the scales included, the CAS-CBT (Bjaastad *et al.,* 2016), the YACS (Carroll *et al.,* 2000) and the CTS (Dobson *et al.*, 1985; Vallis *et al.*, 1986) were the most thoroughly analysed, with data available for four aspects of reliability and validity. The ITIS (Boyle *et al.,* 2019) and the UCL scales (Roth *et al.,* 2019) were the least comprehensively analysed with two measurements of interest being reported for the ITIS and one for the UCL scales (see Table 2). As was also highlighted in the review by Barlow and Brown (2019), the measurements commonly reported were; inter-rater reliability, internal consistency, content validity, convergent validity, sensitivity to change, and factor analysis.

3.5.1 Inter-rater reliability

Inter-rater reliability was the only psychometric property reported by all studies. The inter-rater reliability of the ITIS (Boyle *et al.,* 2019) was assessed using Kendall's W (Legendre, 2010), as ratings on one of the subscales were non-continuous and agreement between its subscales was found to be high (see Table 2). However, the intraclass correlation coefficient (ICC; Shrout & Fleiss, 1979) was the preferred method of inter-rater reliability analysis, with 12 studies reporting ICC. ICC assesses agreement between different members of a group (e.g., raters of therapist competence). All studies, apart from two (Blackburn *et al.,* 2001; Haddock *et al.,* 2001) identified which of the ten variations of ICC (McGraw & Wong, 1996) was used.

Koo & Li (2016) noted that an ICC less than 0.5 is 'poor', 0.5-0.75 is 'moderate', 0.75-0.9 is 'good' and above 0.9 is 'excellent'. Based on this, two of the measures had 'excellent' inter-rater reliability (CTS-Psy: Haddock *et al.,* 2001; CBTS-CYP: Stallard *et al.,* 2014), two were 'good' (YACS: Carroll *et al.,* 2000; CTCS-SP: von Consbruch *et al.,* 2012), four were 'moderate' (CTACS: Barber *et al.,* 2003; CAS-CBT competence subscale: Bjaastad *et al.,* 2016; CTS-R: Blackburn *et al.,* 2001; CBAY-C: McLeod *et al.,* 2018) and the UCL scales (Roth *et al.,* 2019) were 'poor'. Inter-rater reliability of the CTS (Young & Beck, 1980) was measured in two studies. Dobson *et al.* (1985) found it to be 'excellent', whereas Vallis *et al.* (1986) found it to be 'moderate'. Vallis *et al.* (1986) suggested that this might be attributed to the heterogeneity of the therapists used by Dobson *et al.* (1985).

Of the 11 studies reporting ICC, two also assessed inter-rater reliability using Pearson's product moment coefficient. Dobson *et al.* (1985) used it to look at the reliability of individual items on the CTS and found a strong agreement between raters. Blackburn *et al.* (2001) calculated Pearson's for the CTS-R to assess the reliability of both the individual items and the overall score to compare the CTS-R to the CTS (Dobson *et al.,* 1985) on which it is based.

3.5.2 Internal consistency

Internal consistency was reported for five scales, in six studies. Vallis *et al.* (1986) used Pearson's to look at the relationship between item scores and the total scale score and found a strong correlation. The 11 items of the CTS are rationally divided into two

subscales, 'general skills' and 'specific cognitive therapy skills'. However, these subscales are not used separately to assess competence, and the strong correlation suggested that two distinct subscales might not be necessary. The other studies calculated Cronbach's alpha coefficients to assess internal consistency and found coefficients ranging from 0.87 and 0.97 (see Table 2). According to guidelines for interpreting Cronbach's alpha (George & Mallery, 2003) the internal consistency of two scales was 'good' (CAS-CBT: Bjaastad *et al.,* 2016; CTCS-SP: von Consbruch *et al.,* 2012) and three were 'excellent' (CTACS: Barber *et al.,* 2003; CTS-R: Blackburn *et al.,* 2001; CTS: Dobson *et al.,* 1985), suggesting that items in these scales were tapping into the same construct.

3.5.3 Content validity

The CTS-Psy (Haddock *et al.,* 2001) was the only scale for which content validity was reported. This was not formally analysed, but it was decided by 'mental health professionals' that the scale had 'good' content validity. This suggests that, based on expert opinion, the CTS-Psy appears to measure what it was designed to. However, details on the method used to canvas opinion and the experience of the decision makers were not provided.

3.5.4 Construct validity

Aspects of construct validity reported by studies included convergent validity, sensitivity to change and factor analysis. Convergent validity was assessed in the CBTS-CYP (Stallard *et al.*, 2014) and CBAY-C (McLeod *et al.,* 2018). The CBTS-CYP was compared with the CTS-R (Blackburn *et al.,* 2001) and the measures were highly correlated (see Table 2). This was expected as the CBTS-CYP was designed to cover all aspects of the CTS-R (see Appendix D). The CBAY-C was compared with measures of adherence, therapeutic alliance and client involvement, and correlations ranged from small to large.

Sensitivity to change was explored in six studies. The CTS-R (Blackburn *et al.,* 2001), CTS-Psy (Haddock *et al.,* 2001) and CBTS-CYP (Stallard *et al.,* 2014) were sensitive to changes in trainee competence in CBT. The CTACS (Barber *et al.,* 2003) and the YACS (Carroll *et al.,* 2000) were found to have good criterion-related validity. They were used to rate

therapists in CBT and other therapies (e.g., supportive-expressive therapy), and could accurately distinguish CBT sessions from non-CBT. The CTS (Vallis *et al.,* 1986) was found to be sensitive to the quality of therapy protocol administration. Using discriminant function analysis, it correctly classified this in 84.9% of cases. Finally, the CBTS-CYP (Stallard *et al.,* 2014) was compared with the CTS-R (Blackburn *et al.,* 2001) and was found to discriminate between poor- and high-quality CBT equally well.

Factor analysis data was available for three of the measures. The majority of variance on the CAS-CBT (Bjaastad *et al.,* 2016) and the CTS (Vallis *et al.,* 1986) was explained by two factors. The factors identified for the CAS-CBT matched hypotheses about which constructs were measured by the scale but highlighted an overlap between adherence and competence. The results for the CTS (see Table 2) suggested that the overall scale is effective at measuring CT competence, but that individual item scores should not be considered when deciding therapist competence. The majority of variance on the ITIS (Boyle *et al.,* 2019) was explained by one factor, similar to results found for the CTS (Weck *et al.,* 2010) on which the ITIS was based.

4 Discussion

This review synthesised 13 studies to investigate the methods used to develop and psychometrically evaluate therapist competence scales for CBT. It was hoped that recommendations for clinical practice and research would be made based on the findings. Quality appraisal (see Appendix F) indicated that 11 studies fell within the 'good-excellent' range and could therefore be considered with similar confidence. One paper was assessed as 'fair' and one was 'poor'. As discussed, these poorer quality papers did not impact on the findings or conclusions of the review.

4.1 The development and psychometric evaluation of CBT competence scales

Studies described 11 therapist competence scales developed for a range of CBT interventions and patient groups, with five including a measure of adherence given its overlap with competence (Barber *et al.,* 2003). The review found that there was little consensus as to how scales should be developed. For example, some relied on author opinion and others used pre-existing competence scales and consensus building processes. In addition, not all studies adequately described the development process,

making it difficult to conclude which methods might produce more valid and reliable scales. The approaches used to psychometrically evaluate the scales also varied. For example, some used students to rate therapy material and others used CBT experts. Some studies explored a number of psychometric properties whereas others focused on just one or two properties. This variation in aspects of reliability and validity measurement highlights the lack of consensus amongst researchers when developing competence scales. These findings were similar to Barlow and Brown (2019), who highlighted a lack of agreement on methods of developing and evaluating therapist competence scales for interpersonal, dynamic and relational models.

4.2 The reliability and validity of CBT competence scales

The psychometric properties of scales varied independently of the development process in that a comprehensive development strategy did not always lead to stronger reliability and validity. For example, the less systematically developed CTCS-SP had good inter-rater reliability and internal consistency, whereas the comprehensively developed CTACS and CBAY-C were only moderately reliable. This might be because many CBT competence scales are based on existing measures (e.g., CTS; Young & Beck, 1980). Therefore, competence scales might be assessed as reliable if they are not distinct enough from the validated scale on which they are based (Stallard *et al.,* 2014). However, CBT is a well-established intervention with a large evidence base and a similar structure regardless of the specific model (e.g., NICE, 2018, 2019). As CBT models can be similar in their delivery, and in the skills and techniques a competent therapist should display, it could be argued that similarity between competence scales (e.g., those based on existing measures) may not be problematic.

Some scales had stronger reliability and/or validity in some areas but were weaker in others. For example, the CTS-R, used on many CBT training courses (e.g., Reichelt *et al.,* 2003), has excellent internal consistency but moderate inter-rater reliability. These discrepancies might be problematic, as those selecting which competence measure to use might have to prioritise one aspect of reliability or validity over another. However, many types of CBT, such as CBT for psychosis, only have one published scale with limited psychometric evaluation. Inter-rater reliability was the only property explored by all of the studies, suggesting its importance during competence scale development. It is

essential to know the extent to which different raters agree on competence scores to ensure that they are adhering to similar standards of scale use. Therefore, good inter-rater reliability should be prioritised when selecting scales, especially given that these scales are used by different clinicians in different settings.

4.3 Factors impacting the reliability and validity of CBT competence scales

Variation in scale reliability and validity may be attributed to scale, rater and methodological factors. As also noted by Barlow and Brown (2019), most measures used 7-point Likert-scales to rate competence but none described why this was chosen. It may be because the CTS (Young & Beck, 1980) used a 7-point scale and this was used as a basis for several further scales. Having more response options might reduce inter-rater reliability due to the opportunity for more variance. However, this review found variation in psychometric properties regardless of the number of response options.

Consideration should also be given to variation in the raters used in the studies, such as their varying levels of experience in CBT. When using scales to measure competence, raters make inferences based on their own experiences (Kogan *et al.,* 2011). In line with this, Roth *et al.* (2019) suggested that low levels of agreement may be due to a lack of training in the use of the measures. Therefore, rater experience and training, as well as specific scale characteristics, should be considered when developing competence scales.

Methodological factors may have impacted the reliability and validity of the reviewed scales. For example, three scales were evaluated using audio tapes which, obviously, cannot capture non-verbal communications that are often considered essential for the competent delivery of therapy (e.g., CFT-TCRS; Horwood *et al.,* 2019). Non-verbal communication can assist the development of therapeutic alliance (Dowell & Berman, 2013) which has been included in competence frameworks and is measured by competence scales (e.g., CTACS). Conversely, it might be important to have scales which can be reliably used with audio material when assessing competence of therapists providing interventions over the telephone, for example.

4.4 Strengths and limitations

The current review had several strengths. It was systematic and the search process meant that all available published literature was included. The review has drawn attention to the variability in methods used to develop and evaluate therapist competence scales. However, the review had several limitations. First, the quality and data extraction tool was based on published tools and followed a similar development process to a similar review (Barlow & Brown, 2019). However, the papers included in this review were not cross-checked. It may have been useful to have a percentage of the papers assessed 'independently' to ensure that the author was drawing out useful information and correctly assessing quality. Second, a lack of homogeneity in approaches to developing and assessing the reliability and validity of scales led to difficulties in synthesis and thus the use of a narrative approach.

4.5 Clinical implications and future research

Despite limitations, this review highlighted some important issues. Firstly, it showed a lack of consensus in competence scale development and psychometric evaluation which may be linked to the variation in the reliability and validity of CBT competence scales. Therefore, care should be taken when selecting which scales to use. Future researchers should thoroughly describe which methods they used, and why, to provide clarity for other scale developers. If there is no reliable and valid measure of therapist competence, then intervention quality cannot be formally measured or assured. Therefore, future research should focus on developing competence scales for all NICE recommended CBT models (e.g., trauma-focused CBT for PTSD; NICE, 2018). However, basing new measures on scales which may not have been comprehensively developed or examined may continue to produce tools not suitable for widespread use.

Most studies highlighted the need for more comprehensive psychometric evaluation. Therefore, future competence scale developers should assess as many psychometric properties as is feasible to increase reliability and validity, and thus usability. Similar to suggestions made by Barlow and Brown (2019), future studies might focus on recruiting larger, more representative samples to establish psychometric properties such as factor structure. Furthermore, the variation in therapist experience and training noted in this

review and the review by Muse and McManus (2013) has highlighted the importance of scale implementation protocols. These might provide guidance on the level of therapist training and experience required to use the scale (Muse & McManus, 2013), and therefore clarity around who should complete the competence assessment.

4.5.1 Assessing CBT competence

Several CBT competence scales have been recently published, perhaps partly due to suggestions from previous reviews that scales should be refined and developed (Muse & McManus, 2013) and partly due to the publication of Roth and Pilling's (2007) CBT competence framework which described the practices and skills required during all variations of CBT. However, CBT and what constitutes its competent delivery has changed over time. A focus on transdiagnostic and transtheoretical identification of processes of change in psychotherapy (e.g., Hofmann & Hayes, 2018) has impacted CBT practice (Boyle *et al.,* 2019). This has led to revisions of 'classic' CBT and for CBT to encompass an increasing variety of interventions (Boyle *et al.,* 2019). These developments, along with difficulties reliably measuring CBT competence, suggest that competence scales alone might not be sufficient to assess this increasing diversity. In line with this, Muse and McManus (2013) suggested that a multi-method approach might be preferable, as no single validated measure of CBT competence has been agreed.

5 Conclusion

In conclusion, this narrative synthesis highlighted the need for further research into the development and psychometric evaluation of new and existing therapist competence scales for CBT. Future scale developers might focus on using comprehensive development and psychometric evaluation procedures when developing new or existing scales. Despite methodological differences and the need for further research, some of the included scales in the present review can be used reliably to assess therapist competence in CBT and many are currently used in therapist training. However, assessors and therapist training courses using competence scales should be aware of their limitations and other factors impacting on competence ratings and might consider other ways in which therapist competence can be evaluated.

References

Barber, J.P., & Crits-Cristoph, P. (1996). Development of a therapist adherence/competence rating scale for supportive-expressive dynamic psychotherapy: A preliminary report. *Psychotherapy Research, 6*(2), 82-94.

[1]*Barber, J. P., Liese, B. S., & Abrams, M. J. (2003). Development of the cognitive therapy adherence and competence scale. *Psychotherapy Research, 13*(2), 205-221.

Barber, J. P., Krakauer, I., Calvo, N. et al. (1997). Measuring adherence and competence of dynamic therapists in the treatment of cocaine dependence. *The Journal of Psychotherapy Practice and Research, 6*(1), 12-24).

Barber, J. P., Sharpless, B. A., Klosterman, S., & McManus K. S. (2007). Assessing intervention competence and its relation to therapy outcome: A selected review derived from the outcome literature. *Professional Psychology: Research and Practice, 38,* 493-500.

Barlow, I., & Brown, R. J. (2019). A systematic review of measures of therapist competence in psychodynamic, interpersonal, and/or relational models. *Psychology and Psychotherapy: Theory, Research and Practice.* DOI:10.1111/papt.12231

Bennett, D., & Parry, G. (2004). A measure of psychotherapeutic competence derived from cognitive analytic therapy. *Psychotherapy Research, 14*(2), 176-192.

*Bjaastad, J. F., Haughland, B. S. M., Fjermestad, K. W. et al. (2016). Competence and adherence scale for cognitive behavioural therapy (CAS-CBT) for Anxiety Disorders in Youth: Psychometric properties. *Psychological Assessment, 28*(8), 908-916.

[1] * studies selected for review

*Blackburn, I., James, I. A., Milne, D. L. et at. (2001). The revised cognitive therapy scale (CTS-R): Psychometric properties. *Behavioural and Cognitive Psychotherapy, 29*(4), 431-466.

* Boyle, K., Deisenhofer, A. K., Rubel, J. A. et al. (2019). Assessing treatment integrity in personalised CBT: The inventory of therapeutic interventions and skills. *Cognitive Behaviour Therapy.* DOI:10.1080/16506073.2019.1625945

*Carroll, K. M., Nich, C., Sifry, R. L. et al. (2000). A general system for evaluating therapist adherence and competence in psychotherapy research in the addictions. *Drug and Alcohol Dependence, 57*(3), 225-238.

Clark, D. M., Consbruch, K. von, Hinrichs, J., & Stangier, U. (2007). *Cognitive Therapy of Social Phobia: Checklist of therapist competency.* Unpublished manuscript. London: Kings College London.

*Consbruch, K. von, Clark, D. M., & Stangier, U. (2012). Assessing therapeutic competence in cognitive therapy for social phobia: Psychometric properties of the cognitive therapy competence scale for social phobia (CTCS-SP). *Behavioural and Cognitive Psychotherapy, 40*(2), 149-161.

*Dobson, K. S., Shaw, B. F., & Vallis, T. M. (1985). Reliability of a measure of the quality of cognitive therapy. *British Journal of Clinical Psychology, 24*(4), 295-300.

Dowell, N. M., & Berman, J. S. (2013). Therapist nonverbal behaviour and perceptions of empathy, alliance, and treatment credibility. *Journal of Psychotherapy Integration, 23*(2), 158–165.

Downs, S. H., & Black, N. (1998). The feasibility of creating a checklist for the assessment of the methodological quality both of randomised and non-randomised studies of health care interventions. *Journal of Epidemiology and Community Health, 52*(6), 377-384.

Fairburn, C. G., & Cooper, Z. (2011). Therapist competence, therapy quality, and the therapist training. *Behaviour Research and Therapy, 49*, 373-378.

George, D., & Mallery, P. (2003). *SPSS for Windows step by step: A simple guide and reference.* 11.0 update (4th ed.). Boston: Allyn & Bacon.

Gilbert, P. (2009). Introducing compassion-focused therapy. *Advances in Psychiatric Treatment, 15*(3), 199-208.

*Haddock, G., Devane, S., Bradshaw, T. et al. (2001). An investigation into the psychometric properties of the cognitive therapy scale for psychosis (CTS-Psy). *Behavioural and Cognitive Psychotherapy, 29*(2), 221-233.

Hazell, C. M., Hayward, M., Cavanagh, K., & Strauss, C. (2016). A systematic review and meta-analysis of low intensity CBT for psychosis. *Clinical Psychology Review, 45,* 183-192.

Hofmann, S. C., & Hayes, S. G. (Eds.). (2018). *Process-based CBT: The science and core clinical competencies of cognitive behavioral therapy*. Oakland, CA: New Harbinger Publications.

Horwood, V., Allan, S., Goss, K., & Gilbert, P. (2019). The development of the Compassion-Focused Therapy Therapist Competence Scale. *Psychology and Psychotherapy: Theory, Research and Practice.* DOI:10.1111/papt.12230

Kogan, J. R., Conforti, L., Bernabeo, E. et al. (2011). Opening the black box of clinical skills assessment via observation: A conceptual model. *Medical Education, 45,* 1048-1060.

Koo, T. K., & Li, M. Y. (2016). A guideline for selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine, 15*(2), 155-163.

Legendre, P. (2010). Coefficient of concordance. *Encyclopaedia of Research Design. 1,* 164-169.

Linstone, H. A., 7 Turoff, M. (Eds.) (1975). *The Delphi method: Techniques and applications (Vol. 29)*. Reading, MA: Addison-Wesley.

Lucas, N. P., Macaskill, P., Irwig, L., & Bogduk, N. (2010). The development of a quality appraisal tool for studies of diagnostic reliability (QAREL). *Journal of Clinical Epidemiology, 63*(8), 854-861.

* McLeod, B. D., Southam-Gerow, M. A., & Rodríguez, A. et al. (2018). Development and initial psychometrics for a therapist competence instrument for CBT for youth anxiety. *Journal of Clinical Child and Adolescent Psychology, 47*(1), 47-60.

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. Psychological methods, 1(1), 30.

McLeod, B. D., Southam-Gerow, M. A., Tully, C.B. et al. (2013). Making a case for treatment integrity as a psychosocial treatment quality indicator for youth mental health care. *Clinical Psychology: Science and Practice, 20*(1), 14-32.

Muse, K., & McManus, F. (2013). A systematic review of methods for assessing competence in cognitive-behavioural therapy. *Clinical Psychology Review, 33*(3), 484-499.

National Institute for Health and Care Excellence. (2018). *Depression in adults: Recognition and management.* Retrieved 10/10/2019 from https://www.nice.org.uk/guidance/cg178

National Institute for Health and Care Excellence. (2019). Generalised anxiety disorder and panic disorder in adults: Management. Retrieved 10/10/2019 from https://www.nice.org.uk/guidance/cg113

National Institute for Health and Care Excellence. (2018). *Post-traumatic stress disorder.* Retrieved 27/10/19 from https://www.nice.org.uk/guidance/ng116

National Institute for Health and Care Excellence. (2014). *Psychosis and schizophrenia in adults: Prevention and management.* Retrieved 10/10/2019 from https://www.nice.org.uk/guidance/cg178

O'Connor, S. R., Tully, M. A., Ryan, B., Bradley, J. M., Baxter, G. D. & McDonough, S. M. (2015). Failure of a numerical quality assessment scale to identify potential risk of bias in a systematic review: a comparison study. *BioMed Central, 8*, 224.

Perepletchikova, F. (2011). On the topic of treatment integrity. *Clinical Psychology: Science and Practice, 18*(2), 148-153.

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

Reichelt, F. K., James, I. A., & Blackburn, I. (2003). Impact of training on rating competence in cognitive therapy. *Journal of Behaviour Therapy and Experimental Psychiatry, 34*(2), 87-99.

*Roth, A. D. (2016). A new scale for the assessment of competences in cognitive and behavioural therapy. *Behavioural and Cognitive Psychotherapy, 44,* 620-624.

*Roth, A. D., Myles-Hooton, P., & Branson, A. (2019). Judging clinical competence using structured observation tools: A cautionary tale. *Behavioural and Cognitive Psychotherapy, 47,* 736-744.

Roth, A. D., & Pilling, S. (2007). *The competences required to deliver effective cognitive and behavioural therapy for people with depression and anxiety disorders*. London: Department of Health.

Ryle, A. (2005). Cognitive Analytic Therapy. In J.C. Norcross & M. R. Goldfried (Eds.). *Handbook of psychotherapy integration* (2nd edn, pp.196-217). New York: Oxford University Press.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86,* 420-428.

*Stallard, P., Myles, P., & Branson, A. (2014). The cognitive behaviour therapy scale for children and young people (CBTS-CYP): Development and psychometric properties. *Behavioural and Cognitive Psychotherapy, 42*(3), 269-282.

Stewart, R. E., & Chambless, D. L. (2009). Cognitive-behavioural therapy for adult anxiety disorders in clinical practice: A meta-analysis of effectiveness studies. *Journal of Consulting and Clinical Psychology, 77*(4), 595-606.

*Vallis, T. M., Shaw, B. F., Dobson, K. S. (1986). The cognitive therapy scale: Psychometric properties. *Journal of Consulting and Clinical Psychology, 54*(3), 381-385.

Webb, C. A., DeRubeis, R. J., & Barber, J. P. (2010). Therapist adherence/competence and treatment outcome: A meta-analytic review. *Journal of Consulting and Clinical Psychology, 78*(2), 200-211.

Weck, F., Hautzinger, M., Heidenreich, T., & Stangier, U. (2010). Erfassung psychotherapeu- tischer Kompetenzen. *Zeitschrift Für Klinische Psychologie Und Psychotherapie, 39*(4), 244–250.

Yeaton, W., & Sechrest, L. (1981). Critical dimensions in the choice and maintenance of successful treatments: Strength, integrity, and effectiveness. *Journal of Consulting and Clinical Psychology, 49,* 156-167.

Young, J., & Beck, A. T. (1980). *The Development of the Cognitive Therapy Scale.* Unpublished manuscript, Center for Cognitive Therapy, Philadelphia, PA.

Part 2: Empirical Research Project

An initial exploration of the psychometric properties and expert opinion of the Compassion Focused Therapy Therapist Rating Scale

Abstract

Objectives

Compassion-focused therapy (CFT) is an effective treatment for a range of psychological difficulties, but therapists' treatment fidelity in CFT has not been researched (Leaviss & Uttley, 2015). The Compassion Focused Therapy Therapist Rating Scale (CFT-TCRS; Horwood *et al.,* 2019) was developed to assess competence of therapists and to address issues of treatment fidelity. The aim of the study was to assess the inter-rater reliability of the CFT-TCRS using a novel methodology and to elicit feedback on the individual items and the scale as a whole.

Method

CFT experts were recruited to watch clips of simulated CFT sessions showing varying levels of therapist competence. They used the CFT-TCRS to make competence judgments before being interviewed for feedback on the individual items and the scale overall. Results were statistically analysed to determine inter-rater reliability, and content analysis was used to analyse the qualitative data.

Results

Inter-rater reliability between the CFT experts was found to be 'good'. However, the reliability of a single rater and the reliability between participants' and the researchers' 'official' rating was 'poor'. Content analysis revealed *general* and *specific* changes for each item including clearer scoring guidance, less content and more behavioural anchors. Feedback on the CFT-TCRS overall revealed the *benefits* and *difficulties* of using the scale and experts suggested changes to the *content* and *structure* of the scale. For example, it could be made clearer and items could be split into sub-items.

Conclusion

The inter-rater reliability of the CFT-TCRS was good but might be further improved by revising the scale in light of the expert user feedback. However, caution should be taken when interpreting the results due to several methodological problems with the study. The findings suggested that the CFT-TCRS may eventually prove to be a useful tool in clinical practice, research and therapist training following recommendations.

1 Introduction

1.1 Compassion Focused Therapy

Compassion-focused therapy (CFT) is a relatively new psychotherapeutic approach drawing on developmental, evolutionary, neurological, social and Buddhist psychology (Gilbert, 2009). CFT is a multi-modal intervention focused on nurturing a more compassionate inner voice (Leaviss & Uttley, 2015). It proposes that humans have three affect regulation systems known as the threat-protective system, the drive, seeking and reward system, and the contentment-soothing system (Gilbert, 2014). CFT suggests that imbalance between the systems can lead to biopsychosocial problems (Liddell *et al.,* 2017). CFT's principle aim is to enhance someone's ability to regulate their three systems and promotes the development of affiliative feelings which mediate threat responses by enhancing calmness.

A review conducted by Leaviss and Uttley (2015) explored the effectiveness of CFT. They found that overall, CFT can be an effective psychotherapeutic intervention across a range of mental health difficulties. For example, CFT was found to improve experiences of depression and anxiety (e.g., Gilbert & Proctor, 2006), psychosis (e.g., Braehler *et al.,* 2013; Laithwaite *et al.,* 2009), eating disorders (e.g., Gale *et al.,* 2014; Kelly *et al.,* 2017), trauma (e.g., Beaumont *et al.,* 2016) and 'personality disorders' (e.g., Lucre & Corten, 2013). Although further study is required, especially into aspects of treatment fidelity, CFT is becoming increasingly more accepted as an intervention for people with mental health difficulties (Leaviss & Uttley, 2015) and several CFT training courses are offered in the UK and internationally.

1.2 Therapist competence

For therapy to be considered 'evidence-based' it must demonstrate 'treatment fidelity' which is "the degree to which the intervention is implemented as intended" (Vermilyea *et al.*, 1984, p.2). Treatment fidelity comprises 'treatment differentiation', 'therapist adherence' and 'therapist competence' (Perepletchikova *et al.,* 2007). Treatment differentiation refers to whether a therapy is distinct (Southam-Gerow & McLeod, 2013) and adherence refers to whether techniques described in the therapy manual are demonstrated (Webb *et al.,* 2010). Therapist competence is defined as "the extent to

which a therapist has the knowledge and skill required to deliver a treatment to the standard needed for it to achieve its expected effects" (Fairburn & Cooper, 2011, p.374).

It is essential to demonstrate that a therapy can be competently delivered to determine its fidelity (Waltz *et al.,* 1993). Furthermore, competence must be assessed as therapists have a duty to provide skilled intervention (Fairburn & Cooper, 2011). For an intervention to be well executed the therapist must have knowledge of the treatment, the capability to deliver it, the willingness and drive to provide the treatment as intended, and the ability to provide the treatment to a variety of patients (Barber *et al.,* 2007). Whist some research on cognitive therapy found that higher levels of therapist competence better predicted patient outcomes (e.g. Shaw *et al.,* 1999), studies on this relationship have not produced consistent results (e.g., Berman & Norton, 1985; Hattie *et al.,* 1984).

The skills required to competently deliver therapy have been identified for a variety of psychological interventions, such as cognitive behavioural therapy (CBT), cognitive analytic therapy (CAT) and psychodynamic therapy (e.g., Bennett & Parry, 2004; Lemma *et al.,* 2008; Roth & Pilling, 2007). However, methods of assessing competence have not been extensively researched (Sharpless & Barber, 2009) and the standard of therapy provision is rarely evaluated (Perepletchikova *et al.,* 2007). Nonetheless, understanding and being able to measure competence can aid therapist training and can maximise the provision of effective, evidence-based training and treatment (McHugh & Barlow, 2010). This can lead to improved outcomes (Boyle *et al.,* 2019) and minimised risk for patients (Bennett & Parry, 2004).

1.3 Therapist competence scales

Several methods have been used to assess therapeutic competence. These have included multiple-choice questionnaires, essays, role plays and therapist competence scales (Muse & McManus, 2013). Of these, competence scales have been the most commonly used measure of therapeutic skill (Muse & McManus, 2013). Competence scales have been developed for several therapeutic modalities such as cognitive therapy (Blackburn *et al.,* 2001), supportive-expressive therapy (Barber *et al.,* 1997), CAT (Bennett & Parry, 2004) and CBT for children and young people (Stallard *et al.*, 2014).

Most competence scales are 'observer rated', meaning that a therapist will be watched and assigned a competence score. These scores can provide useful feedback to therapists on areas for improvement, and scales can be used as a supervision tool and to promote self-reflection.

It is important that competence scales are well developed and evaluated in order that their reliability and validity can be established. However, there is little consensus about the methods of competence scale development and psychometric evaluation that are most appropriate in order to produce valid and reliable scales. For example, some scales are subject to comprehensive and multi-faceted development procedures (e.g., Barber *et al.,* 2003) whereas others have, more simply, been based on existing competence scales (e.g., von Consbruch *et al.,* 2012). Furthermore, there is little consistency in the psychometric properties assessed and reported by scale developers, aside from inter-rater reliability which is widely explored.

1.4 Therapist competence in CFT and the Compassion Focused Therapy Therapist Competence Rating Scale (CFT-TCRS)

Liddell *et al.* (2017) generated a CFT competence framework (CFT-CF) comprising 25 competencies falling within six categories; creating safeness, meta-skills, non-phase-specific skills, phase-specific skills, knowledge/understanding, and use of supervision. Some competencies were specific to CFT while others were more general competencies required by many forms of therapy (e.g., socratic questioning and summarising).

Subsequently, Horwood *et al*. (2019) developed the Compassion Focused Therapy Therapist Competence Rating Scale (CFT-TCRS) (see Appendix G). The CFT-TCRS was based on the competency research by Liddell *et al*. (2017) with additional suggestions from an informal CFT assessment guide by Gilbert *et al*. (2012, unpub). It was developed using the Delphi method and involved multiple meetings and online surveys with 11 CFT experts to define and operationalise the scale items (competences) and scoring. It was designed to be used by experts to rate trainee CFT therapists.

The CFT-TCRS is divided into two subscales; unique CFT competencies and generic microskills. Unique CFT competencies are theory-driven and are required to effectively carry out the intervention. They are not expected to be observed during every session

or every stage of the therapy. The generic microskills are essential therapeutic skills required to deliver psychological intervention and are expected to be observed during every session. However, the scale is yet to have its psychometric properties explored and may need further revision before it can be widely disseminated and utilised.

1.5 Rationale and aims

The CFT-TCRS (Horwood *et al.,* 2019) was designed to address issues of therapists' treatment fidelity. The psychometric properties of the scale must be established for it to be reliably used in research, evaluation, training and audits (Koo & Li, 2016). Therefore, similar to previous studies assessing psychometric properties of therapist competence scales (e. g. Barber *et al.,* 1997; Bennett & Parry, 2004; Blackburn *et al*., 2001; Stallard *et al.,* 2014), the current research aimed to explore whether the CFT-TCRS could be used reliably by experts to rate therapist competence. Given that generic microskills have been included in many measures of therapist competence, and because the CFT-TCRS is the first competence measure for CFT, it was decided that the unique CFT competences would be the focus of the research.

The present study also developed a novel methodology for assessing inter-rater reliability of competence scales using simulated therapy videos. It was hoped that creating videos would allow the researchers to control the level of competence displayed and make a reliable judgment as to the level of competence displayed by the therapist in each video for the purpose of statistical analysis.

Previous research into therapist competence scales has not typically sought the opinions of those attempting to use the scale for the first time. Therefore, the present research also aimed to capture qualitative information from experts about changes they would suggest to specific items in the scale, as well as eliciting feedback about the scale overall. It was hoped that this information might help inform future improvements to specific items and to the scale overall [2].

[2] Qualitative information on the decision making behind the allocation of competence scores on the CFT-TCRS was also collected. However, the word limit of the project precluded the inclusion of the analysis (which is available on request).

Therefore, the aims of this research were:

1. To develop video materials of simulated CFT sessions for use during the assessment of the inter-rater reliability of the CFT-TCRS.
2. To use this novel methodology to assess whether experts can reliably use the CFT-TCRS to make decisions on the competence of therapy delivery and to ascertain the scale's inter-rater reliability.
3. To identify changes which could be made to each of the items of the CFT-TCRS included in this study and make recommendations for improvement.
4. To explore the overall feedback for the CFT-TCRS given by the participants and make recommendations for improvement.

2 Method

2.1 Design

A mixed methods design was used to address the aims of the study. The analysis of quantitative information was used to provide information about whether CFT experts could use the CFT-TCRS to make reliable competence judgments, and it was hoped that the analysis of qualitative information would be useful for future improvements to the recently developed scale. This study was approved by the University of Leicester ethics board (see Appendix H).

A common approach to establishing reliability is for experts to use competence scales to rate recorded therapy sessions (Bennett & Parry, 2004). Using real session footage can be problematic in terms of quality control and ethics so it was decided that videos of simulated CFT sessions would be used to determine inter-rater reliability. Inter-rater reliability was addressed using a 'balanced' design, meaning that each participant viewed and rated all the available material and was compared with all other raters. The dependent variable was therefore the 'judgment of competence' made by expert raters.

2.2 Measures and materials

2.2.1 The CFT-TCRS

As previously described, the CFT-TCRS is a 23-item scale for measuring therapist competence in CFT (Appendix G). The first 14-items measure CFT-specific skills and the final 9-items measure generic therapeutic skills which would be expected during all therapy sessions. Each item is titled with the competence it is measuring (e.g., Item 1: Psychoeducation), followed by a short description of the competence and how skilful enactment might look. Each item also includes 'points to consider when scoring' which indicate what the rater should look out for (e.g., 'the therapist demonstrates skilfulness in the methods used'). Examples of specific topics and skills are also provided in this section. These guidelines for scoring are followed by the rating scale.

Items on the CFT-TCRS are measured on a 5-point Likert scale, which provides a score between 0-4. A score of '0' is used when the competence is 'absent or inappropriate', and a score of '4' is assigned for 'skilful enactment'. Scores of 1-3 are not specifically described in the current version of the CFT-TCRS. At either end of the Likert scale for each item, 'behavioural anchors' are provided to describe what might be seen during 'less competent' and 'more competent' delivery of the specific skill. An 'unable to rate' option is also provided for when a competence is not present, as it would be unlikely to observe all competencies during a single CFT session.

2.2.2 Video material to be rated

Ten video clips of simulated CFT were designed and filmed by the principal researcher, their supervisors and a senior expert in the CFT community. The senior expert played the 'therapist' as he was able to comment on and demonstrate both competent and non-competent CFT. The 'patient' was played by a professional actor who had experience creating CFT training videos and working with the Compassionate Mind Foundation. On the day of filming researchers discussed what should be included in the clips. These ranged from 4-12 minutes long and were designed to demonstrate a segment of a therapy session. The videos were edited by the principal researcher and were uploaded to YouTube as 'unlisted', meaning they were not available to the public and could only be viewed via a private link.

The videos covered six of the most important competencies (as decided by a senior expert in CFT and another CFT expert) due to the exploratory nature of the research and novel methodology. Five videos showed the six competencies delivered competently, and the other five displayed less-competent enactment (one of the videos encompassed two competencies). The competencies included in the videos were: 'psychoeducation', 'recognising motives and emotions', 'understanding the relationship between the three systems' (threat, drive and soothe), 'building motivation', 'functional analysis', and 'fears, blocks and resistances'.

2.3 Participants

Data was collected from nine experts in CFT who were recruited using purposive sampling. This was in line with Barber and Crits-Cristoph (1996) who argued that experts should assess whether a treatment was delivered competently, especially during scale development. Clinicians were considered expert if they had significant knowledge, training, experience and supervisory practice in CFT. The specific inclusion criteria were similar to those applied by Horwood *et al.,* (2019) who used CFT experts during the development of the CFT-TCRS. Thus, participants were required to: have been trained or supervised by a member of the Compassionate Mind Foundation, have a minimum of two years of CFT clinical practice, and be experienced in supervising or training CFT clinicians or trainees. As CFT is a contemporary therapy, the pool of potential participants known to the researchers was relatively small (*n*=27). They were identified and approached via email and were provided with an information sheet (Appendix I) which included an explanation of the study, what participation would involve and the potential risks and benefits of taking part. Participants could then volunteer to take part if they were able to.

2.4 Procedure

Once experts had agreed to participate, they were contacted to arrange a date and time for study completion. Depending on the location of the participants, the consent form (Appendix J) and 'data collection pack' were sent either via post (UK participants) or email (EU and international participants) by the principal researcher. The data collection pack (Appendix K) included standardised instructions, relevant scale items and 'notes'

pages. Once the materials were received by the participants, the secure links to the 10 video clips, in randomised order, were sent by email to the participant. The data collection was completed in one session with participants first completing the video rating, immediately followed by a short semi-structured interview. Participants were asked to either post or email their completed consent form and ratings back to the principal researcher.

2.4.1 Video rating

The first part of the study was the video rating, which was estimated to take three hours. For each of the 10 videos, participants were instructed to familiarise themselves with the relevant scale item/s, paying attention to the 'points to consider when scoring' and the behavioural anchors. They were directed to complete the note pages (one for each item) whilst watching the videos, as these would be used during the semi-structured interview. The note pages included questions designed to prompt their thinking around their competence decision making, whether they would change anything about the item and their general feedback on the scale. Finally, they were asked to provide a competence rating for the therapist on the scale item/s associated with each video.

2.4.2 Semi-structured interview

Immediately following the video rating, the principal researcher completed a short semi-structured interview with the rater which was estimated to take 30 minutes. These were conducted over the telephone, or via video-link and were directed by an interview guide (Appendix L). The videos were discussed one by one. Firstly, the competence rating and the decision making behind it was discussed, followed by questions around what changes the rater would make to the items. Once all 10 videos had been discussed, the rater was asked about their overall impression of the scale. The interviews were recorded onto an encrypted audio recording device and were transcribed into word processing documents by the principal researcher.

2.5 Analysis

Quantitative data, in the form of 12 competence ratings per participant, was collected to assess inter-rater reliability. Qualitative data from the interviews was used to collect

expert opinion on the need for changes to specific items and to the scale overall, and to explore decision making.

2.5.1 Quantitative analysis: Inter-rater reliability

The inter-rater reliability of a measure is the degree of agreement between raters using that measure. In the case of the CFT-TCRS, it is important to determine inter-rater reliability to highlight the extent to which variation in competence ratings is representative of differences in the therapist's skill, rather than differences between the raters. To establish inter-rater reliability between the participants, and between the participants' and the researchers' 'official' competence rating, the intraclass correlation coefficient (ICC) was calculated (Shrout & Fleiss, 1979).

ICC is an acceptable measure of reliability (e.g., Koo & Li, 2016) and has 10 different forms (McGraw & Wong, 1996) which vary depending on the model, type and definition (see Appendix M). It is essential that the correct form of ICC is selected, but it has been suggested that absolute agreement should be prioritised when considering inter-rater reliability, as focusing on the consistency has similar problems to Pearson's which was historically used for assessing inter-rater reliability (Koo & Li, 2016). Absolute agreement considers whether different raters give the same competence ratings to a therapist, and therefore absolute agreement was selected for this research. A secondary analysis focusing on consistency between raters was completed and presented in Appendix N.

Statistical analyses were calculated using IBM SPSS Statistics 25. First, ICC (2, *k*) (Shrout & Fleiss, 1979) was calculated. This is a two-way random effects model which considers the rater as a random effect and *k* as the number of raters (*k* =9). It uses the mean competence rating of participants to determine the inter-rater reliability. Second, ICC (2, 1) (Shrout & Fleiss, 1979) was calculated to determine the reliability of a single rater. This was calculated because in practice competence scores are usually assigned by a single clinician. Finally, the reliability between the 'official' rating assigned by the researchers and the mean of participants' competence scores for each item was calculated using ICC (2, *k*) where *k* is 2 and represents the participants and the researchers. Researchers assigned a rating of 'competent' or 'less competent' to videos, so participants' mean ratings were transformed into ratings of 'competent' (a mean

score of 2 or above) or 'less competent' (a mean score of 2 or below). The confidence intervals of the ICCs were interpreted using Koo & Li's (2016) guidelines which propose that a reliability coefficient of <.5 is 'poor', .5-.74 is 'moderate', .75-.89 is 'good' and >.9 is 'excellent'.

2.5.2 Qualitative analysis (content analysis)

The qualitative data was analysed using content analysis, which is defined as 'the subjective interpretation of the content of text data through the systematic classification process of coding and identifying themes or patterns' (Hsieh & Shannon, 2005). Conventionally, content analysis follows specific steps which include the researcher immersing themselves in the data and assigning codes to patterns of words or phrases. This is typically followed by grouping codes into categories based on how the codes are linked, and organisation of these categories into hierarchical clusters if appropriate. Finally, each of the categories, subcategories and codes are defined and examples are taken from the qualitative data which represents each of the categories (Hsieh & Shannon, 2005).

In the present study, the content analysis process was deductive as the principal researcher coded the data based on the specific questions which were being asked about the CFT-TCRS. Please see Appendix O for an example of the initial coding process, and Appendix P for higher-level coding. The data was semantically interpreted, in that it was taken at 'face value' and the latent meaning was not explored. Three distinct analyses were carried out on the qualitative data. Firstly, data was coded for item-by-item changes suggested by participants. This was followed by coding of the comments on the scale overall. It was hoped that this would allow procedure specific recommendations for improving the scale. Finally, the data was coded to capture aspects of participant decision-making that informed the allocation of their competence rating.

For each of the above three analyses, initial codes were grouped into categories. Codes were then counted to determine how many participants made comments associated with each code and subsequent category, which was transformed into a percentage. Finally, direct quotes were taken from the data to illustrate each of the categories and

their associated codes. (See Appendix Q for details of each stage of the content analysis.) Issues of quality assurance for content analysis are addressed through a statement of the principal researcher's epistemological position (Appendix R) and through a discussion in Appendix S. The principal researcher also kept a reflexive diary throughout the research process (see Appendix T for an extract.)

3 Results

The inter-rater reliability of the CFT-TCRS is presented first, for agreement between the participants (both for the average scores and a single rater), and agreement between participants and researchers. This is followed by the results of the content analysis which explored changes to items suggested by the participants and feedback on the scale 'overall'[3]. Results of an initial internal consistency analysis can be found in Appendix U.

3.1 Inter-rater reliability

To establish inter-rater reliability, each participant watched the ten simulated therapy clips and provided a competence rating for each item associated with the clip. This provided 12 competence scores per participant.

3.1.1 Inter-rater reliability between participants

The ICC (2, 9) was carried out on the competence ratings assigned by participants to assess the degree of agreement. It found that inter-rater reliability between participants was 'good' (ICC = .87), with confidence intervals suggesting that reliability fell between 'moderate' and 'excellent' (ICC= .72 - .95). Reliability for a single average rater, calculated using ICC (2, 1) was found to be 'poor' (ICC = .42), with confidence intervals between 'poor' and 'moderate' (ICC= .22 - .7). The results for both analyses were significant ($p < .0005$). See Table 1. (See Appendix N for ICC results using 'consistency' which found similar coefficients.)

[3] The content analysis exploring aspects of decision making is included in an addendum, along with summary tables for the analysis presented in the report. This addendum is available on request.

Table 1. Results of ICC calculations for inter-rater reliability between participants. For multiple raters and single rater, with absolute-agreement, two-way random effects model

		95% Confidence Interval		F Test with True Value 0			
	ICC	Lower Bound	Upper Bound	Value	*df1*	*df2*	Sig (p<)
Average-measures	.866	.716	.954	7.827	11	88	.0005
Single-measures	.418	.219	.699	7.827	11	88	.0005

3.1.2 Inter-rater reliability between the participants' rating and the 'official' rating assigned by researchers

The data analysed was the mean participant rating (transformed into a rating of 'competent' or 'less competent') and the official researcher rating (see Table 2). The mean competence rating of participants matched the official rating of the researchers for seven out of twelve (58%) items across the ten videos. Furthermore, participants assigned a rating of 'competent' to nine of the items (75%). They assigned the correct rating in 83% of the competent videos, and the correct rating in 33% of the less competent videos.

Table 2. The official competence rating assigned by researchers, mean participant rating and over all competence rating (using the cut-off of 2)

Official rating of video/item assigned by researchers	**Mean rating of participants (n = 9)**	**Overall competence rating of participants**	**Does participant rating match researcher rating?**
Item 1: Psychoeducation - competent	2	Less competent	No
Item 1: Psychoeducation – less competent	3	Competent	No
Item 2: Recognising motives and emotions - competent	4	Competent	Yes
Item 2: Recognising motives and emotions – Less competent	1.33	Less competent	Yes
Item 4: Understanding the relationship between the three systems - competent	3.44	Competent	Yes
Item 4: Understanding the relationship between the three systems – Less competent	1.55	Less competent	Yes
Item 6: Building motivation - competent	3.17	Competent	Yes
Item 6: Building motivation – less competent	2.11	Competent	No
Item 10: Functional analysis - competent	2.22	Competent	Yes
Item 10: Functional analysis – less competent	3	Competent	No
Item 11: Fears, blocks and resistances - competent	2.17	Competent	Yes
Item 11: Fears, blocks and resistances – less competent	3	Competent	No

ICC (2, *2*) was used to determine the interrater reliability between participants and the official rating that the researchers assigned to each video clip. ICC was not calculated for a single rater. The inter-rater reliability was found to be 'poor' (ICC = .3) with confidence intervals falling within the 'poor' to 'good' range (ICC = -1.1 - .79). These results were not significant (p = .267) (see Table 3).

Table 3. Results of ICC calculations for inter-rater reliability between participants and researchers. For multiple raters, with absolute-agreement, two-way random effects model

		95% Confidence Interval		F Test with True Value 0			
	ICC	Lower Bound	Upper Bound	Value	*df1*	*df2*	Sig
Average-measures	.304	-1.083	.790	1.471	11	11	.267

3.2 Qualitative content analysis

The nine semi-structured interviews were analysed using content analysis to answer three distinct questions. First, section 1 summarises the changes to each of the six individual items that were suggested by participants. Section 2 summarises feedback about the CFT-TCRS overall. The code counts present information on the number (percentage) of participants who mentioned the codes. (Given the extent of the qualitative findings, only a selection of illustrative comments are highlighted in the text)[4]

3.2.1 Section 1: Changes to items of the CFT-TCRS

The content analysis revealed a variety of suggestions for changes to the items. There were two categories of suggestions for each item and these were labelled *general changes* and *specific changes.* A summary of the results is presented in Table 4 which outlines the *general* and *specific* changes suggested for each item, the frequency of the individual suggestions, and the number of participants who made comments relating to each category.

[4] Further examples are included in the appendices.

Table 4. General and specific changes suggested for each of the items, code counts (number of raters who identified each code), categories and the number of raters who identified each category

Item	Codes	Number of raters identifying code	Category	Number of raters identifying category
Item 1: Psychoeducation	Use sub-items	1 (11%)	General changes to item	4 (44%)
	More explicit scoring guidance	3 (33%)		
	More consistent CFT language	3 (33%)		
	Split into 2 items	1 (11%)	Specific changes to item	4 (44%)
	Include "not your fault" as a key concept	1 (11%)		
	Include an embodiment/ relational anchor	1 (11%)		
	More examples of key concepts	1 (11%)		
Item 2: Recognising motives and emotions	More explicit scoring guidance	1 (11%)	General changes to item	1 (11%)
	Include a physiology anchor	1 (11%)	Specific changes to item	5 (56%)
	Include an embodiment/ relational anchor	2 (22%)		
	Include therapeutic skill anchors	1 (11%)		
	Include a 'diagram use' anchor	1 (11%)		
Item 4: Understanding the relationship between the three systems	Split into 2 items	2 (22%)	Specific changes to item	4 (44%)
	Include an embodiment/ relational anchor	2 (22%)		
Item 6: Building motivation	Anchors should be opposites	1 (11%)	General changes to item	2 (22%)
	More explicit scoring guidance	1 (11%)		
	Include an anchor for defining compassionate motivation	1 (11%)	Specific changes to item	4 (44%)
	Anchor for recognising barriers to motivation	1 (11%)		

Item	Codes	Number of raters identifying code	Category	Number of raters identifying category
	Change “or others” to “and others”	1 (11%)		
	Anchor for “encouraging motivational strategies in and out of therapy”	2 (22%)		
Item 10: Functional analysis	More ‘less competent’ anchors	1 (11%)	General changes to item	2 (22%)
	More explicit scoring guidance	1 (11%)		
	Split first point to consider (forms and functions/self-criticisms link to safety)	1 (11%)	Specific changes to item	2 (22%)
	Include an embodiment/ relational anchor	1 (11%)		
	Remove shame/self-criticism as safety strategies	1 (11%)		
	Clarify the difference between FBRs and safety strategies	1 (11%)		
Item 11: Fears, blocks and resistances	Use sub-items	1 (11%)	General changes to item	4 (44%)
	More explicit scoring guidance	3 (33%)		
	Split into 2 items: identifying and addressing FBRs	3 (33%)	Specific changes to item	4 (44%)
	Include an anchor for explicit definition of FBRs	3 (33%)		
	Clarify FBRs	3 (33%)		
	Include a less competent anchor for “intellectual understanding only”	1 (11%)		
	Add “in the therapeutic relationship” to the last competent anchor	1 (11%)		

As can be seen in Table 4, the most common *general* change highlighted across items was for more explicit scoring guidance. This was suggested for all items apart from item 4 (building motivation). This concerned the need for more clarity around what would be

expected in a good CFT session (e.g., whether all 'points to consider' and anchors need to be displayed for full marks) and the need for more detail in the items (e.g., examples of skills required by the therapist). Item 1 (psychoeducation) and item 11 (fears, blocks and resistances; FBRs) were both highlighted as needing to be described more explicitly by 33% of the participants. For example, for psychoeducation the feedback included "[W]e need to be a bit more exhaustive with the examples we give or we need to not give them" and "[B]e a bit more clear about which things we're talking about, what would we expect to see in a good psychoeducation session". Some specific changes to items also related to increasing the clarity of items. For example, for item 10 (functional analysis) the need for more clarity around the differences between FBRs and safety strategies in CFT was suggested.

Most of the *specific* changes suggested by participants related to the behavioural anchors or the size of the item. New anchors, or changes to existing ones, were suggested for all 6 items. The need for an anchor capturing embodiment and relational aspects was the most common and this was suggested for four of the items, for example "[I]nclude a relational aspect, so how connected do the client and the therapist appear?". Other suggestions for changes to behavioural anchors were typically suggested by just one participant each. However, 33% of participants suggested that item 11 (FBRs) would benefit from an anchor capturing the therapist's explicit definition of FBRS. Furthermore, 22% of participants suggested that item 6 (building motivation) should include an anchor for encouraging motivational strategies in and out of therapy.

Finally, several participants suggested changes related to the size (or perceived compound nature) of the items. Suggestions included splitting criteria into sub-items which can be individually rated, or splitting items into two distinct ones. For example, item 11 (FBRs) covers both identifying a client's fears and addressing them. Participants found this difficult to rate and splitting the item into two was suggested by 33% of participants, for example "I would just split it into two and say that's being able to identify the FBRs and then being able to work with the FBRs". It was clear from the analysis that item 11 (FBRs) was highlighted most often as needing to be improved.

3.2.2 Section 2: Feedback on the CFT-TCRS as a whole

The content analysis on the feedback about the CFT-TCRS yielded five categories; *benefits of the CFT-TCRS, difficulties with the CFT-TCRS, changes to the structure of the CFT-TCRS, changes to the content of the CFT-TCRS*, and *the CFT-TCRS in practice*. Table 5 shows the categories, the associated codes and code counts, and the percentage of participants who made comments relating to that category.

Table 5. Overall scale feedback, codes, code counts (number of raters who identified each code), categories and the number of raters who identified each category

Codes	Number of raters identifying code	Category	Number of raters identifying category
Easy to use	5 (56%)	Benefits of the CFT-TCRS	8 (89%)
Useful scale	4 (44%)		
Items are precise	3 (33%)		
Well worded	6 (67%)		
Items are too big	4 (44%)	Difficulties with the CFT-TCRS	5 (56%)
Difficult to rate	3 (33%)		
Anchors repeat points to consider	1 (11%)		
Easier to rate competent CFT	2 (22%)		
Looks like visual analogue scale	1 (11%)	Changes to the structure of the CFT-TCRS	5 (56%)
Use sub-items	2 (22%)		
Include anchors for all scores	2 (22%)		
Capture microskills in every item	1 (11%)		
Use a 7-point Likert scale	1 (11%)		
Include relational components	2 (22%)	Changes to the content of the CFT-TCRS	8 (89%)
Include process issues	2 (22%)		
Needs more CFT language	1 (11%)		
Make items more explicit	4 (44%)		
Clarify scoring	5 (56%)		
Use with segments of sessions	4 (44%)	The CFT-TCRS in practice	6 (67%)
Needs trainee direction	1 (11%)		
Useful for training	3 (33%)		

Table 5 shows that two categories were discussed most frequently. Feedback relating to the *benefits of the CFT-TCRS* was brought up by 89% of the participants and included that the scale was well worded, precise and useful. *Changes to the content of the CFT-TCRS* were suggested by 89% participants, with most feedback relating to improving the clarity of the CFT-TCRS. For instance, 56% of participants suggested that it was unclear how they should score the items, for example "[I] cannot say 'unable to rate' to all the criteria but I need to say it for some criteria". Further, 44% of participants suggested that items should be more explicit in content and should "[B]e a bit more exhaustive with the examples". In line with section 1 of the content analysis, this category also highlighted the need to include relational and process issues in the CFT-TCRS.

Feedback on *the CFT-TCRS in practice* was provided by 67% of participants. The most common suggestion (44%) was that the scale should be used to rate segments of therapy sessions rather than full sessions. For example "[W]atching the videos individually split into 10 minutes made it a lot easier to concentrate on each item". One participant suggested that, for this to be useful, trainees who submit video segments should provide raters with some direction as to which competencies they displayed in that segment. For example "[Y]ou'd need some focusing from the person to say 'at this point I was doing this'". Finally, this category included feedback from 33% of participants that the CFT-TCRS would be a useful tool for therapist training.

Comments related to *structural changes to the CFT-TCRS* were suggested by 56% of participants. The most common suggestion was that items should be broken down into sub-items, for example "Break it into sub-items so the person can just go through and circle things really quickly". Furthermore, 22% of participants suggested that all possible scores on the scale should be accompanied by a descriptive anchor, rather than just having an anchor for each extreme on the Likert scale. The last category concerned *difficulties with the CFT-TCRS* and was mentioned by 56% of participants. The most frequent feedback was that the items were too big (44%), for example "There's too many points". In addition, 33% of participants suggested that they found it difficult to rate, for example "It's incredibly difficult to rate something using more than one criteria". Finally, some participants (22%) commented that competent CFT was easier to rate on the scale than less competent CFT.

4 Discussion

The present study aimed to explore the psychometric properties of the CFT-TCRS and to elicit expert feedback. A summary of the main findings, strengths and limitations, and implications for clinical practice and future research are discussed below.

4.1 Psychometric properties of the CFT-TCRS

The inter-rater reliability between CFT experts was 'good', suggesting they agreed with each other when assigning competence ratings. However, inter-rater reliability did not reach 'excellence', perhaps because agreement is often higher with generic therapeutic competencies (Morrison & Barratt, 2010) which are observable across therapeutic modalities, and the current study focused on a subset of CFT-specific competencies. However, the level of inter-rater agreement for this subset of items was similar to that of the full CTS-R (Blackburn *et al.,* 2001), a reliable and widely used CBT competence scale, which is a promising finding. The agreement between the raters in the current study might have been due to their similar levels of experience and thus a similar level of skill in detecting competent and less competent performance, which was expected given the stringent inclusion and exclusion criteria.

However, the inter-rater reliability between the CFT experts and the 'official' competence rating of the researchers was 'poor'. The participants did not consistently agree with the ratings assigned by the researchers. Participants sometimes rated 'less competent' videos as 'competent' and 'competent' videos as 'less competent'. This suggested that the video performances designed to display either 'competent' or 'less competent' CFT were not completely successful. Perfect agreement between the participants' ratings and researchers' intentions was not expected, but the poor inter-rater reliability suggested that the video performances of competent and less competent CFT were not as differentiated as intended.

One possible reason for the relatively poor agreement between participants' ratings and researchers' intentions was that a senior expert in the CFT community acted as the therapist in the videos. This may have resulted in a "halo" effect whereby participants may have rated "individual items too similarly as a result of their overall impression of the therapist's competence" (Schmidt *et al.,* 2018, p.370). This might have increased the

likelihood of a 'competent' rating where it was not warranted and this possible confound might have been avoided by using a different therapist or by using an actor.

The reliability for a single, average rater was found to be 'poor'. This suggested that a single person making competence ratings on the CFT-TCRS would not be reliable. However, this finding should be interpreted with caution as it was likely to be due to the relatively small number of participants.

4.2 Expert feedback on the CFT-TCRS and suggested changes

The content analysis highlighted several areas for improvement for the CFT-TCRS.

4.2.1 Overly complex items

Participants reported that they were sometimes unsure how to assign a single score to some items. The raters suggested that several items could be split into two distinct ones (e.g., item 11 – fears, blocks and resistances). Some participants also recommended that some could usefully be split into sub-items, allowing a score for each specific skill or criteria (e.g., item 1 – psychoeducation).

4.2.2 More examples and more scoring labels

Some participants suggested that items should include more examples of key skills, more clearly defined concepts, and clearer differentiation between similar concepts (e.g., safety strategies vs. fears, blocks and resistances). It was also highlighted that having a label for each possible score on the Likert-scale (not just the anchors) would make the CFT-TCRS easier to use and score.

4.2.3 Behavioural anchors and relational processes

The feedback about individual items of the CFT-TCRS mainly focused on the behavioural anchor labels at each end of the item scale. Participants indicated that aspects of some of the anchors seemed to be missing. For example, feedback indicated that item 2 (recognising motives and emotions) required more specific anchors to capture what might constitute 'competent' and 'less competent' delivery of the skill. For several of the items, the experts suggested that their anchors might be modified to account for relational and process components of CFT delivery. Although some relational aspects

are picked up in the microskills section of the CFT-TCRS, this feedback suggested that some important competencies of CFT might not have been covered or were not explicit enough.

Overall, participants tended to report similar (common) difficulties with the scale. This supports that they might have experienced the scale in a similar way, which was also reflected in the high level of agreement between the competence scores assigned by the participants. Furthermore, the feedback about the lack of clarity might partly account for the low agreement between the participants and researchers. Uncertainty might have meant that when assigning competence scores, participants relied on idiosyncratic interpretations of the items and their own clinical skills, which have been found to be associated with the rating of trainee competence (Kogan *et al.,* 2010). This more subjective approach might have reduced reliability, similarly to the suggestions of Schmidt *et al.,* (2018).

More generally, feedback suggested that the scale was useful and easy to use. However, it is important to address this expert feedback as competence scales are often criticised for unclear instructions, ambiguity and lack of specificity in behavioural anchors (Muse & McManus, 2015).

4.3 Strengths and limitations

This study was the first to explore the psychometric properties of the CFT-TCRS. It used a novel methodology and provided recommendations for possible refinements to the scale. The research highlighted some advantages and disadvantages of using simulated CFT sessions in an inter-rater reliability study. The mixed methods approach allowed for qualitative data to be considered alongside the inter-rater reliability results and enabled a richer exploration of the factors (e.g., difficulties with using the scale) which might have impacted on how the scale was used. The inclusion and exclusion criteria for recruitment meant that the eventual participants were representative of the type of expert therapists who might be expected to use the scale when training others in CFT.

However, there were a number of limitations. The use of video and actors meant that the generalisability of the results to real therapy sessions was limited. In addition, the inter-rater reliability results were likely to have been impacted by the small sample size,

due to the small pool of clinicians eligible for inclusion, and the relatively low number of ratings obtained from each participant. Whilst a small number of raters is not atypical for studies into the psychometric properties of therapist competence scales (e.g., Blackburn *et al.,* 2001; Haddock *et al.* 2001), Koo and Li (2016) suggest that at least 30 ratings should be obtained from a minimum of three raters. Therefore, caution should be taken when interpreting the results of this study.

The study asked participants to rate the same 'therapist-patient' dyad multiple times. Therefore, as with research using similar procedures (e.g., von Consbruch *et al.,* 2012), the data set was not independent and so rater confounds such as fatigue and practice effects cannot be ruled out. Another methodological limitation related to the design and development of the video clips. The videos were not piloted to ensure they reliably displayed what was intended, which may have played a role in the low agreement between the participants' ratings and researchers' intentions.

4.4 Implications for research and clinical practice

The subset of CFT-TCRS items, even in its current form, had 'good' inter-rater reliability when used by a group of CFT experts. This suggested that the scale as a whole might be suitable for use during training, research and audits. Indeed, it has been suggested that measuring competence can maximise the provision of effective, evidence-based training and treatment (McHugh & Barlow, 2010). However, as concluded by Muse and McManus (2013), a multi-method approach to assessing competence is beneficial. Therefore, services and training courses might consider using the CFT-TCRS alongside other measures such as essays, role-plays and multiple-choice questionnaires. The 'poor' reliability of a single rater also has implications for competence assessment and suggests that caution should be taken when only using one assessor. This finding might have been due to methodological flaws and warrants further investigation.

Clinicians understandably vary in terms of their background, training and experiences, and perhaps their ideas around what constitutes therapist competence. Although scales such as the CFT-TCRS aim to operationalise the specific competencies required for the delivery of a therapy, it is likely that clinical experience impacts competence decisions (Kogan *et al.,* 2010). Although variation in clinicians is unavoidable, it is important that

consistency can be ensured when using competence scales. This, along with participant feedback, indicated that the CFT-TCRS might benefit from a 'user manual' to clarify how to score items and provide further information around expected skills.

Participants highlighted that the CFT-TCRS felt easier to use with segments of therapy sessions, but that trainee therapists would need to indicate which competencies they were displaying in each segment. Therefore, for trainees and less experienced CFT clinicians to reliably use the CFT-TCRS, training might be beneficial. This could be service specific or more general training for all CFT clinicians. It is important to highlight that the CFT-TCRS was not designed to be used in its entirety for each competence assessment, as not all competencies would be displayed in one session of CFT.

4.5 Future research

The CFT-TCRS would benefit from being assessed further for reliability, using a larger sample, and collecting a larger number of competence ratings. The reliability of the scale could be further explored by asking CFT experts to rate therapists in real therapy sessions, which is the most common approach when assessing the psychometric properties of a competence scale (Bennet & Parry, 2004). If simulated therapy sessions are used, they should be carefully designed and piloted. Researchers may consider using scripts and an actor who is not trained in CFT in order to control for factors which might impact competence rating (e.g., the "halo" effect).

Future research might continue to explore the psychometric properties of the CFT-TCRS, such as internal consistency and predictive ability. Discriminant validity could also be assessed by rating trainees at different points during training. If the CFT-TCRS is a valid measure of CFT competence, then trainees' competence scores on the scale should increase over the course of training (Blackburn *et al.,* 2001).

Measures of treatment fidelity are important for use in outcome studies, as it has been found that higher levels of therapist competence predict improved treatment outcomes (Shaw *et al.,* 1999). The initial inter-rater reliability results for the CFT-TCRS were promising and suggested it might be a useful tool in outcome research. However, the reliability of the CFT-TCRS might well be improved by altering the scale in line with

expert feedback and this might increase the likelihood of valid and reliable conclusions from outcome research (Perepletchikova *et al.,* 2009).

It is important to note that research has been conducted into the efficacy of group based CFT (e.g., Kelly *et al.,* 2017). It would therefore be beneficial for future research to assess whether the CFT-TCRS can be used to reliably assess the competence of therapists providing group CFT, and whether it captures the additional competencies required when working with multiple patients.

5 Conclusion

CFT experts demonstrated high agreement with each other when assigning a competence rating to a therapist in a simulated CFT session. However, agreement between the raters and the researchers' competence rating was 'poor'. Participant feedback on the overall scale was generally positive in terms of its usefulness, but it was suggested it could be made more explicit and specific, where more anchor points and less complex items would make it easier to use. For individual items of the CFT-TCRS, experts suggested that more inclusive behavioural anchors would make them easier to use. Future researchers might assess the psychometric properties of the CFT-TCRS when used with real therapy sessions. With some revision and further assessment, it is anticipated that the CFT-TCRS will be of value in clinical practice, research and therapist training.

References

Barber, J. & Critis-Christoph, P. (1996). Development of a therapist adherence/competence rating scale for supportive-expressive dynamic psychotherapy: a preliminary report. *Psychotherapy Research, 6*(2), 81-94.

Barber, J. P., Liese, B. S., & Abrams, M. J. (2003). Development of the cognitive therapy adherence and competence scale. Psychotherapy Research, 13(2), 205-221.

Barber, J., Krakauer, I., Calvo, N. et al. (1997). Measuring adherence and competence of dynamic therapists in the treatment of cocaine dependence. *The Journal of Psychotherapy Practice and Research, 6*(1), 12-24.

Barber, J. P., Sharpless, B. A., Klostermann, S. & McCarthy, K. S. (2007). Assessing intervention competence and its relation to therapy outcome: A selected review derived from the outcome literature. *Professional Psychology: Research and Practice, 38*(5), 493-500.

Beaumont, E., Durkin, M., McAndrew, S., & Martin, C. R. (2016). Using compassion focused therapy as an adjunct to trauma-focused CBT for fire service personnel suffering with trauma-related symptoms. *The Cognitive Behaviour Therapist, 9*(34), 1-13.

Bennett, D. & Parry, G. (2004). A measure of psychotherapeutic competence derived from cognitive analytic therapy. *Psychotherapy research, 14*(2), 176-192.

Berman, J. S., & Norton, N. C. (1985). Does professional training make a therapist more effective? *Psychological Bulletin, 89,* 401–402.

Blackburn, I. M., James, I. A., Milne, D. L. et al. (2001). The revised cognitive therapy scale (CTS-R): psychometric properties. *Behavioural and Cognitive Psychotherapy, 29*(04), 431-446.

Boyle, K., Deisenhofer, A. K., Rubel, J. A. et al. (2019). Assessing treatment integrity in personalised CBT: The inventory of therapeutic interventions and skills. *Cognitive Behaviour Therapy.* DOI:10.1080/16506073.2019.1625945

Braehler, C., Gumley, A., Harper, J. et al. (2013). Exploring change processes in compassion focused therapy in psychosis: Results of a feasibility randomized controlled trial. *British Journal of Clinical Psychology, 52*(2), 199-214.

Consbruch, K. von, Clark, D. M., & Stangier, U. (2012). Assessing therapeutic competence in cognitive therapy for social phobia: Psychometric properties of the cognitive therapy competence scale for social phobia (CTCS-SP). Behavioural and Cognitive Psychotherapy, 40(2), 149-161.

Fairburn, C. G., & Cooper, Z (2011). Therapist competence, therapy quality, and therapist training. *Behaviour Research and Therapy, 49,* 373-378.

Gale, C., Gilbert, P., Read, N., & Goss, K. (2014). An evaluation of the impact of introducing compassion focused therapy to a standard treatment programme for people with eating disorders. *Clinical Psychology and Psychotherapy, 21*(1), 1-12.

Gilbert, P. (2009). Introducing compassion-focused therapy. *Advances in Psychiatric Treatment, 15*(3), 199-208.

Gilbert, P. (2014). The origins and nature of compassion focused Therapy. *British Journal of Clinical Psychology, 53,* 6–41.

Gilbert, P., & Proctor, S. (2006). Compassionate mind training for people with high shame and self-criticism: Overview and pilot study of a group therapy approach. *Clinical Psychology and Psychotherapy, 13*(6), 359-379.

Gilbert, P., Wood, W., & Gale, C. (2012). *Therapy Assessment Guide*. Compassionate Mind Foundation. Unpublished.

Haddock, G., Devane, S., Bradshaw, T. et al. (2001). An investigation into the psychometric properties of the cognitive therapy scale for psychosis (CTS-Psy). *Behavioural and Cognitive Psychotherapy, 29*(2), 221-233.

Hattie, J. A., Sharpley, C. F., & Rogers, H. J. (1984). Comparative effectiveness of professional and paraprofessional helpers. *Psychological Bulletin, 95,* 534–541.

Horwood, V., Allan, S., Goss, K., & Gilbert, P. (2019). The development of the Compassion-Focused Therapy Therapist Competence Scale. *Psychology and Psychotherapy: Theory, Research and Practice.* DOI:10.1111/papt.12230

Hsieh, H. F., & Shannon, S. E. (2005). Three approaches to qualitative content analysis. *Qualitative health research, 15*(9), 1277-1288.

Kelly, A. C., Wisniewski, L., Martin-Wagar, C., & Hoffman, E. (2017). *Group-based compassion focused therapy as an adjunct to outpatient treatment for eating disorders: A pilot randomized controlled trial. Clinical Psychology and Psychotherapy, 24*(2), 475-487.

Kogan, J. R., Hess, B. J., Conforti, L. N. et al. (2010). What drives faculty rating of residents' clinical skills? The impact of faculty's own clinical skills. *Academic Medicine, 85*(10), S25-S28.

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine, 15*(2), 155-163.

Laithwaite, H., O'Hanlon, M., Collins, P. et al. (2009). Recovery after psychosis (RAP): A compassion focused programme for individuals residing in high security settings. *Behavioural and Cognitive Psychotherapy, 37*(5), 511-526.

Leaviss, J., & Uttley, L. (2015). Psychotherapeutic benefits of compassion-focused therapy: an early systematic review. *Psychological Medicine, 45*(5), 927-945.

Lemma, A., Roth, A.D. & Pilling, S. (2008). The competencies required to deliver effective Psychoanalytic/ Psychodynamic Therapies. Retrieved 2 April 2018 from http://www.ucl.ac.uk/ clinical-psychology/CORE/psychodynamic_framework.htm

Liddell, A. E., Allan, S., & Goss, K. (2017). Therapist competencies necessary for the delivery of compassion-focused therapy: A Delphi study. *Psychology and Psychotherapy: Theory, Research and Practice, 90,* 156-176.

Lucre, K. M., & Corten, N. (2013). An exploration of group compassion-focused therapy for personality disorder. *Psychology and Psychotherapy: Theory, Research, and Practice, 86*(4), 387-400.

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. Psychological methods, 1(1), 30.

McHugh, R. K., & Barlow, D. H. (2010). The dissemination and implementation of evidence-based psychological treatments. *American Psychologist, 65,* 73-84.

Morrison, A. P., & Barratt, S. (2010). What are the components of CBT for psychosis? A Delphi study. *Schizophrenia Bulletin, 36*(1), 136-142.

Muse, K., & McManus, F. (2013). A systematic review of methods for assessing competence in cognitive-behavioural therapy. *Clinical Psychology Review, 33*(3), 484-499.

Muse, K., & McManus, F. (2015). Expert insight into the assessment of competence in cognitive-behavioural therapy: A qualitative exploration of experts' experiences, opinions and recommendations. *Clinical Psychology and Psychotherapy, 23,* 246-259.

Perepletchikova, F., Hilt, L. M., Chereji, E., & Kadzin, A. E. (2009). Barriers to implementing treatment integrity procedures: Survey of treatment outcome researchers. *Journal of Consulting and Clinical Psychology, 77*(2), 212-218.

Perepletchikova, F., Treat, T. A. & Kazdin, A. E. (2007). Treatment integrity in psychotherapy research: Analysis of the studies and examination of the associated factors. *Journal of Consulting and Clinical Psychology, 75*(6), 829-841.

Roth, A. D., Pilling, S. (2007). The competencies required to deliver effective cognitive and behavioural therapy for people with depression and with anxiety disorders. Department of Health: London.

Schmidt, I. D., Strunk, D. R., DeRubeis, R. J. et al. (2018). Revisiting how we assess therapist competence in cognitive therapy. *Cognitive Therapy and Research, 42*(4), 369-384.

Sharpless, B. A., & Barber, J. P. (2009). A conceptual and empirical review of the meaning, measurement, development, and teaching of intervention competence in clinical psychology. *Clinical Psychology Review, 29,* 47-56.

Shaw, B. F., Elkin, I., Yamaguchi, J. 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.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*, 420-428.

Southam-Gerow, M. A., & McLeod, B. D. (2013). Advances in applying treatment integrity research for dissemination and implementation science: Introduction to special issue. *Clinical Psychology: Science and Practice, 20*(1), 1-13.

Stallard, P., Myles, P., & Branson, A. (2014). The cognitive behaviour therapy scale for children and young people (CBTS-CYP): Development and psychometric properties. *Behavioural and Cognitive Psychotherapy, 42,* 269-282.

Vermilyea, B. B., Barlow, D. H., & O'Brien, G. T. (1984). The importance of assessing treatment integrity: An example in the anxiety disorders. *Journal of Behavioural Assessment, 6,* 1-11.

Waltz, J., Addis, M. E., Koerner, K., & Jacobson, N. S. (1993). Testing the integrity of a psychotherapy protocol: Assessment of adherence and competence. *Journal of Consulting and Clinical Psychology, 61*(4), 620-630.

Wampold, B. E. (2001). *The Great Psychotherapy Debate: Models, Methods and Findings.* Lawrence Erlbaum Associates: Mahwah, NJ.

Webb, C. A., DeRubeis, R. J., & Barber, J. P. (2010). Therapist adherence/competence and treatment outcome: A meta-analytic review. *Journal of Consulting and Clinical Psychology, 78*(2)), 200-211.

Part 3: Appendices

Appendix A: Rationale for databases and search terms used for the review

Table 1. Rationale for databases used.

Database used:	Rationale:
PsychINFO	Provides up to date literature which is specific to psychology but includes related disciplines.
Scopus	Provides access to a comprehensive search of resources from wider multidisciplinary and scientific disciplines.
Web of Science	Provides access to multidisciplinary research across a variety of subjects.
Medline	Provides a broader search of medical articles.

Table 2. Search terms and filters used on 14th September 2019, and initial results (terms also used in January 2020).

Search terms used	Database				Total
"Therap*" OR "Psychotherap*" **AND** "Competen*" **AND** "Scale" OR "rate" OR "rating" OR "measure" **AND** "Psychometric" OR "Develop*"	PsychINFO	Scopus	Web of Science	Medline	
	2357	2797	1513	6718	13385
Filter 1: Published in a peer reviewed journal	1927	2671	1460	6614	12672
Filter 2: written in English	1758	2407	1319		12098

*(blank cell = filter not available on database)

Appendix B: Bespoke quality appraisal and data extraction form

Quality appraisal and data extraction form

Score 1 for YES and 0 for NO / PARTIALLY - Total score out of 20:

Basic Information

Title:		
Authors:		
Publication Date:	Country of study:	
Journal:		
Volume:	Number:	Pages:

Aims and Objectives

1. Are the aims/objectives of the study clearly described? YES / NO

What are the aims/objectives of the study?

2. Is the aim/objective related to the development or validation of a therapist competence scale? YES / NO / PARTIALLY

3. Is the scale designed to rate therapists who are using cognitive behavioural therapy? YES / NO / PARTIALLY

Method

4. Is the method used to develop the scale adequately described and based in evidence? YES / NO / PARTIALLY

What method was used to develop the scale?

5. Are the characteristics of those involved in scale development described? YES / NO / PARTIALLY

6. Were those involved in scale development representative of a population? YES / NO / PARTIALLY

Who was involved in scale development? (number, profession)

7. Is the scale adequately described? YES / NO / PARTIALLY

Final scale characteristics

8. Were methods used to measure reliability and validity of the scale adequately described? YES / NO / PARTIALLY

9. Were methods used to measure reliability and validity of the scale suitable/based in evidence? YES / NO / PARTIALLY

What methods were used to assess the reliability and validity of the scale? (videos, audio, real therapy, simulated therapy etc.)

10. Was the test evaluated in a sample of subjects who were representative of those whom the authors intended the therapy to be applied? YES / NO / PARTIALLY

11. Were these subjects adequately described? YES / NO / PARTIALLY

Who were the 'subjects' being rated? (therapists, trainees) Who were the 'patients'? (real patients, actors, stage of therapy)

12. Was the test performed by raters who were representative of those to whom the authors intended the results to be applied? YES / NO / PARTIALLY

13. Were raters adequately described? YES / NO / PARTIALLY

Who were the raters? (number, profession, therapists, experts etc.)

14. Were raters blind to the findings of other raters in the study? YES / NO / PARTIALLY

15. Were methods of data analysis adequately described? YES / NO / PARTIALLY

16. Were appropriate measures of agreement used? YES / NO / PARTIALLY

What methods were used to analyse qualitative/quantitative data?

Results and Discussion

17. Are the main findings adequately described? YES / NO / PARTIALLY

What are the main findings?

18. Are the conclusions adequately described? YES / NO / PARTIALLY

What are the conclusions?

19. Were the strengths/ limitations of the study adequately described? YES / NO / PARTIALLY

What are the strengths and limitations?

20. Are the clinical implications/implications for future research adequately described? YES / NO / PARTIALLY

What are the implications?

Other

Key references not identified from search strategy

Any other comments?

Appendix C: Items included in the bespoke quality appraisal and data extraction form

Item	Origin of item	Data extraction?
1	Item 1 from the Downs and Black (1998) checklist for measuring study quality	Yes – aims and objectives
2	Generated by the researcher to cover quality of scale development/validation	No
3	Generated by the researcher to cover quality of scale development/validation	No
4	Generated by the researcher to cover quality of scale development/validation	Yes – methods of scale development
5	Generated by the researcher to cover quality of scale development/validation	No
6	Generated by the researcher to cover quality of scale development/validation	Yes – scale developers
7	Generated by the researcher to cover quality of scale development/validation	Yes – scale characteristics
8	Generated by the researcher to cover quality of scale development/validation	No
9	Generated by the researcher to cover quality of scale development/validation	Yes – methods of assessing reliability/validity
10	Item 1 from the QAREL (Lucas *et al.*, 2010)	No
11	Item 3 from the Downs and Black (1998) checklist for measuring study quality	Yes – subjects/patients
12	Item 2 from the QAREL (Lucas *et al.*, 2010)	Yes - raters
13	Generated by the researcher to cover quality of scale development/validation	No
14	Item 3 from the QAREL (Lucas *et al.*, 2010)	No
15	Generated by the researcher to cover quality of discussion	No
16	Item 11 from the QAREL (Lucas *et al.*, 2010)	Yes - analysis
17	Item 6 from the Downs and Black (1998) checklist for measuring study quality	Yes – main findings
18	Generated by the researcher to cover quality of discussion	Yes – conclusions
19	Generated by the researcher to cover quality of discussion	Yes – limitations
20	Generated by the researcher to cover quality of discussion	Yes – implications

***** The tool also provided space for extraction of relevant references not found during the initial search and space for further comments.

Appendix D: Summary of aims, sample and design of included papers

Author/ date/ country	Title/scale	Aims	Patient sample/intervention	Therapist sample	Therapy sessions rated	Raters of therapist competence	Study design
Barber *et al.,* (2003) - USA	Development of the cognitive therapy adherence and competence scale (CTACS)	Assess psychometric properties of the CTACS in cocaine dependent patients. Assess whether the CTACS can distinguish between CT and other interventions.	129 randomly selected cocaine dependent patients from the NIDA CCTS. Randomly assigned to CT, SE, IDC GDC alone.	40 therapists from the NIDA CCTS study with either a PhD, master's in social work or a medical degree. *n*= 18 CT therapists (22% female) *n*= 12 SE therapists (33% female) *n*= 10 IDC therapists (60% female)	134 audio tapes (92 from CT, 20 from SE and 22 from IDC)	2 expert cognitive therapists (1 clinical psychologist and 1 psychiatric nurse).	Secondary data analysis/ RCT
Bjaastad *et al.* (2016) - Norway	Competence and adherence scale for cognitive behavioural therapy (CAS-CBT) for anxiety disorders in youth: psychometric properties	Develop a scale to measure competence in CBT for children/ adolescents. Establish psychometric properties.	182 patients (mean age = 11.5, range = 8-15, 53% female) diagnosed with: separation anxiety, social phobia or generalised anxiety.	17 therapists (94% female). 5 had formal 2-year CBT training. All received 2-day CBT training, 2-day FRIENDS programme training (Barrett, 2004, 2008) and 8-days training on anxiety in youth.	127 videotapes.	2 licensed CBT therapists and 2 psychology graduates (trained in the use of the CAS-CBT).	Secondary data analysis/ RCT

Author/ date/ country	Title/scale	Aims	Patient sample/intervention	Therapist sample	Therapy sessions rated	Raters of therapist competence	Study design
Blackburn *et al.,* (2001) - UK	The revised cognitive therapy scale (CTS-R): Psychometric properties	Devise an updated version of the CTS and reduce overlap between items. Improve the scaling system. Define items more clearly.	34 patients receiving weekly CBT (mean age= 37.0) diagnosed with depression (*n*= 16), social phobia (*n*= 6), panic disorder (*n*= 5), OCD (*n*= 4) and generalised anxiety disorder (*n*= 3).	20 trainee CBT therapists (55% female) made up of specialist nurses (*n*= 7), clinical psychologists (*n*= 6), psychiatrists (*n*= 5), a senior registrar and a senior nurse.	102 videotapes (3 for each patient) reflecting 3 stages of therapy (start, middle and end).	4 experts in CBT who were trainers and supervisors from the course which the therapists were training on.	Non-RCT/ data from training program
Boyle *et al.,* (2019) Germany	Assessing therapeutic integrity in personalised CBT: The inventory of therapeutic interventions and skills (ITIS)	Devise a tool to assess adherence and competence in modern, personalised CBT.	70 patients (mean age= 36.8, 64.3% female) with DSM-IV diagnoses who were randomly assigned to a therapist.	30 therapists (86.7% female) with a master's degree in clinical psychology (either training in CBT or licensed CBT practitioners).	185 videotapes (1-3 for each patient).	4 graduate clinical psychology students and 2 post-graduate clinicians trained in the use of the ITIS.	Secondary data analysis/ RCT
Carroll *et al.,* (2000) - USA	A general system for evaluating therapist adherence and competence in psychotherapy	Develop a general psychotherapy rating scale. Report development and	117 patients (mean age= 30.0, 27% female) with cocaine dependence. Patients were randomised to	Therapists (reported in Carroll *et al.,* 1998) provided weekly sessions over a 12-week period.	576 videotapes (no more than 1 from any given week)	5 raters, mostly master's level clinicians with experience treating substance users with either CBT,	Secondary data analysis/ RCT

Author/ date/ country	Title/scale	Aims	Patient sample/intervention	Therapist sample	Therapy sessions rated	Raters of therapist competence	Study design
	research in the addictions	validation of the Yale Adherence and Competence Scale (YACS)	either CBT with medication, TSF with medication, CM with medication, CBT only or TSF only.			TSF or CM. Trained in YACS use.	
Dobson *et al.,* (1985) - USA	Reliability of a measure of the quality of cognitive therapy.	Assess reliability and validity of the CTS (Young & Beck, 1980). To move towards evaluating predictive validity of the CTS.	21 patients diagnosed with depression.	21 psychotherapists (10 psychiatrists and 11 psychologists) with a minimum of 2 years clinical experience and who were applying for specialist training in CBT.	21 videotapes (1 per therapist submitted as part of the selection process for CBT training).	4 CBT therapists who had made contributions to the development of CBT and were trained in the CTS.	Secondary data analysis/ RCT
Haddock *et al.,* (2001) - UK	An investigation into the psychometric properties of the cognitive therapy scale for psychosis (CTS-Psy)	Investigate inter-rater reliability and validity of the CTS-Psy.	84 patients (81 diagnosed with schizophrenia) receiving individual/ family CBT.	21 trainee CBT therapists from 2 cohorts of the Manchester University Thorn Initiative training (20 mental health nurses and 1 occupational therapist).	5 audio tapes selected by the authors to display a range of competence.	4 raters (authors of the paper) with CTS-Psy training and specialist CBT training.	Non-RCT/ data from training program

Author/ date/ country	Title/scale	Aims	Patient sample/intervention	Therapist sample	Therapy sessions rated	Raters of therapist competence	Study design
McLeod *et al.,* (2018) - USA	Development and initial psychometrics for a therapist competence instrument for CBT for youth anxiety	Develop a tool to measure global/discrete therapist competence in CBT for youth. Test psychometric properties of the CBAY-C.	68 patients (mean age = 10.6, 47.1% female, 82.3% Caucasian) from an RCT comparing individual CBT to family CBT. All met the diagnostic criteria for an anxiety disorder.	29 therapists (86.8% female) delivering the 'Coping Cat' CBT programme. Therapists were clinical psychology doctoral trainees, licensed psychologists and clinical employees.	744 videotapes taken from a sample of 1098 (first and last sessions were excluded).	4 clinical psychology doctoral trainees with training/ experience in CBT for anxiety. Used CBAY-C. 4 trainees used other scales to rate videos.	Secondary data analysis/ RCT
Roth (2016) - UK	A new scale for the assessment of competences in cognitive and behavioural therapy (UCL scales)	Develop a scale to measure competence of CBT therapists.	Described fully in another study (Roth *et al.,* 2019)	Described fully in another study (Roth *et al.,* 2019)	Videotapes of therapy sessions. Described fully in another study (Roth *et al.,* 2019)	Described fully in another study (Roth *et al.,* 2019)	Non-RCT/ data from training program
Roth *et al.,* (2019) – UK	Judging clinical competence using structured observation tools: A cautionary tale	Examine inter-rater reliability of the UCL scales and the CTS-R.	Adults between the ages of 19 and 62 who were accessing therapy in the service where the therapists were employed. Primary diagnosis of	14 therapists (78.5% female) undertaking a one-year Postgraduate Diploma in CBT as part of the Improving Access to Psychological	25 videotapes submitted by the therapists. Purposively selected from a sample of 76 tapes by an independent research assistant	6 tutors from the IAPT Postgraduate Diploma programme who were CBT accredited Clinical Psychologists.	Non-RCT/ data from training program

Author/ date/ country	Title/scale	Aims	Patient sample/intervention	Therapist sample	Therapy sessions rated	Raters of therapist competence	Study design
	(UCL scales)		depression or an anxiety disorder.	Therapies (IAPT) programme. All had at least two years of clinical experience, but varied in profession, experience and exposure to CBT interventions.		Raters were trained to use the CTS-R through work. Training on the UCL scales was limited.	
Stallard *et al.,* (2014) - UK	The cognitive behaviour therapy scale for children and young people (CBTS-CYP): Development and psychometric properties	Develop/evaluate a scale for assessing competence in therapists providing CBT for children and young people.	1 male adolescent patient with anxiety disorder (video). 48 patients (mean age= 14.44, 64.6% female). Diagnosed with depression (*n*= 20), separation anxiety (*n*= 9), social anxiety (*n*= 6), OCD (*n*= 5), panic (*n*= 3), generalised anxiety (*n*= 3), phobia (*n*= 1) and PTSD (*n*= 1).	1 trainee undertaking training in CYP-IAPT CBT. 18 trainees undertaking training inn CYP-IAPT CBT.	1 videotape and 48 audio tapes used in 2 separate studies.	12 markers from the CYP-IAPT CBT training course.	Non-RCT/ data from training program
Vallis *et al.,* (1986) - USA	The cognitive therapy scale: Psychometric properties	Evaluate and present psychometric	Patients diagnosed with unipolar depression and who met the criteria for	9 psychotherapists (PhD or medical degree). Three from each of the 3 sites of	10 videotapes from a sample of 94 for inter-rater reliability study.	7 experts in CT (6 PhD and 1 M.D) with clinical experience and	Secondary data analysis/ RCT

Author/ date/ country	Title/scale	Aims	Patient sample/intervention	Therapist sample	Therapy sessions rated	Raters of therapist competence	Study design
		properties of the CTS.	major depressive disorder (exact number unknown).	the NIMH TDCRS in the USA. Each was trained in CT for at least 18-months and treated 4-5 patients.	90 tapes from a sample of 725 for internal consistency study, and 53 from the same sample for discriminant validity study.	experience training others in CT. 3 were trainees in the NIMH TDCRS programme and the remaining 4 were consultants.	
Von Consbruch *al.,* (2012) - Germany	Assessing therapeutic competence in cognitive therapy for social phobia: Psychometric properties of the cognitive therapy competence scale for social phobia (CTCS-SP)	Investigate psychometric properties of the CTCS-SP (Clark *et al.,* 2007). Determine if the CTCS-SP is reliable.	98 patients diagnosed with social phobia. Patients were part of the SOPHO-NET multi-centre trial in Germany which compared CT to SE (Leichsenring *et al.,* 2009).	51 therapists from the SOPHO-NET trial, trained in CT for social phobia (Stangier *et al.,* 2006).	161 videotapes randomly chosen from three stages of treatment (initial, middle and final).	6 trainee clinical psychologists and 1 psychotherapist. All trained in CT for social phobia and on use of the CTCS-SP. First author was a rater and rated all tapes.	Secondary data analysis/ RCT

Appendix E: Summary of scale development and validation, and results of interest

Study	Scale and constructs measured	Scale development	Final scale characteristics (CBT items only)	Methods used to assess validity and reliability	Analysis and results	Conclusions
Barber *et al.,* (2003)	CTACS Measures: adherence and competence in CT	Developed by 27 CT therapists and 5 psychology graduates for use during therapy for drug dependence (Liese *et al.,* 1995). CSPRS, CTS and other CT treatment manuals reviewed to generate items. CT therapists made amendments before psychology graduates used a second version and made suggestions. CT therapists decided on the final scale.	21 items split into 5 scales: CT structure, developing a therapeutic relationship, case conceptualisation, CT techniques and overall performance. 7-point Likert scale for competence and adherence on each item.	Raters listened to randomly assigned audio tapes of therapy sessions (CT, SE and IDC) and use the CTACS to rate competence and adherence.	Inter-rater reliability: ICCs were .67 for adherence and .73 for competence (moderate agreement) Internal consistency: Cronbach's alpha coefficients were very high (α = .92 for adherence, α = .93 for competence). Relationship between adherence and competence: Pearson's r = .96	The CTACS can be used to measure adherence/ competence with moderate reliability. The CTACS has good internal consistency and inter-rater reliability correlates well with other adherence and competence scales for cognitive therapy.
Bjaastad *et al.* (2016)	CAS-CBT Measures: adherence and competence in CBT	Developed by 3 trained CBT therapists/ supervisors and 1 expert in adult CBT. Scale based on the CTACS (Barber *et al.,* 2003). Developers chose relevant items and generated more to	11 items covering: CBT structure, goals and process. 3 supplementary items providing scores on: overall adherence, overall competence and session difficulty.	Students rated 127 videotapes and 20% were re-rated by CBT therapists. One CBT therapist re-rated videos 12 months later to	Inter-rater reliability: ICCs were .83 for adherence and .64 for competence. Rater stability: ICCs were .89 for adherence and .92 for competence.	The CAS-CBT can reliably measure competence in CBT for anxiety in youth. Training is required for the scale to be used properly/ reliably.

Study	Scale and constructs measured	Scale development	Final scale characteristics (CBT items only)	Methods used to assess validity and reliability	Analysis and results	Conclusions
		cover: CBT structure, facilitation of patient goals and process/ relational skills. A preliminary version was used and revised.	7-point Likert scale for competence and adherence on each item.	assess rater stability.	Internal consistency: Cronbach's α = .87 for the scale.	Important to differentiate between adherence and competence.
Blackburn *et al.,* (2001)	CTS-R Measures: competence in CT	Developed by 2 raters (a clinical psychologist and an educationalist) and 4 CBT experts. CBT videos were rated to redefine competence. 4 expert CBT therapists revised it and incorporated ideas from the 2 raters. They collapsed some items and added 3 new ones.	13 items covering: agenda setting, feedback, collaboration, pacing, interpersonal effectiveness, guided discovery, conceptualisation, key cognitions, change methods, behavioural techniques and homework. 7-point Likert scale for each item.	Each videotape was double rated by pairs (4 raters) in a balanced design. Each rater used the scale on 51 tapes.	Inter-rater reliability: ICC for scale was .63, Pearson's r= .66 For items the ICC ranged between -.14 and .084. Pearson's *r*= .42 - .84 Internal consistency: Cronbach's alpha ranged between α = .92- α = .97. Discriminant validity: ANOVA showed the CTS-R is sensitive to improved skill	The CTS-R can be used reliably and measures CBT competence. Training on the CTS-R is required as raters make inferences about therapist skill despite clear definitions of competence.
Boyle *et al.,* (2019)	ITIS Measures: adherence and competence in CBT	Developed by the authors of the paper. The CBT-AS (Weck *et al.,* 2014) and the CTS-R (Blackburn *et al.,* 2001) were used as a starting point. Third	2 sub-scales make up the ITIS. Interventions scale: 19 items measuring adherence. Measured on a 3-point Likert scale apart from	Clinical psychology students each independently rated 14 randomly assigned videotapes.	Inter-rater reliability: Kendall's W calculated pairwise. Excellent for interventions scale (students = .783, clinicians = .832) and skills scale (students = .703, clinicians = .700).	The ITIS might be a valid/reliable tool for assessing adherence and competence in modern CBT. It may help to make individual

Study	Scale and constructs measured	Scale development	Final scale characteristics (CBT items only)	Methods used to assess validity and reliability	Analysis and results	Conclusions
		wave approaches and other measures were reviewed to define 'modern CBT'. Developers went through a process of integration/synthesis/ summarisation.	overall adherence (6-point scale). Skills scale: 11 items measuring competence rated on 7-point Likert scales. Included an item rating overall competence.	Post-graduate clinicians each independently rated 48 randomly assigned videotapes.	Item correlations: interventions items had low inter-correlations; skills items were significantly correlated. Predictive validity: competence/adherence positively correlated with session outcome.	recommendations for patients and can successfully predict therapy outcome. The scale is short, easy to use and covers several skills.
Carroll *et al.,* (2000)	YACS Measures: adherence and competence in general drug abuse treatment, CBT, TSF and CM	Developed by the authors of the paper. Videotapes of sessions and treatment manuals (CBT, TSF and CM) were reviewed to generate items.	55 items each rated for competence and adherence. Subscales for: general drug abuse treatment, CBT, 12-step and CM. CBT scale: 6 items covering cue skills such as decision making, relapse prevention and confronting thoughts about substance use.	Randomly selected tapes were rated by each of the raters for different parts of the analysis.	Inter-rater reliability: ICC for the sub-scales, including CBT, showed high reliability (.80 - .95 for adherence, .71 - .95 for competence). Concurrent validity: Pearson's correlations between the scales showed significant negative correlations. Discriminant validity: ANOVA showed that the scale could discriminate between treatments.	The YACS can be used reliably to assess adherence and competence to several forms of therapy. It can be used to rate therapists providing CBT for cocaine dependent patients. The scale might enable treatment integrity to be more easily measured.

Study	Scale and constructs measured	Scale development	Final scale characteristics (CBT items only)	Methods used to assess validity and reliability	Analysis and results	Conclusions
Dobson *et al.,* (1985)	CTS Measures: competence in CBT	Developed by Young and Beck (1980). Development methods are not reported.	11 items split into 2 sub-scales: general skills and specific CBT skills. Measured on 7-point Likert scales providing a score out of 66.	Each videotape was rated by randomly assigned pairs of raters in a balanced design Ratings were made independently.	Inter-rater reliability: Pearson's was .94 and the ICC was .96. Internal consistency: Cronbach's alpha was α = .95 Correlations between items and total score: mean Cronbach's alpha was α = .72	The CTS seems to measure one construct (CBT competence) and was used reliably.
Haddock *et al.,* (2001)	CTS-Psy Measures: competence in CBT for psychosis	Developed by the authors of the paper. The CTS (Young & Beck, 1980) was modified based on author experience. After piloting, irrelevant items were removed. Authors used the scale and discussed before agreeing the final version.	10 items split into 2 sub-scales: general skills (5 items) and technical skills (5 items). Items rated on a 7-point Likert scale to provide a total score out of 60.	5 audio tapes (selected by an independent CTS-Psy rater from a larger sample) were rated independently by each of the 4 raters.	Inter-rater reliability: ICCs excellent for the general subscale (.95), the technical subscale (.80) and the total scale (.94). Face and content validity: good as judged by a ranger of mental health professionals with skills in CBT for psychosis.	The CTS-Psy is easy to use and can reliably measure therapist competence in CBT for psychosis. It allows raters to reliably distinguish between general and specific skills. The scale needs further validation.
McLeod *et al.,* (2018)	CBAY-C Measures: competence	Developed by the authors in 4 steps:	25 items split into 4 categories: standard interventions (5 items), model specific	Videotapes were randomly assigned to raters who used	Inter-rater reliability: ICC for the whole scale was .67 (good), and 20/25 items fell	The CBAY-C can be used reliably to measure

Study	Scale and constructs measured	Scale development	Final scale characteristics (CBT items only)	Methods used to assess validity and reliability	Analysis and results	Conclusions
	in CBT for youth anxiety	1.Scale based on existing measures and developers. 2.Items developed by consulting manuals, measures and reviews. 3.Scoring strategy based on previous measures. 4.Scoring manual developed and piloted to refine it.	interventions (12 items), how the intervention was delivered (6 items) and overall skill and responsiveness (2 global items). Items rated on a 7-point Likert scale.	the scale to determine the competence displayed by the therapist.	between the good-excellent range. Construct validity: Correlation analysis showed that there was moderate overlap between CBAY-C items and items from adherence measures. Overall scores support that construct validity of the CBAY-C but indicate that the 2 global items might not be necessary.	competence in CBT for youth anxiety. Global items should be combined to provide one score for overall competence.
Roth (2016)	UCL scales Measures: CBT and generic therapeutic competence	Based on a pre-existing CBT competence framework. The author identified items for the scales which were amended based on feedback from clinicians during pilot studies on the scale.	26 item CBT scale with 4 sections: CBT techniques, change techniques based on experiential methods, change techniques for specific conditions, and review. Measured on a 5-point Likert scale.	6 raters watched videotapes of therapy sessions to assess inter-rater reliability. Described fully in Roth *et al.,* (2019)	Described in Roth *et al.,* (2019)	Psychometric properties should be assessed as the next stage of development.

Study	Scale and constructs measured	Scale development	Final scale characteristics (CBT items only)	Methods used to assess validity and reliability	Analysis and results	Conclusions
Roth *et al.,* (2019)	UCL scales Measures: CBT and generic therapeutic competence	Described in Roth (2016)	Described in Roth (2016)	6 raters used the CTS-R and the UCL scales to rate 25 videotapes. Raters provided reasons for each rating so that the decision-making process could be assessed.	Inter-rater reliability: ICC for the UCL CBT scale was .39, suggesting poor reliability. ICC for the UCL generic scale was .27, suggesting poor reliability. ICC for the CTS-R was .42, suggesting poor reliability.	There are significant difficulties with getting high levels of reliability. These scales must be triangulated with other measures.
Stallard *et al.,* (2014)	CBTS-CYP Measures: competence in CBT for children and young people.	Developed by the authors, 27 child CBT therapists, 18 CBT trainees and 16 clinical psychology trainees. Based on the CTS-R (Blackburn *et al.,* 2001). Items added based on a literature review. 27 therapists used the scale and gave feedback. Final scale was discussed with CBT and clinical psychology trainees.	14 item scale including CTS-R items plus items specific to CBT with children and young people. Items included: assessment, formulation, emotions, general skills and investigation. Items rated on a 7-point Likert scale to provide a total score out of 84, which can be transformed into a percentage.	Two studies: 1. 12 raters rated a video of a CBT session with an adolescent. Used both the CBTS-CYP and the CTS-R. 2. 12 raters irated 48 audio recordings of CYP-IAPT CBT. Used both the CBTS-CYP and the CTS-R.	Inter-rater reliability: ICC for the whole CBTS-CYP was .96 suggesting excellent reliability. Convergent validity: Pearson coefficients were high between the CBTS-CYP and the CTS-R in study 1 (r= .98) and study 2 (r= .93). Discriminative ability: agreement on competence between the CBTS-CYP and CTS-R was 77% and the scale picked up on trainee improvements.	The CBTS-CYP compares well with the CTS-R and can be reliably used to assess competence. Training in the use of the CBTS-CYP is required to ensure consistency between raters. Competence must be clearly defined when evaluating therapists.

Study	Scale and constructs measured	Scale development	Final scale characteristics (CBT items only)	Methods used to assess validity and reliability	Analysis and results	Conclusions
Vallis *et al.,* (1986)	CTS Measures: competence in CT.	Developed by Young and Beck (1980). Development methods are not reported.	11 items split into 2 sub-scales: general skills and specific CBT skills. Measured on 7-point Likert scales providing a score out of 66.	5 raters used CTS on 10 videotapes randomly chosen from a sample of 94 for inter-rater reliability. Raters then paired to rate a further 90 tapes from a sample of 725 for internal consistency. Finally, 53 tapes from the same sample of 725 were rated with the CTS and given an overall competence rating.	Inter-rater reliability: ICC for a single rater was .59. A one-way ANOVA showed a significant difference in CTS scores across tapes. Internal consistency: items correlated moderately – highly with the subscale/total score and subscales were highly correlated (*r*= .85). Discriminant validity: a MANOVA and t-tests showed that the scale can discriminate between competent and less competent CT.	The CTS can be used with moderate reliability. However, it was reliably used to rate a heterogenous sample of therapists. High correlation between subscales suggests homogeneity and thus 2 distinct scales are unnecessary. The CTS is not a highly differentiated scale.
Von Consbruch *al.,* (2012)	CTCS-SP Measures: competence in CT for social phobia.	Developed by Clark *et al.* (2007). The CTS was adapted to become the CTCS-SP. Developers added items related to treatment of anxiety disorders, as well as an	16 items e.g. agenda setting, pacing, focus on cognitions, rationale, change strategies, and experiential techniques. 2 items to assess overall competence	The first author of the paper rated all submitted videotapes using the CTCS-SP, and other raters only assessed some of the tapes.	Inter-rater reliability: ICC high for total CTCS-SP score (.73 - .88), and between the first author scores and other raters (.62 - .92). Internal consistency: Cronbach's alpha ranged from α = .82 to α = .92	The CTCS-SP can be used reliably to assess competence. Multidimensional methods of assessing competence (e.g. a scale with items

Study	Scale and constructs measured	Scale development	Final scale characteristics (CBT items only)	Methods used to assess validity and reliability	Analysis and results	Conclusions
		item for general therapeutic skill.	and degree of difficulty of working with the patient. Items rated on a 7-point Likert scale.		Retest reliability: was calculated by re-assessing 15 videos from the original sample. Retest reliability was high for the total score on the CTCS-SP (.92).	covering various skills) are necessary to obtain high levels of reliability.

Key for summary tables

CAS-CBT: Competence and Adherence Scale for Cognitive Behaviour Therapy; CBAY-C: Cognitive Behaviour Treatment for Anxiety in Youth Competence Scale; CBT: cognitive behavioural therapy; CBT-AS: Cognitive Behavioural Therapy Adherence Scale; CBTS-CYP: Cognitive Behaviour Therapy Scale for Children and Young People; CM: Clinical Management; CSPRS: Collaborative Study Psychotherapy Rating Scale; CTACS: Cognitive Therapy Adherence and Competence Scale; CT: cognitive therapy; CTCS-SP: Cognitive Therapy Competence Scale for Social Phobia; CTS: Cognitive Therapy Scale; CTS-Psy: Cognitive Therapy Scale for Psychosis; CTS-R: Cognitive Therapy Scale – Revised; CYP-IAPT: Improving Access to Psychological Therapies Programme for Children and Young People; GDC: Group Drug Counselling; ICC: Intraclass Correlation Coefficient; IDC: Individual Drug Counselling; ITIS: Inventory of Therapeutic Interventions and Skills; NIDA CCTS: National Institute on Drug Abuse Collaborative Cocaine Treatment Study; NIMH TDCRS: National Institute of Mental Health Treatment of Depression Collaborative Research Study; OCD; Obsessive-compulsive disorder; PTSD: post-traumatic stress disorder; RCT: randomised controlled trial; SE: Supportive-expressive therapy; SOPHO-NET: Social Phobia Psychotherapy Research Network; TSF: Twelve-Step Facilitation

Appendix F: Summary of quality appraisal using a bespoke tool based on QAREL and the Downs and Black (1998) checklist for measuring study quality

First Author (Date)		Barber (2003)	Bjaastad (2016)	Blackburn (2001)	Boyle (2019)	Carroll (2000)	Dobson (1985)	Haddock (2001)	McLeod (2018)	Roth (2016)	Roth (2019)	Stallard (2014)	Vallis (1986)	Von Consbruch (2012)
Introduction	Aims described?	√	√	√	√	√	√	√	√	√	√	√	√	√
	Aims related to scale development or validation?	√	√	√	√	√	√	√	√	√	√	√	√	√
	Scale designed to rate CBT?	√	√	√	√	√	√	√	√	√	√	√	√	√
Methods	Is scale development method described?	√	√	√	√	√	X	√	√	√	P	√	U	√
	Developer characteristics described?	√	√	√	X	X	U	X	P	P	P	√	U	U
	Did developers represent a specific population?	√	√	√	U	U	P	P	√	U	U	√	U	U
	Scale described?	√	√	√	√	√	√	√	√	√	P	√	√	√
	Are reliability/ validity methods described?	√	√	√	√	√	√	√	√	P	√	√	√	√
	Reliability/validity methods suitable?	√	√	√	√	√	√	√	√	U	√	√	√	√

	Was the sample appropriate to apply the scale to?	√	√	√	√	U	√	√	√	√	√	√	√	√
	Sample described?	√	√	√	√	P	√	√	√	X	√	√	P	P
	Were raters used who scale was designed for use by?	√	√	√	P	√	√	√	U	U	√	√	√	U
	Raters described?	√	√	P	√	√	√	√	√	P	√	P	√	√
	Raters blind to each other's findings?	U	U	U	√	√	√	√	√	U	√	U	U	√
	Data analysis described?	√	√	√	√	√	√	√	√	X	√	√	√	√
	Data analysis suitable?	√	√	√	√	√	√	√	√	U	√	√	√	√
Results and discussion	Main findings described?	√	√	√	√	√	√	√	√	√	√	√	√	√
	Conclusions described?	√	√	√	√	√	√	√	√	√	√	√	√	√
	Strengths and limitations described?	√	√	√	√	√	√	√	√	X	√	√	P	√
	Implications described	√	√	√	√	√	√	√	√	P	√	√	√	√
Total score out of 20		**19**	**19**	**18**	**17**	**16**	**17**	**18**	**18**	**8**	**16**	**18**	**14**	**16**

√ = YES, X = NO, U = UNCLEAR, P = PARTIALLY

Appendix G: Extract of the Compassion Focused Therapy Therapist Competence Rating Scale (CFT-TCRS, Horwood *et al.,* 2019)*

CFT Unique Therapy Skills

ITEM 1: Psychoeducation

The therapist provides CFT focused psycho-education. The therapist demonstrates an understanding of and is able to convey to the client how the human brain has evolved with built-in biases and problems that make humans very susceptible to harmful behaviours/reactions to ourselves and others.

These points should be considered when scoring:

- **The therapist discusses key concepts such as old-new brain loops.**
- **Therapist shows appropriateness of delivery. (e.g. their understanding of the model, understanding of the client).**
- **The therapist demonstrates skilfulness in the methods used (appropriate reflections, uses clients own experience, supports client to reflect on own experience, manages client's responses).**

Unable to rate: X

Absent or inappropriate				**Skilful enactment**
0	**1**	**2**	**3**	**4**

Less competent	**More competent**
Absence or highly inappropriate discussion about psychoeducation. The therapist shows little understanding of the key concepts. The therapist shows little skill in responding to the client's own experiences.	The therapist shows a good understanding of the key psycho-education concepts. The therapist uses CFT material to make appropriate links with the client's key issues. The therapist ensures that the client understands the material being discussed. The psycho-education material is used to help the client move forward.

ITEM 2: Recognising motives and emotions

The therapist helps the client to distinguish between motives and emotions that can be categorised as threat-focused, drive-reward focused and soothing-contentment focused and their evolved functions.

These points should be considered when scoring:

- **The three-circles model is correctly understood and explained.**
- **Skilful and appropriate feedback is given.**
- **The content is delivered alongside reflection, guided discovery, summarising.**

Unable to rate: X

Absent or inappropriate				Skilful enactment
0	1	2	3	4

Less competent

The therapist does not to make reference to the three-circle model, uses inappropriate feedback and makes no links between theory and client's experience.

More competent

The therapist appropriately explores the three-circle model and uses this to help the client understand their experience and move the client forward in therapy. The therapist relates the three-circles model to examples in the client's life.

ITEM 4: Understanding the relationship between three systems

The therapist helps the client to understand the relationship between their threat, drive, affiliative soothing system. E.g. they are able to use their affiliative soothing to regulate their threat system. This is used to manage the client's distress.

These points should be considered when scoring:

- **The therapist demonstrates knowledge and understanding about the balance and interplay between the three systems.**
- **The therapist collaboratively works with the client to help understand the relationship between their three-systems and helps the client understand how their systems work (e.g. uses Socratic questions, guided discovery, CFT-psychoeducation)**
- **Appropriate examples are provided and linked with the client's experiences and three-systems (e.g. how we might regulate threat if we have underdeveloped soothing system)**

Unable to rate: X

Absent or inappropriate				**Skilful enactment**
0	**1**	**2**	**3**	**4**

Less competent	**More competent**
The therapist does not show an understanding about the relationship between the three systems. The therapist describes information that is not relevant or didactically teaches them without checking their understanding. The therapist offers generalisations that do not fit with the client's presenting issues.	The therapist demonstrates a thorough knowledge of the relationship between the three systems. The therapist collaboratively engages the client to help them understand their own interplay between their systems. The therapist uses appropriate and meaningful examples to instruct the client in using their soothing system to regulate threat.

ITEM 6: Building motivation

The therapist helps the client to build their compassionate motivational system. E.G. the therapist provides CFT psychoeducation, guided discovery and skills training to develop the compassionate mind. The therapist helps the client to develop their motivation to offer compassion to themselves and others and to receive compassion.

These points should be considered when scoring:

- **The therapist helps the client explore the way their current compassionate mind works.**
- **The therapist helps the client explore how compassionate motivation could help the client engage with and alleviate/prevent the suffering of themselves or others.**
- **The therapist helps the client to explore how the compassionate mind of others may be helpful for the client.**

Unable to rate: X

Absent or inappropriate				**Skilful enactment**
0	**1**	**2**	**3**	**4**

Less competent	**More competent**
The therapist shows no recognition of the client's current compassionate motivations. The therapist does not help the client develop their compassionate motivations.	The therapist helps the client to recognise and reflect on their current compassionate motivations and how these may be helpful to the client and others. The therapist helps the client to develop compassionate motivations within a CFT framework. The therapist encourages in session and out of session practice and reflection on the development of compassionate motivations.

ITEM 10: Functional analysis

The therapist is able to help the client functionally analyse the forms and functions of safety behaviours. E.G. the forms and functions of self-criticism or shame and how these link to safety strategies.

These points should be considered when scoring:

- **The therapist supports the client to understand their safety strategy and the therapist elicits and explores the function it has served.**
- **The therapist skilfully facilitates exploration with the client about their fear of removing a safety strategy (e.g. self-criticism) to establish its function.**
- **The therapist uses normalising, validation, 'it is not your fault', common humanity, and understanding complex brain processes to work with the client's self-criticism/self-attack toward their safety strategy if necessary.**
- **The therapist explores the intended and the unintended consequences of the safety strategy.**
- **Therapist can distinguish between shame and guilt and checks the client's understanding.**

Unable to rate: X

Absent or inappropriate				**Skilful enactment**
0	**1**	**2**	**3**	**4**

Less competent — **More competent**

The therapist does not help the client understand the functions of their safety behaviours.

The therapist helps the client identify their safety strategies. The therapist explores the functions of their safety strategies and their intended and unintended consequences. If appropriate they link this back to a wider formulation. The therapist addresses self-criticism/shame in relation to safety strategies and their consequences.

ITEM 11: Fears, blocks and resistances.

The therapist helps the client to recognise, understand and work with any fears, blocks and resistances to compassionate motives and emotions and change.

These points should be considered when scoring:

- **The therapist notices and helps the client notice fears, blocks and resistances as they arise in therapy.**
- **The therapist skilfully explores the blocks to compassion and discusses with the client the function and nature of these blocks (e.g. for self-protection) and addresses any shame and self-criticism the client may have in relation to these.**
- **The therapist helps the client to develop skills in recognising and understanding their fears, blocks and resistances.**
- **The therapist reflects on this process with the client and helps the client to understand the function of blocks.**
- **Therapist uses skills to help the client work on fears, blocks and resistances, e.g. Socratic questioning, validation, reflective listening, allying with the defensive function of the resistance, exploring ambivalence, functional analysis, affect matching, limited emotional self-disclosure).**

Unable to rate: X

Absent or inappropriate				Skilful enactment
0	1	2	3	4

Less competent	More competent
The therapist does not recognise or address the clients fears, blocks and resistances. The therapist shows a lack of understanding about these concepts.	The therapist sensitively recognises and addresses the client's fears, blocks and resistances. They explore this with the client and address shame and self-criticism in relation to fears, blocks and resistances as necessary. The therapist helps the client develop skills in recognising and addressing their fears, blocks and resistances outside of therapy.

Appendix H: Confirmation of ethical approval*

The study was granted ethical approval by the University of Leicester on 18th February 2019. Below is the letter of approval.

University Ethics Sub-Committee for Psychology

18/02/2019

Ethics Reference: [redacted]psych&behaviour,deptof

TO:

Name of Researcher Applicant: [redacted]

Department: Psychology

Research Project Title: Appraisal and Validation of the Compassion-Focused Therapy Therapist Rating Scale (CFT-TRS)

Dear [redacted]

RE: Ethics review of Research Study application

The University Ethics Sub-Committee for Psychology has reviewed and discussed the above application.

1. Ethical opinion

The Sub-Committee grants ethical approval to the above research project on the basis described in the application form and supporting documentation, subject to the conditions specified below.

2. Summary of ethics review discussion

The Committee noted the following issues:

Good luck with your study.

3. General conditions of the ethical approval

The ethics approval is subject to the following general conditions being met prior to the start of the project:

As the Principal Investigator, you are expected to deliver the research project in accordance with the University's policies and procedures, which includes the University's Research Code of Conduct and the University's Research Ethics Policy.

If relevant, management permission or approval (gate keeper role) must be obtained from host organisation prior to the start of the study at the site concerned.

4. Reporting requirements after ethical approval

You are expected to notify the Sub-Committee about:

- Significant amendments to the project
- Serious breaches of the protocol
- Annual progress reports
- Notifying the end of the study

5. Use of application information

Details from your ethics application will be stored on the University Ethics Online System. With your permission, the Sub-Committee may wish to use parts of the application in an anonymised format for training or sharing best practice. Please let me know if you do not want the application details to be used in this manner.

Best wishes for the success of this research project.

Yours sincerely,

[redacted]

Chair

Appendix I: Participant information sheet*

Participant Information Sheet

Study title

"Appraisal and validation of the Compassion-Focused Therapy Therapist Competence Rating Scale (CFT-TCRS)"

Invitation:

You are being invited to take part in a research study that aims to assess whether the CFT-TCRS, a recently developed scale, can be used reliably to aid decision making around therapist competence in compassion-focused therapy (CFT). Before you consent to participate in this study, please read the following information which explains the purpose of the study, what your participation would involve, risks/benefits of participation, and contact details. Please ask questions if anything is unclear, or you would like further information. Take some time to decide whether you would like to take part.

What is the purpose of the study?

This study is being completed in partial fulfilment of the Doctorate in Clinical Psychology (DClinPsy) at the University of Leicester. Building on previous research into the competencies required to deliver CFT, the project has been designed as an initial exploration of the scale's suitability for use within CFT training courses. The CFT-TCRS includes items for rating both CFT-specific skills and more generic microskills (which are expected to be seen during any form of therapy). At present, there are no validated CFT therapist rating scales. To bring CFT in line with other treatment modalities which use such scales, the CFT-TCRS needs to be psychometrically evaluated. Therefore, the purpose of this study is to assess the psychometric properties of the CFT-TCRS (interrater reliability) and ascertain its usability.

Why have I been invited?

You have been selected for participation because you are considered to have significant knowledge, training and experience within CFT. Furthermore, you have been involved in the training of other clinicians in CFT or in the development of treatment protocols.

Do I have to take part?

No, your participation in the study is entirely voluntary. There will be no adverse consequences should you decide not to participate, or if you would like to withdraw at a later stage.

What would taking part involve?

You will be required to participate in the study for around 4 hours. You will not be required to travel to take part, data collection will be completed remotely and via post. All materials will be provided. You will be asked to watch a series of short, simulated CFT sessions. You will be made aware of the competencies displayed in the simulated session so that you can use the correct items on the CFT-TRS to rate the therapy in the video. You will use the CFT-TRS to make decisions around the level of competence displayed by the therapist in the videos for each of the identified competencies.

Furthermore, short semi-structured interviews will allow the researcher to understand the reasons behind your competency decisions and the usability of the scale. Interviews will be conducted via skype and recorded onto a transportable, password protected audio recording device which will be stored in a locked cabinet. The audio files will be stored on a secure computer at the University and interview transcripts will also be stored securely. Consent forms will be stored in a locked cabinet at the University. Data will be held on a password protected system at the University of Leicester for 6 years before being destroyed in line with University regulations. Anonymised, numerical data will be kept by the Compassionate Mind Foundation for potential use in further studies into the CFT-TRS.

What are the benefits of taking part?

Whilst the outcomes are unknown and there are no immediate benefits to participating, it is hoped that the study will provide indication as to whether the CFT-TRS is fit for use within therapist training. Wider benefits of participation may include:

- Contributing to research necessary for the standardisation of CFT training courses and assessment.
- Advancing the evidence base for CFT.
- Improvements to the scale, such as increasing ease of use.

Results will be shared with participants in order to inform their understanding of the CFT-TRS, which they may use in their professional work.

What are the risks of taking part?

There are no lasting risks anticipated as a result of participation in the study. However, participation will require your time. Data collected will be kept confidential and will be stored securely.

Contact details for further information

Should you have any further questions, or would like to withdraw from the study at any point, please contact:

[redacted]

Thank you for considering taking part in this study.

Appendix J: Consent form*

Consent Form

Study title: *"Appraisal and validation of the Compassion-Focused Therapy Therapist Rating Scale (CFT-TCRS)"*

Name of researcher: [redacted]

Please read each statement in the table below and initial to indicate your agreement:

Consent statement	Initials
I have read the participant information sheet for the above study.	
I have had the opportunity to consider the information, ask questions and have them answered satisfactorily.	
I understand that my participation is voluntary and I have the right to withdraw from the study at any time.	
I have been made aware of any potential risks involved with participating in this study.	
I understand that all information collected will be anonymised and stored securely.	
I understand that interviews will be conducted via Skype, recorded and stored in line with University requirements. However, this data will not be used in future research.	
Anonymous numerical data will be kept for use in future research.	
I have been provided with contact details for both the principal researcher and their field supervisor, should I have any other questions or wish to withdraw.	
I agree to take part in the above study.	

Participant:

Name	
Signature	
Date	

Person obtaining consent:

Name	
Signature	
Date	

Appendix K: Excerpt from the data collection pack*

CFT-TCRS Study

Thank you for agreeing to take part in this study. In this pack, you will find:

1. Participant information sheet
2. Consent form
3. CFT-TCRS items which correspond with YouTube videos
4. Pages with prompts to make notes
5. Addressed and stamped return envelope

General Instructions – please complete this study in one session

1. Carefully read the participant information sheet.
2. Complete the consent form if you are happy to take part.
3. There are ten videos to watch and rate, the YouTube links to these videos have been sent to you in an email titled 'Videos for CFT Study'
4. For each video you will have to familiarise yourself with the relevant scale items, paying attention to the 'points to consider when scoring' and the scale anchors.
5. All videos will be rated on one item, apart from 'Video 1' and 'Video 8' which will be rated on two.
6. Whilst watching each video, please feel free to make notes in the page margins of the scale item.
7. Once you have finished watching each video, please circle the competence score you feel fits best.
8. After rating the video using the scale items, please complete the questions on the notes page and make any other comments. You will need to refer to this during the short interview at the end of the study.
9. Once you have completed all 10 ratings and have made notes, please call [redacted] This will have been pre-arranged. A short semi-structured interview will be completed and recorded onto a secure audio recorder.
10. Please use the stamped envelope to return the completed consent form, scale items and notes pages. Alternatively you can scan and email these to [redacted]

Thank you for taking part in this study.

CFT Video 1

For this video there are two scale items. Please spend 2-3 minutes familiarising yourself with item 2 and item 4. Please open CFT video 1 from the email. Please watch the video and provide a rating on the two items. Following this, please complete the notes page.

ITEM 2: Recognising motives and emotions

The therapist helps the client to distinguish between motives and emotions that can be categorised as threat-focused, drive-reward focused and soothing-contentment focused and their evolved functions.

These points should be considered when scoring:

- **The three-circles model is correctly understood and explained.**
- **Skillful and appropriate feedback is given.**
- **The content is delivered alongside reflection, guided discovery, summarising.**

Unable to rate: X

Absent or inappropriate				Skillful enactment
0	1	2	3	4

Less competent **More competent**

The therapist does not to make reference to the three-circle model, uses inappropriate feedback and makes no links between theory and client's experience.

The therapist appropriately explores the three-circle model and uses this to help the client understand their experience and move the client forward in therapy. The therapist relates the three-circles model to examples in the client's life.

ITEM 4: Understanding the relationship between three systems

The therapist helps the client to understand the relationship between their threat, drive, affiliative soothing system. E.g. they are able to use their affiliative soothing to regulate their threat system. This is used to manage the client's distress.

These points should be considered when scoring:

- **The therapist demonstrates knowledge and understanding about the balance and interplay between the three systems.**
- **The therapist collaboratively works with the client to help understand the relationship between their three-systems and helps the client understand how their systems work (e.g. uses Socratic questions, guided discovery, CFT-psychoeducation)**
- **Appropriate examples are provided and linked with the client's experiences and three-systems (e.g. how we might regulate threat if we have underdeveloped soothing system)**

Unable to rate: X

Absent or inappropriate				Skillful enactment
0	1	2	3	4

Less competent **More competent**

The therapist does not show an understanding about the relationship between the three systems. The therapist describes information that is not relevant or didactically teaches them without checking their understanding. The therapist offers generalisations that do not fit with the client's presenting issues.

The therapist demonstrates a thorough knowledge of the relationship between the three systems. The therapist collaboratively engages the client to help them understand their own interplay between their systems. The therapist uses appropriate and meaningful examples to instruct the client in using their soothing system to regulate threat.

Notes for CFT Video 1

1. Why did you give the therapist the rating you did on item 2? Why did you give the therapist the rating you did on item 4?

2. Would you change anything about the behavioural anchors for item 2? And item 4?

3. How easy was it to use item 2 and its anchors? And item 4 and its anchors?

Any other comments?

Appendix L: Semi-structured interview guide*

Interview Guide

Questions 1-3 below will be asked for each of the videos watched and rated by participants in this study. The final question will be asked at the end of the interview.

Q1: What rating did you give the therapist in the video? Why did you give the therapist this rating? (purpose of this question is to draw out the reasons behind the decision made).

PROMPTS:

- What would the therapist needed to have done to get a higher rating?

Q2: Would you include anything else in the list of behavioural anchors for this item? (purpose of this question is to ascertain whether the scale covers the skills/strategies experts would expect to see from competent/less competent delivery)

PROMPTS:

- Did you notice anything the therapist did well which wasn't included in the anchors?

Q3: How easy was it to use the item and anchors? (purpose of this question is to work out if the language needs changing)

PROMPTS:

- Would you change any of the language used? What would you change?

Final Q: Overall, was there anything that worked well or that you liked about the scale? Overall, is there anything that could be improved or changed?

Appendix M: Intraclass Correlation Coefficient (ICC) model, definition and type

McGraw and Wong (1996) detailed 10 different forms of ICC which can be used in reliability research. Each of these types of analysis depends on which 'model', 'type' and 'definition' is required. Koo and Li (2016) describe ICC selection in detail and suggest the following:

Model: Researchers can select from one of three models. A *one-way random-effects* model where each subject is rated by a different group of raters, a *two-way random-effects* model should be used when raters are randomly selected to represent a larger population, or a *two-way mixed-effects* model which should be utilised when the raters in the study are the only raters of interest.

Type: There are two types of ICC which can be selected from following model definition. If the researcher is interested in the reliability of the mean rating of several raters they should use the *mean of k raters* type. If they are interested in the reliability of a single, average rater they should use the *single rater* type.

Definition: Researchers must finally define the relationship they are interested in. If the researcher is interested in whether different raters give similar scores to a subject they should choose the *absolute agreement* definition. If they are interested in consistency between raters then they should select the *consistency* definition.

Appendix N: SPSS outputs for ICC with the consistency definition for both mean of *k* raters and a single rater

Below is the SPSS output from the ICC with the *consistency* definition. As can be seen, the results are very similar to the results of the ICC using the *absolute agreement* definition which is presented in the main body of the report. The ICC with *absolute agreement* found that for average measures ICC was 'good' (.87) and for a single rater it was 'poor' (.42).

Intraclass Correlation Coefficient

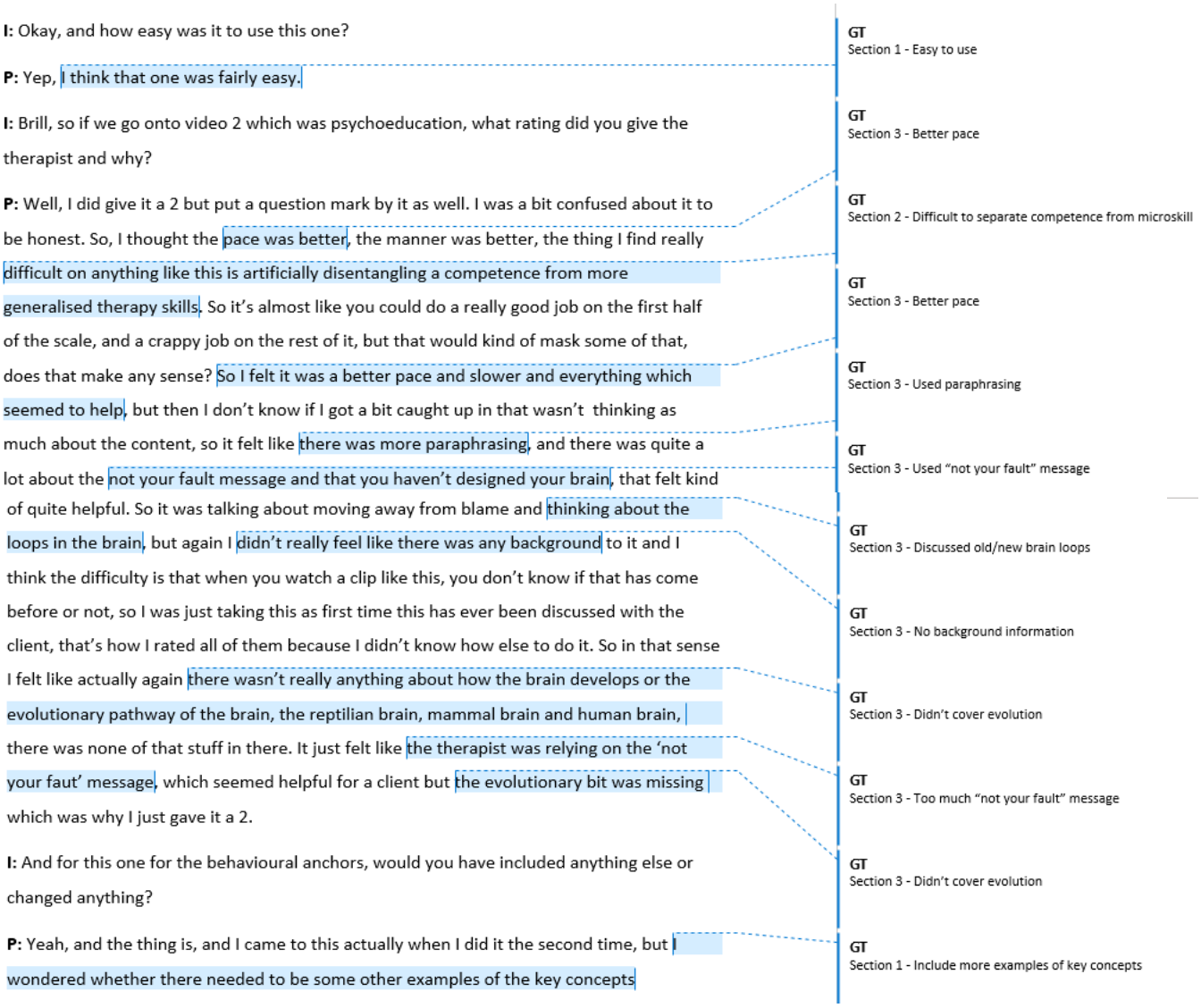
	Intraclass Correlation[b]	95% Confidence Interval		F Test with True Value 0			
		Lower Bound	Upper Bound	Value	df1	df2	Sig
Single Measures	.431[a]	.228	.712	7.827	11	88	.000
Average Measures	.872	.726	.957	7.827	11	88	.000

Two-way random effects model where both people effects and measures effects are random.

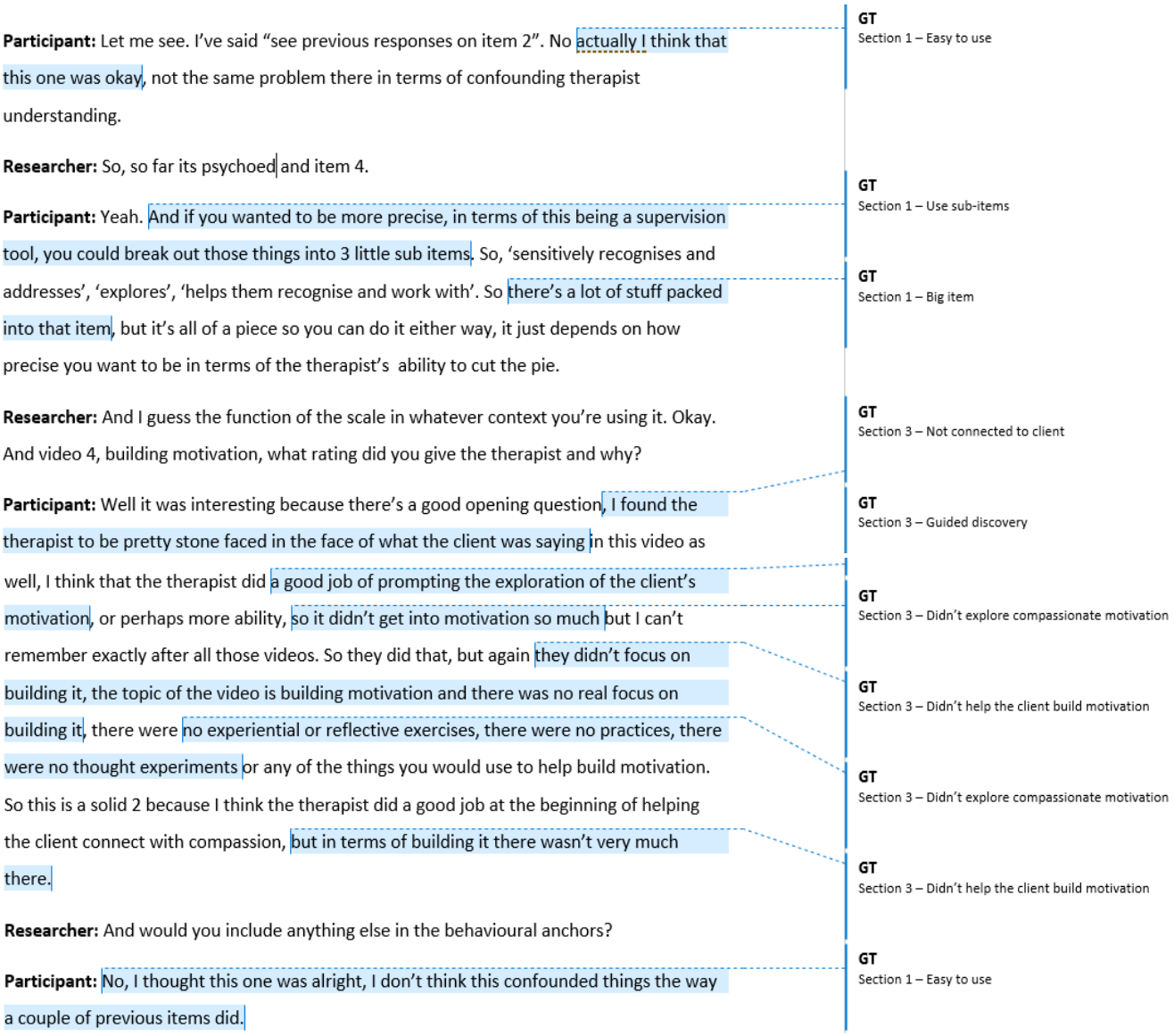
a. The estimator is the same, whether the interaction effect is present or not.

b. Type C intraclass correlation coefficients using a consistency definition. The between-measure variance is excluded from the denominator variance.

Appendix O: Excerpt of interview with initial coding*



Appendix P: Excerpt of interview with higher level coding*



Appendix Q: Content analysis – coding and categorisation process

Data familiarisation

The researcher conducted, transcribed, reviewed and analysed each of the interviews meaning that they were very familiar with the data. During this initial process, but before formal analysis of data, the researcher made noted some possible categories.

Generating initial codes

Following immersion in the data, the researcher assigned initial codes to the data for each of the distinct sections of the content analysis. This was done by hand, using different colours to signify the three sections of analysis. Coding was extensive to ensure all relevant data was included. Throughout this stage the researcher continued to note down possible categories.

Generating higher level codes

Once initial coding had been completed, the researcher printed the codes and began to group them into higher level codes which encompassed similar initial codes. Once higher-level codes were identified, the researcher re-coded transcripts using these new, more inclusive codes. Again, this was done for each of the three sections of the analysis.

Grouping codes into categories

For each section of analysis, higher-level codes were printed, and the researcher grouped them into categories. For part 1 of the content analysis which addressed changes to the items of the CFT-TCRS, codes were grouped into categories for each of the items. For part 2, which explored feedback on the overall scale, all relevant higher-level codes were grouped into categories. For part 3 which considered aspects of participant decision making, codes were grouped for each video/item. Categories were reviewed on an ongoing basis.

Appendix R: Statement of epistemological position*

Throughout both the literature review and the empirical report, the researcher adopted a 'critical-realist' position. Critical realism, which emerged in the 1970s, is based on the ideas of the philosopher Bhaskar (e.g., 1975) and other social theorists (e.g., Archer *et al.,* 1998) and offers an alternative to positivism and constructionism (Denzin & Lincoln, 2011). In proposes that human knowledge is limited and does not capture all of 'reality' (Bhaskar, 1978). Critical realism acknowledges the real social world which we attempt to understand through social science, but also recognises that some knowledge is closer to the reality of this social world than other knowledge (Danermark *et al.,* 2002).

By taking a critical-realist stance, the researcher considered the development of therapist competence scales and their psychometric properties as realities that could be accessed and thus measured. However, they acknowledged the human element to these measures, in that they are understood through interpretation (Fletcher, 2017). They considered how this might impact on what might constitute an acceptable measure of competence or what might be considered as sufficient evidence of a psychometric property. During interviews and analysis the researcher acknowledged that their own and the participants' ideas will have been impacted by their cultural and social contexts and considered how this might limit the understanding of the data to the context of the study, the participants and the researcher.

References

Archer, M., Bhaskar, R., Collier, A. et al. (1998). *Critical Realism: Essential Readings.* London: Routledge.

Bhaskar, R. (1975). *A Realist Theory of Science.* Leeds, UK: Leeds Books.

Danermark, B., Ekström, M., Jakobsen, L., & Karlsson, J. C. (2002). *Explaining society: An introduction to critical realism in the social sciences*. London: Routledge.

Denzin, N. K. & Lincoln, Y. S. (2011). The Sage handbook of qualitative research. Thousand Oaks, CA: Sage.

Fletcher, A. J. (2017). Applying critical realism in qualitative research: methodology meets method. International Journal of Social Research Methodology, 20(2), 181-194.

Appendix S: Quality assurance of content analysis*

Quality issues were considered throughout the research process. However, the following discussion relates to the qualitative content analysis. Content analysis is regarded as a flexible method which can be used to analyse qualitative data (Cavanagh, 1997), but the specific approach is dependent on the research topic and the researcher's theoretical interests (Weber, 1990). It is important to acknowledge that the flexibility of the approach can inhibit the use of content analysis due to a lack of a definition and specific procedures (Tesch, 1990). Due to this lack of specificity with the content analysis approach, it is essential that issues of quality are considered, and processes of quality assurance are undertaken. To ensure good practice in relation to aspects of quality assurance (reflexivity, transferability, credibility, and verification) a range of sources have been drawn upon (Elliot *et al.,* 1999; Morrow, 2005).

Reflexivity refers to the process of systematically attending to the context in which knowledge is constructed and considering the impact of the researcher (Malterud, 2001). Reflexivity was engaged in throughout the research. For example, the principal researcher detailed their epistemological position (see Appendix O) and completed a reflective diary (see Appendix T). This was kept regularly throughout the research process and supported reflexivity, especially during data collection and analysis. These ongoing practices were particularly useful given the researcher's lack of experience in completing qualitative research.

Transferability of the findings are discussed in the limitations section of the discussion. They acknowledge that the generalisability of the results is limited to the population from which the sample was drawn and may not be generalisable beyond the sample due to methodological issues. To ensure that the content analysis and results were credible, the principal researcher engaged in regular discussions and reviews with their research supervisor who was able to comment on the quality and relevance of the work. Further, the project was applied research, in that it sought to address practical issues, and regular supervision enabled the researcher to hold this in mind during analysis.

Appendix Q details the methodology of the content analysis from coding to categorisation.

Finally, verification of the results was considered to ensure that the content analysis was not overly subjective. Again, this was addressed through regular discussions between the researcher and their supervisor. Additionally, the process of analysis was detailed (Appendix R) and the researcher provided excerpts of their analysis to ensure transparency (see Appendix P and Q). It is hoped that the tables detailing the codes, categories and direct quotes from participants (see Appendix V, W and X) show how the findings are grounded in the qualitative data and not overly influenced by the preconceptions of the researcher.

It is hoped that attending to and acknowledging issues around quality assurance has reduced bias during the content analysis process. As stated by Malterud (2001; p.484), 'preconceptions are not the same as bias, unless the researcher fails to mention them'.

References

Cavanagh, S. (1997). Content analysis: concepts, methods and applications. *Nurse Researcher, 4*(3), 5-16.

Elliott, R., Fisher, C. T., Rennie, D. L. (1999). Evolving guidelines for publication for qualitative research studies in psychology and related fields. *British Journal of Clinical Psychology, 97*, 483-498.

Malterud, K. (2001). "Qualitative research: Standards, challenges and guidelines. The Lancet 358: pp. 483-488.

Morrow, S. L. (2005). Quality and trustworthiness in qualitative research in counseling psychology. *Journal of Counseling Psychology, 52*(2), 250-260.

Tesch, R. (1990). *Qualitative research: Analysis types and software tools*. Bristol, PA: Falmer.

Weber, R. P. (1990). *Basic content analysis*. Beverly Hills, CA: Sage.

Appendix T: Sample extract from reflexive research diary *

The following is an extract from the researcher's diary to demonstrate the process of reflexivity. The entry was made whilst the researcher was conducting and transcribing interviews with participants.

"I'm noticing myself becoming somewhat defensive of the scale when participants are offering ideas on how it could be improved, almost feeling personally responsible for the scale. I think some of this feeling comes from perceived pressure to 'get things right' and make a helpful contribution to the CFT evidence base. Although I don't explicitly express these thoughts, it feels important to note and I wonder how they might impact the analysis of my data. What I know is that I should try to maintain this awareness of the potential biases which could impact on my 'sense making' of the data. Perhaps I should write a list of the thoughts and feelings I've had which could potentially skew the way I interpret the interview transcripts. I'm reminding myself that my work is supervised, that the purpose of this research is to advance the CFT evidence base and to check in with the ways in which validity and reliability can be maintained during the qualitative research process, especially as this is the first time I've conducted qualitative analysis"

This extract highlights the researcher's perceived lack of expertise around conducting qualitative analysis, and the process of developing awareness of their expectations of themself, the data collection and analysis process, and the final research report. Reflexivity allowed the researcher to step away from their feelings enough to complete the qualitative analysis, but not so much that the analysis became a purely objective process.

Appendix U: Internal consistency analysis and results

Internal consistency was calculated to indicate whether the items of the CFT-TCRS included in the study measured the same construct. In line with Nunnally and Bernstein (1994) a Cronbach's alpha between .70 and .90 was considered 'good'. However, due to the small sample size and because the whole scale was not analysed, results were cautiously interpreted. The internal consistency of the six items of the CFT-TCRS included in this study was 'good' (Cronbach's alpha = .87). See below for the SPSS output from the internal consistency analysis.

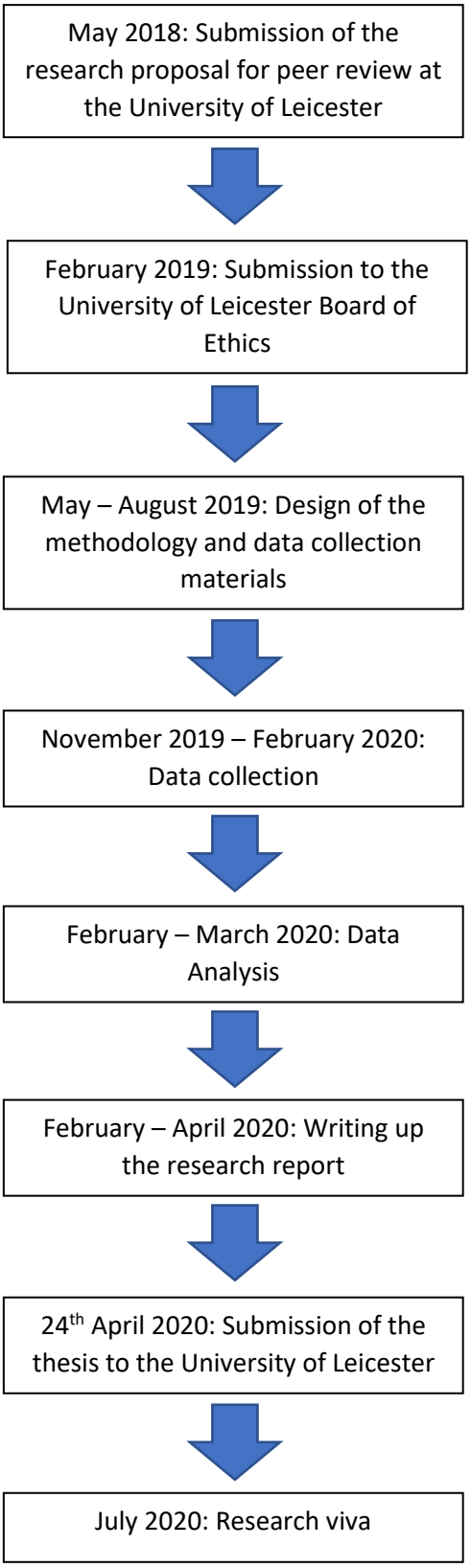
Reliability Statistics

Cronbach's Alpha	N of Items
.872	9

References

Nunnally, J.C. and Bernstein, I.H. (1994). *The Assessment of Reliability*. Psychometric Theory, 3, 248-29

Appendix V: Chronology of the research process*



Appendix W: Guidelines for authors for target journal for the systematic review and empirical project*

Target journal: Psychology and Psychotherapy: Theory, Research and Practice

Excerpt from 'Author Guidelines' retrieved on 14th April 2020 from:

https://onlinelibrary.wiley.com/page/journal/20448341/homepage/forauthors.html

1. SUBMISSION

Authors should kindly note that submission implies that the content has not been published or submitted for publication elsewhere except as a brief abstract in the proceedings of a scientific meeting or symposium.

Once the submission materials have been prepared in accordance with the Author Guidelines, manuscripts should be submitted online at http://www.editorialmanager.com/paptrap

Click here for more details on how to use **Editorial Manager**.

All papers published in the *Psychology and Psychotherapy: Theory Research and Practice* are eligible for Panel A: Psychology, Psychiatry and Neuroscience in the Research Excellence Framework (REF).

Data protection:

By submitting a manuscript to or reviewing for this publication, your name, email address, and affiliation, and other contact details the publication might require, will be used for the regular operations of the publication, including, when necessary, sharing with the publisher (Wiley) and partners for production and publication. The publication and the publisher recognize the importance of protecting the personal information collected from users in the operation of these services, and have practices in place to ensure that steps are taken to maintain the security, integrity, and privacy of the personal data collected and processed. You can learn more at **https://authorservices.wiley.com/statements/data-protection-policy.html**.

Preprint policy:

This journal will consider for review articles previously available as preprints. Authors may also post the submitted version of a manuscript to a preprint server at any time. Authors are requested to update any pre-publication versions with a link to the final published article.

2. AIMS AND SCOPE

Psychology and Psychotherapy: Theory Research and Practice is an international scientific journal with a focus on the psychological aspects of mental health difficulties and well-being; and psychological problems and their psychological treatments. We welcome submissions from mental health professionals and researchers from all relevant professional backgrounds. The Journal welcomes submissions of original high quality empirical research and rigorous theoretical papers of any theoretical provenance provided they have a bearing upon vulnerability to, adjustment to, assessment of, and recovery (assisted or otherwise) from psychological disorders. Submission of systematic reviews and other research reports which support evidence-based practice are also welcomed, as are relevant high quality analogue studies and Registered Reports. The Journal thus aims to promote theoretical and research developments in the understanding of cognitive and emotional factors in psychological disorders, interpersonal attitudes, behaviour and

relationships, and psychological therapies (including both process and outcome research) where mental health is concerned. Clinical or case studies will not normally be considered except where they illustrate particularly unusual forms of psychopathology or innovative forms of therapy and meet scientific criteria through appropriate use of single case experimental designs.

All papers published in *Psychology and Psychotherapy: Theory, Research and Practice* are eligible for Panel A: Psychology, Psychiatry and Neuroscience in the Research Excellence Framework (REF).

3. MANUSCRIPT CATEGORIES AND REQUIREMENTS

- Articles should adhere to the stated word limit for the particular article type. The word limit excludes the abstract, reference list, tables and figures, but includes appendices.

Word limits for specific article types are as follows:

- Research articles: 5000 words
- Qualitative papers: 6000 words
- Review papers: 6000 words
- Special Issue papers: 5000 words

In exceptional cases the Editor retains discretion to publish papers beyond this length where the clear and concise expression of the scientific content requires greater length (e.g., explanation of a new theory or a substantially new method). Authors must contact the Editor prior to submission in such a case.

Please refer to the separate guidelines for **Registered Reports**.

All systematic reviews must be pre-registered.

4. PREPARING THE SUBMISSION

Free Format Submission

Psychology and Psychotherapy: Theory, Research and Practice now offers free format submission for a simplified and streamlined submission process.

Before you submit, you will need:

- Your manuscript: this can be a single file including text, figures, and tables, or separate files – whichever you prefer. All required sections should be contained in your manuscript, including abstract, introduction, methods, results, and conclusions. Figures and tables should have legends. References may be submitted in any style or format, as long as it is consistent throughout the manuscript. If the manuscript, figures or tables are difficult for you to read, they will also be difficult for the editors and reviewers. If your manuscript is difficult to read, the editorial office may send it back to you for revision.
- The title page of the manuscript, including a data availability statement and your co-author details with affiliations. *(Why is this important? We need to keep all co-authors informed of the outcome of the peer review process.)* You may like to use this template for your title page.

Important: the journal operates a double-blind peer review policy. Please anonymise your manuscript and prepare a separate title page containing author details. *(Why is this important? We need to uphold rigorous ethical standards for the research we consider for publication.)*

- An ORCID ID, freely available at https://orcid.org. *(Why is this important? Your article, if accepted and published, will be attached to your ORCID profile. Institutions and funders are increasingly requiring authors to have ORCID IDs.)*

To submit, login at **https://www.editorialmanager.com/paptrap/default.aspx** and create a new submission. Follow the submission steps as required and submit the manuscript.

If you are invited to revise your manuscript after peer review, the journal will also request the revised manuscript to be formatted according to journal requirements as described below.

Revised Manuscript Submission

Contributions must be typed in double spacing. All sheets must be numbered.

Cover letters are not mandatory; however, they may be supplied at the author's discretion. They should be pasted into the 'Comments' box in Editorial Manager.

Parts of the Manuscript

The manuscript should be submitted in separate files: title page; main text file; figures/tables; supporting information.

Title Page

You may like to use **this template** for your title page. The title page should contain**:**

- A short informative title containing the major key words. The title should not contain abbreviations (see Wiley's best practice SEO tips);
- A short running title of less than 40 characters;
- The full names of the authors;
- The author's institutional affiliations where the work was conducted, with a footnote for the author's present address if different from where the work was conducted;
- Abstract;
- Keywords;
- Data availability statement (see Data Sharing and Data Accessibility Policy);
- Acknowledgments.

Authorship

Please refer to the journal's Authorship policy in the Editorial Policies and Ethical Considerations section for details on author listing eligibility. When entering the author names into Editorial Manager, the corresponding author will be asked to provide a CRediT contributor role to classify the role that each author played in creating the manuscript. Please see the **Project CRediT** website for a list of roles.

Abstract

Please provide an abstract of up to 250 words. Articles containing original scientific research should include the headings: Objectives, Design, Methods, Results, Conclusions. Review articles should use the headings: Purpose, Methods, Results, Conclusions.

Keywords

Please provide appropriate keywords.

Acknowledgments

Contributions from anyone who does not meet the criteria for authorship should be listed, with permission from the contributor, in an Acknowledgments section. Financial and material support should also be mentioned. Thanks to anonymous reviewers are not appropriate.

Practitioner Points

All articles must include Practitioner Points – these are 2-4 bullet point with the heading 'Practitioner Points'. They should briefly and clearly outline the relevance of your research to professional practice. (The Practitioner Points should be submitted in a separate file.)

Main Text File

As papers are double-blind peer reviewed, the main text file should not include any information that might identify the authors.

The main text file should be presented in the following order:

- Title
- Main text
- References
- Tables and figures (each complete with title and footnotes)
- Appendices (if relevant)

Supporting information should be supplied as separate files. Tables and figures can be included at the end of the main document or attached as separate files but they must be mentioned in the text.

- As papers are double-blind peer reviewed, the main text file should not include any information that might identify the authors. Please do not mention the authors' names or affiliations and always refer to any previous work in the third person.
- The journal uses British/US spelling; however, authors may submit using either option, as spelling of accepted papers is converted during the production process.

References

References should be prepared according to the *Publication Manual of the American Psychological Association* (6th edition). This means in text citations should follow the author-date method whereby the author's last name and the year of publication for the source should appear in the text, for example, (Jones, 1998). The complete reference list should appear alphabetically by name at the end of the paper. Please note that for journal articles, issue numbers are not included unless each issue in the volume begins with page 1, and a DOI should be provided for all references where available.

For more information about APA referencing style, please refer to the **APA FAQ**.

Reference examples follow:

Journal article

Beers, S. R. , & De Bellis, M. D. (2002). Neuropsychological function in children with maltreatment-related posttraumatic stress disorder. *The American Journal of Psychiatry*, *159*, 483–486. doi:**10.1176/appi.ajp.159.3.483**

Book

Bradley-Johnson, S. (1994). *Psychoeducational assessment of students who are visually impaired or blind: Infancy through high school* (2nd ed.). Austin, TX: Pro-ed.

Internet Document

Norton, R. (2006, November 4). How to train a cat to operate a light switch [Video file]. Retrieved from **http://www.youtube.com/watch?v=Vja83KLQXZs**

Tables

Tables should be self-contained and complement, not duplicate, information contained in the text. They should be supplied as editable files, not pasted as images. Legends should be concise but comprehensive – the table, legend, and footnotes must be understandable

without reference to the text. All abbreviations must be defined in footnotes. Footnote symbols: †, ‡, §, ¶, should be used (in that order) and *, **, *** should be reserved for P-values. Statistical measures such as SD or SEM should be identified in the headings.

Figures

Although authors are encouraged to send the highest-quality figures possible, for peer-review purposes, a wide variety of formats, sizes, and resolutions are accepted.

Click here for the basic figure requirements for figures submitted with manuscripts for initial peer review, as well as the more detailed post-acceptance figure requirements.

Legends should be concise but comprehensive – the figure and its legend must be understandable without reference to the text. Include definitions of any symbols used and define/explain all abbreviations and units of measurement.

Colour figures. Figures submitted in colour may be reproduced in colour online free of charge. Please note, however, that it is preferable that line figures (e.g. graphs and charts) are supplied in black and white so that they are legible if printed by a reader in black and white. If an author would prefer to have figures printed in colour in hard copies of the journal, a fee will be charged by the Publisher.

Supporting Information

Supporting information is information that is not essential to the article, but provides greater depth and background. It is hosted online and appears without editing or typesetting. It may include tables, figures, videos, datasets, etc.

Click here for Wiley's FAQs on supporting information.

Note: if data, scripts, or other artefacts used to generate the analyses presented in the paper are available via a publicly available data repository, authors should include a reference to the location of the material within their paper.

General Style Points

For guidelines on editorial style, please consult the **APA Publication Manual** published by the American Psychological Association. The following points provide general advice on formatting and style.

- **Language:** Authors must avoid the use of sexist or any other discriminatory language.
- **Abbreviations:** In general, terms should not be abbreviated unless they are used repeatedly and the abbreviation is helpful to the reader. Initially, use the word in full, followed by the abbreviation in parentheses. Thereafter use the abbreviation only.
- **Units of measurement:** Measurements should be given in SI or SI-derived units. Visit the Bureau International des Poids et Mesures (BIPM) website for more information about SI units.
- **Effect size:** In normal circumstances, effect size should be incorporated.
- **Numbers:** numbers under 10 are spelt out, except for: measurements with a unit (8mmol/l); age (6 weeks old), or lists with other numbers (11 dogs, 9 cats)

Appendix X: Checklist to ensure anonymity (coursework handbook Appendix D)*

	Checked in Executive Summary/Abstract/ Overview (if included in assignment)	Checked in main text	Checked in appendices
Pseudonym or false initials used	✓	✓	✓
Reference to pseudonym/false initials as a footnote	✓	✓	✓
Removed any reference to names of Trusts/hospitals/clinics/services (including letterhead if including letters in appendices)	✓	✓	✓
Removed any reference to names/specific dates of birth/specific date of clinical appointments/addresses/ location of client(s), participant(s), relatives, caregivers, and supervisor(s). [For research thesis – supervisors can be named in the research thesis "acknowledgements" section]	✓	✓	✓
Removed/altered references to client(s) jobs/professions/nationality where this may potentially identify them. [For research thesis – removed potential for an individual research participant to be identifiable (e.g., by a colleague of the participant who might read the thesis on the internet and be able to identify a participant using a combination of the participants specific job title, role, age, and gender)]	✓	✓	✓
Removed any information that may identify the trainee (consult with course staff if this will detract from the points the trainee is making)	✓	✓	✓
No Tippex or other method has been used to obliterate the original text – unless the paper is subsequently photocopied and the trainee has ensured that the obliterated text cannot be read	✓	✓	✓
The "find and replace" function in word processing has been used to check the assignment for use of client(s) names/other confidential information	✓	✓	✓