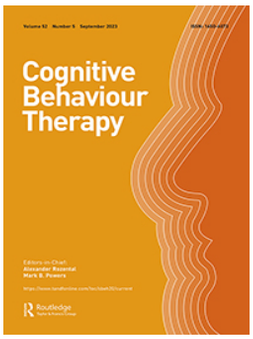
Article:
Beard, J. orcid.org/0000-0002-4330-4536, Cooper, Z. orcid.org/0000-0001-7963-656X, Masson, P. orcid.org/0009-0003-9878-0438 et al. (8 more authors) (2024) Assessing clinician competence in the delivery of cognitive-behavioural therapy for eating disorders: development of the Cognitive-Behavioural Therapy Scale for Eating Disorders (CBTS-ED). Cognitive Behaviour Therapy, 53 (1). pp. 29-47. ISSN 1650-6073

https://doi.org/10.1080/16506073.2023.2263640



Cognitive Behaviour Therapy

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/sbeh20

Assessing clinician competence in the delivery of cognitive-behavioural therapy for eating disorders: development of the Cognitive-Behavioural Therapy Scale for Eating Disorders (CBTS-ED)

Jessica Beard, Zafra Cooper, Philip Masson, Victoria A. Mountford, Rebecca Murphy, Bronwyn Raykos, Madeleine Tatham, Jennifer J. Thomas, Hannah M. Turner, Tracey D. Wade & Glenn Waller

To cite this article: Jessica Beard, Zafra Cooper, Philip Masson, Victoria A. Mountford, Rebecca Murphy, Bronwyn Raykos, Madeleine Tatham, Jennifer J. Thomas, Hannah M. Turner, Tracey D. Wade & Glenn Waller (09 Oct 2023): Assessing clinician competence in the delivery of cognitive-behavioural therapy for eating disorders: development of the Cognitive-Behavioural Therapy Scale for Eating Disorders (CBTS-ED), Cognitive Behaviour Therapy, DOI: 10.1080/16506073.2023.2263640

To link to this article: https://doi.org/10.1080/16506073.2023.2263640

COGNITIVE BEHAVIOUR THERAPY
https://doi.org/10.1080/16506073.2023.2263640

Routledge
Taylor & Francis Group

Assessing clinician competence in the delivery of cognitive-behavioural therapy for eating disorders: development of the Cognitive-Behavioural Therapy Scale for Eating Disorders (CBTS-ED)

Jessica Beard[a], Zafra Cooper[b], Philip Masson[c], Victoria A. Mountford[d,e], Rebecca Murphy[f], Bronwyn Raykos[g], Madeleine Tatham[a], Jennifer J. Thomas[h], Hannah M. Turner[i], Tracey D. Wade[j] and Glenn Waller[a]

[a]Department of Psychology, University of Sheffield, Sheffield, UK; [b]Department of Psychiatry, Yale University, New Haven, USA; [c]Department of Psychology, Western University, London, Canada; [d]Sage Clinics, Dubai, UAE; [e]Institute of Psychiatry, Psychology & Neuroscience, King's College London, London, UK; [f]Department of Psychiatry, University of Oxford, Oxford, UK; [g]Centre for Clinical Interventions, Northbridge, Western Australia; [h]Department of Psychiatry, Harvard Medical School Massachusetts General Hospital, Boston, USA; [i]Eating Disorders Service, Southern Health NHS Foundation Trust, Southampton, UK; [j]College of Education, Psychology and Social Work, Flinders University, Bedford Park, South Australia

ABSTRACT

Evidence-based cognitive-behaviour therapy for eating disorders (CBT-ED) differs from other forms of CBT for psychological disorders, making existing generic CBT measures of therapist competence inadequate for evaluating CBT-ED. This study developed and piloted the reliability of a novel measure of therapist competence in this domain—the Cognitive Behaviour Therapy Scale for Eating Disorders (CBTS-ED). Initially, a team of CBT-ED experts developed a 26-item measure, with general (i.e. present in every session) and specific (context- or case-dependent) items. To determine statistical properties of the measure, nine CBT-ED experts and eight non-experts independently observed six role-played mock CBT-ED therapy sessions, rating the therapists' performance using the CBTS-ED. The inter-item consistency (Cronbach's alpha and McDonald's omega) and inter-rater reliability (ICC) were assessed, as appropriate to the clustering of the items. The CBTS-ED demonstrated good internal consistency and moderate/good inter-rater reliability for the general items, at least comparable to existing generic CBT scales in other domains. An updated version is proposed, where five of the 16 "specific" items are reallocated to the general group. These preliminary results suggest that the CBTS-ED can be used effectively across both expert and non-expert raters, though less experienced raters might benefit from additional training in its use.

ARTICLE HISTORY
Received 27 October 2022
Accepted 20 September 2023

KEYWORDS
Eating disorders; cognitive behaviour therapy; competence; therapist skill; reliability

CONTACT Glenn Waller g.waller@sheffield.ac.uk Clinical and Applied Psychology Unit, Department of Psychology, University of Sheffield, Cathedral Court, 1 Vicar Lane, Sheffield S1 2LT, UK

Supplemental data for this article can be accessed online at https://doi.org/10.1080/16506073.2023.2263640

Introduction

Cognitive-behavioural therapy for eating disorders (CBT-ED) is one of the treatments recommended by the National Institute for Health and Care Excellence (NICE, 2017) for adults with anorexia nervosa, and is the only approach recommended for adults with binge-eating disorder, bulimia nervosa, and related presentations. When therapists deliver the treatment competently in routine clinical settings, delivering manual-based treatment under supervision, CBT-ED has demonstrated strong levels of effectiveness in a range of eating disorders (Byrne et al., 2011; Knott et al., 2015; Mountford et al., 2021; Signorini et al., 2018; Turner et al., 2015), comparable to outcomes in controlled research settings (e.g. Fairburn et al., 2009, 2015). However, as with National Institute for Health and Care Excellence, 2017 all therapies for all psychological disorders, CBT-ED is not a universal solution, as it is not effective for all patients, and the degree to which it is delivered competently is likely to influence that level of effectiveness.

Recovery and remission rates vary (e.g. Atwood & Friedman, 2020; Lampard & Sharbanee, 2015; Moberg et al., 2021; Södersten et al., 2017; Waller, 2016). In addition to client factors and overall treatment effectiveness, therapists' competence, manifesting as a lack of requisite skills in, and fidelity to, the CBT-ED model and treatment tasks, may impact outcome. It has been demonstrated that simple adherence to treatment protocols is not related to clinical outcomes, but that therapist competence and fidelity/integrity are associated with outcomes across adult psychotherapies, particularly for cognitive-behavioural therapy (Power et al., 2022). Therefore, an important element in improving the effectiveness of CBT-ED might require attending to therapists' competence, in terms of both their skills and their standard of implementation.

Competence is defined in different ways in the psychotherapy literature. For example, Waltz et al. (1993) state that competence is based on the clinician's understanding of the relevance of specific intervention methods, their skill in delivering interventions, and their ability to use those methods in a way that ensures that the potential outcomes are as expected. Cooper et al. (2015) use a similar definition, based on the combination of theoretical and applied knowledge of treatments, as well as the ability to apply it in practice. In contrast, Muse and McManus (2013, 2022) refer less overtly to outcomes, defining competence as "the degree to which a [practitioner] demonstrates the general therapeutic and treatment-specific knowledge and skills required to appropriately deliver CBT interventions". Thus, competence as defined here is not simply a matter of knowing what skills should be implemented, but delivering those techniques skilfully and appropriately.

Taking a somewhat broader view, Rodriguez-Quintana et al. (2021) summarise the necessary elements of competence as being: 1) knowledge of the empirically supported treatment; 2) knowledge about when and how to use the skills that make up that therapy; 3) belief in the value of the therapy; and 4) the actual implementation of the skills. They suggest that these four elements all need to be assessed when training therapists to be competent, though they also stress the context and setting of the therapy (e.g. being able to apply it in schools vs clinics). Clearly, without all four of these components, there is the danger that clinicians will "drift" from the competent delivery of therapy, due to their lack of knowledge about or understanding of the appropriate intervention and when it should be delivered for the individual patient, or due to reluctance to deliver specific interventions (Waller, 2009;

Waller & Turner, 2016). Therefore, competence should be regarded as involving all four of the elements identified by Rodriguez-Quintana et al. (2021).

Given this diversity in definitions of competence in psychological therapies (e.g. whether it should be measured by achievement of outcomes; distinctions between knowledge and implementation; the need to implement the skills appropriately and in context), it is perhaps unsurprising that a link between competence and outcomes is not always found in clinical research. For example, while Power et al. (2022) showed an overall association between rated competence and clinical outcomes, there was a high level of heterogeneity, indicating a substantial proportion of individual studies did not show that effect, even showing effects that were in the opposite direction (e.g. Barber et al., 2008; Gibbons et al., 2010). For example, a moderator analysis showed that competence was associated with positive clinical outcomes when using cognitive therapy and when working with anxiety disorders, but with negative clinical outcomes when using emotion-focused therapy and when working with substance use and addictions.

While the construct of therapist competence can be operationalised in the way identified by Rodriguez-Quintana et al. (2022), measurement of those elements is not equally meaningful. For example, measures of more academic knowledge would be immediately relevant to the first three of those criteria and are very scalable (e.g. Cooper et al., 2015), but they do not ensure that the clinician delivers the therapy. Similarly, clinicians could have the necessary skills, but might deliver them inappropriately (knowing how to deliver the therapy, but not using the skills as or when appropriate). Finally, clinicians might have the necessary academic knowledge, but not the belief in the value or appropriateness of the intervention (e.g. Addis & Krasnow, 2000), so that the methods are not used. These discrepancies can mean that the therapy's delivery is rendered ineffective or even harmful (Waller & Turner, 2016). Consequently, the majority of measures of therapist competence do not focus on knowledge but are based on observation of the appropriate behavioural delivery of the therapy skills (the fourth of Rodriguez-Quintana et al.'s [2022] criteria). It is implied that the ability to deliver an observable skill in the right way at the right time means that the individual clinician has an implicit or explicit knowledge base about the skills, the evidence, and the value of the therapy.

In CBT more widely, several tools have been developed to assess therapist competence using this observer-based approach, and these tools have been used in training and monitoring therapist skills. The most extensively used of these tools is the Cognitive Therapy Scale-Revised (CTS-R; Blackburn et al., 2001), developed to evaluate competence when working with anxiety and depression. The CTS-R assesses both adherence to therapy method and skill of the therapist in doing so. Raters evaluate recorded or directly observed therapy sessions, assessing general and cognitive therapy-specific skills (e.g. agenda-setting, eliciting key cognitions).

The CTS-R has important limitations (Muse & McManus, 2013). It was developed for depression and anxiety rather than other disorders, making its wider use problematic (Barber et al., 2007; Roth, 2016). It has limited and variable inter-rater reliability. For example, Loades and Armstrong (2016) found Intraclass Correlation Coefficients (ICCs) that varied from 0.40 to 0.98, with only a moderate median of 0.65 (Koo & Li, 2016). Consequently, other transdiagnostic observational measures of clinician competence have been developed (e.g. Muse & McManus, 2013; Sudak, 2015). For example, the Assessment of Core CBT Skills scale (Muse et al., 2017) has better psychometric

properties than the CTS-R and shows good validity of self-ratings relative to observer ratings. Another alternative measure is the Cognitive Behavioral Therapy Competence Scale (CCS; Rodriguez-Quintana et al., 2021202), which was developed for young people, and which also has strong psychometric properties. The Cognitive Therapy Adherence and Competence Scale (CTACS; Barber et al., 2003) adds further domains of clinician activity to those measured by the CTS-R. There have also been developments of validated alternative versions of the CTS-R, including a brief version (Alfonsson et al., 2022).

These transdiagnostic measures have limited utility for rating skills in specific disorders because competence frameworks differ from one disorder to another, though they also contain common elements (e.g. Roth & Pilling, 2007). Consequently, other CBT competence measures have been developed with specific diagnoses and populations in mind, including measures specific to: social phobia (Cognitive Therapy Competence Scale for Social Phobia [CTCS-SP] - von Consbruch et al., 2012); anxiety disorders in young people (Competence and Adherence Scale for Cognitive Behavioral Therapy [CAS-CBT] - Bjaastad et al., 2016); young people with post-traumatic stress disorder (Gutermann et al., 2015); and psychosis (Haddock et al., 2001). These disorder-specific measures have generally reported good psychometric properties, supporting their use with those disorders. To date, however, there is no comparable measure of observed competence in CBT-ED, limiting our ability to understand the link between competence and outcomes of CBT for patients with eating disorders.

The need for a disorder-specific measure of competence in CBT-ED is underpinned by the nature of CBT-ED itself. As with other measures of competence based on specific disorders, some CBT competences and metacompetences are generic enough to be relevant to eating disorders as much as to any other disorder (e.g. agenda-setting, questioning, homework-setting). However, others are more specific to the pathology and change methods of eating disorders (e.g. changes in eating patterns, body image interventions, weighing, addressing bingeing and compensatory behaviours). Existing measures such as the CTS-R are not capable of identifying the need to work with these tasks of therapy. Consequently, more generic measures yield relatively high competence scores without the therapist using any eating disorder-oriented assessment and change methods. This is problematic, as such disorder-specific methods are commonly omitted by therapists who state that they are delivering CBT-ED (e.g. Cowdrey & Waller, 2015; Mulkens et al., 2018; Wallace & von Ranson, 2011; Waller et al., 2012). Achieving the goals of improving therapist delivery of CBT-ED would be greatly assisted by effective and reliable methods of measuring competence, to support training, supervision, self-monitoring, and clinical research.

Of most importance in clinical practice is establishing that therapists have the required skills to implement CBT-ED, and that they consistently employ these skills to deliver high quality evidence-based treatment. Assessing the former requires an assessment of the capacity of the therapist, while assessing the latter requires an assessment of the quality of the actual therapy or sessions delivered (Fairburn & Cooper, 2011). Studies to date have focused on the first of these assessment targets in the field of eating disorders. For one form of CBT-ED, enhanced CBT (CBT-E; Fairburn, 2008), two measures of therapist skill have been developed. One is a highly scalable online applied knowledge measure (Cooper et al., 2015), and the other is a performance-based measure involving role-plays that can be rated by non-

experts (Cooper et al., 2017). Reaching a predefined competence cut-off on the online measure predicts similar levels of competence to the performance-based measure (Cooper et al., 2017). These measures aim to assess a wide range of skills and are potentially useful for research purposes and for assessing the outcome of scalable training. However, by their very nature, they do not assess the quality of therapy delivered, but simply the therapist's potential to deliver it, as they do not allow for direct observation of the therapist using their skills to deliver the treatment. Taking a different perspective, self-rating checklists can be used relatively easily. Serfaty et al. (2020) have highlighted the practical and economic value of measuring therapy quality in cognitive therapy for depression using such checklists, reporting promising levels of agreement between therapists' self-report and the report of an independent expert. However, while such a checklist has recently been developed for CBT-E (Bailey-Straebler et al., 2022), it has yet to be evaluated.

To summarise, therapists and supervisors need an observation-based tool to assess clinician competence in delivering CBT-ED (Mulkens & Waller, 2021; Power et al., 2022). Existing measures such as the CTS-R (Blackburn et al., 2001) and the Cognitive Therapy Adherence and Competence Scale (CTACS; Barber et al., 2003) are not suitable for that specific purpose, as they do not assess CBT-ED specific competences (Muse & McManus, 2013). This study aimed to develop a disorder-specific measure to assess CBT-ED competence directly—the CBTS-ED—and to test its preliminary psychometric properties (focusing on reliability). We also aimed to identify which items should be carried forwards into future iterations of the measure.

Method

Design

This study was designed to develop the new CBTS-ED and to test its reliability. A fully crossed design was utilised with the same set of raters rating each therapist. Recruited participants were therapists working in the field of eating disorders.

Phase 1 was the preliminary development of the CBTS-ED, a measure designed to be used across different evidence-based CBT-ED models (e.g. CBT-BN – Fairburn et al., 1993; CBT-E - Fairburn, 2008; CBT for EDs – Waller et al., 2007; CBT-T - Waller et al., 2018). This early version of the measure consisted of general CBT-ED items, relevant across all sessions and rated every time (Part A), and CBT-ED session-specific items, rated only when relevant to the session being observed (Part B). To enable this process, the team developed a set of mock CBT-ED therapy session video recordings ("vignettes").

Phase 2 aimed to assess a key psychometric property of the CBTS-ED—namely its inter-rater reliability. More specifically, we evaluated the inter-rater reliability between expert and non-expert raters (Intraclass Correlation Coefficients; ICC) and the internal consistency (Cronbach's alpha and McDonald's omega) of Part A. The inter-rater reliability (ICC) for Part B mean scores was also assessed, and we identified any Part B items that can be rated consistently depending on the session being observed. Any Part B items that might be better placed in the "general" part of the measure (Part A) were identified. It was hypothesised that the inter-rater reliability and the internal consistency of the scales would be acceptable to strong.

Ethical considerations

This study was reviewed and approved by the University of Sheffield Ethics Committee (039900).

Phase 1: development of the CBTS-ED and material for ratings

The CBTS-ED was developed by an international group of 10 clinical researchers, all of whom had extensive experience in the delivery and training/supervision of CBT-ED. The authors generated items for the measure through an iterative process of nomination, collapsing similar items into one, and reviewing and agreeing on the wording to ensure that each item reflected good CBT-ED practice. For example, open weighing of patients was nominated and agreed as a key competence, then several variants were proposed on how to word the item, these were merged into a single item, and the wording was refined until the team agreed on the version to be used. Thus, in keeping with the different forms of evidence-based CBT-ED, with different theoretical backgrounds and content, items were developed on a phenomenological basis (rather than being based on a single theoretical model), so that they would represent that range of forms of CBT-ED. Items were assumed to have equal weight at this early point, in the absence of factor loadings that might indicate otherwise.

Scoring system and final measure

The scale for scoring items was based on that used in the CTS-R, though the rating term "incompetent" was not used as it was agreed that this was not a constructive descriptor for training or supervision purposes and might deter raters from using the lowest mark on the scale. Instead, the term "no evidence" was used. Mirroring the CTS-R, each item was scored from 0–6 (any midpoint scoring was rounded down), with each score relating to a qualitative definition, agreed upon by the expert team. No effort was made to identify cut-off scores at this early stage in the CBTS-ED's development, as such cut-offs would be speculative before a subsequent validation stage.

The CBTS-ED was designed to be used as an observational instrument to assess therapy quality in the delivery of CBT-ED. The final set of 26 items is detailed in Table 1. The scoring manual and scoresheet are included as supplementary materials.

The generated items were grouped into two types. The first group ("General") were general competences that one would expect to see evidenced during every session (e.g. collaborative engagement; measurement of progress and outcomes). The second group ("Specific") consisted of specific items reflecting competences that one would expect to see in CBT-ED, but not in every session (e.g. focus on early behavioural change) or with every patient (e.g. focus on weight regain). These groups became Parts A and B of the new measure, respectively. Part A consists of 10 items, with each rated from 0 = *no evidence* to 6 = *expert*. Part B has 16 items rated 0 = *no evidence* to 6 = *expert*. As Part B items are not all relevant to rate in every session, any item in Part B can be scored as *"not relevant"*, and thus have no score assigned to it.

The overall score for Part A is the total of the relevant 10 "general" item scores, divided by 10 to give an item mean score (possible range = 0–6, where a higher score indicates

Table 1. Items of the cognitive-behavioural therapy scale for eating disorders (CBTS-ED).

Item	Description
Part A: General CBT-ED competences and meta-competences (rate all items every time)	
1	Starts session well
2	Selection and use of core assessment, progress and outcome measures
3	Enhances patient's self-efficacy
4	Engages patient in collaborative work
5	Appropriate pacing and efficient use of time
6	Clarity and honesty in communication style
7	Combines firmness and empathy
8	Weighs the patient collaboratively and openly
9	Stresses 168-hour-per-week therapy
10	Summarises sessions collaboratively
Part B: Specific CBT-ED competences and meta-competences (rate items relevant to the session being observed)	
11	Addresses essential or non-negotiable aspects of treatment
12	Focuses on early behavioural change
13	Elicits and validates patient's emotions
14	Uses individualised formulation
15	Retains or re-establishes focus on treating issues that maintain the eating disorder
16	Provides appropriate psychoeducation
17	Guidance on necessary dietary change
18	Monitor eating and other behaviour
19	Regular and sufficient eating
20	Prioritise weight regain
21	Encourages change/do things differently
22	Behaviour experiments or surveys
23	Works with different levels of cognition/belief
24	Works with emotional states
25	Addresses negative body image
26	Handling of endings and relapse prevention

stronger competence). While all Part A items should be completed, in the case of error in completion, up to two items can be missing and the scale can still be scored, using pro-rating to yield an overall item mean score. Again, for Part B ("Specific"), scoring can still be applied if up to two items are accidentally omitted in the scoring (i.e. not scored at all, rather than being rated as "not relevant"). The overall score for Part B is the item mean score for the number of items identified as relevant and scored.

Experimental material for ratings

To determine the reliability of CBTS-ED ratings, a series of videotaped role-plays, or "vignettes" were developed, designed to simulate CBT-ED sessions, for independent ratings. Whilst the CBTS-ED will need to be tested in real therapy sessions in the future, in the first instance it was important to pilot the measure to gain an understanding of the items that would remain in the updated version to test with real sessions later. It is also important to note that the video-recordings were of role-plays based on therapy sessions that spanned different stages of CBT-ED (sessions 2–16), as different competences are more likely to be used at different points in the therapy (e.g. body image work, which tends to come later in CBT-ED). Therefore, it was not expected that all role-plays would encompass every Part B competences, though the competences in Part A should be present in all sessions.

Three experienced eating disorders therapists role-played CBT-ED therapists, and two actors (not therapists) role-played adult patients with different eating disorder presentations (e.g. atypical anorexia nervosa and binge-eating disorder), yielding six videos/

vignettes as experimental materials (each of approximately 50 minutes). The vignettes were filmed remotely over an online platform (due to COVID restrictions in place at the time). The video recordings are available on request from the authors.

Phase 2: reliability of the CBTS-ED

Following the development of the CBTS-ED and the experimental materials, a group of expert and non-expert therapists were asked to rate all six of the vignettes.

Rater characteristics

A total of 17 participants were recruited. Approximately half of the sample were those involved in the development of the CBTS-ED. The remaining participants were recruited through professional contacts of those in the CBTS-ED development team (e.g. colleagues, trainee therapists in their service etc.). All participants were therapists working in the field of eating disorders (15 female, one male, and one did not provide gender data). The rating therapists' mean age was 37.53 years (*SD* = 10.44; range = 26–58 years). The number of years qualified ranged from 0 to 36 years (mean = 12.89, SD = 11.06).

Raters were broadly categorised into two groups, nine experts (52.9%) and eight non-experts (47.1%). Expert raters were those who were involved in the development of the CBTS-ED measure and are highly experienced in the field of eating disorders (10 to 40 years of experience). Non-experts were recruited through professional contacts of those in the CBTS-ED development team. Non-experts worked with eating disorders but had limited experience. For example, assistant psychologists or trainee therapists with no more than four years of experience in the field. Participants in both groups were recruited from a range of countries, including the UK, USA, Canada, and Australia.

Procedure

Each rater was sent the six vignettes to observe and rate independently. Raters were also sent a copy of the CBTS-ED measure and a spreadsheet to record scores on. Raters were not trained in how to use the measure, as such training is not routinely available for such competence ratings. Given that lack of such routine training, we decided that the rating materials would be made available for participants, rather than undertaking full training of the raters. It will be necessary to conduct a separate study in the future to evaluate the effect of training in the CBTS-ED.

Data analysis

All analyses were conducted using IBM SPSS Statistics 26 or 28. Due to missing data in one case, a total of 96 completed CBTS-ED ratings were available (16 participants completing six ratings each).

Part A of the CBTS-ED. Internal consistency of the Part A items was examined using Cronbach's alpha and McDonald's omega. Inter-rater reliability for total scores and individual CBTS-ED items were determined using Intraclass Correlation Coefficients (ICC). Koo and Li's (2016) guidance on selection and reporting of appropriate ICC methods was used to determine the ICC model used. ICC estimates and their 95% confidence intervals were calculated based on

a mean-rating of *k* raters, absolute agreement, two-way random-effects model, as the raters were a random selection from a larger population with similar characteristics (i.e. clinicians with varying years of experience in working with eating disorders), and rated all vignettes. ICCs were calculated across all raters combined, expert raters only, and non-expert raters only. ICCs of .75 and above are considered good levels of rater agreement (Koo & Li, 2016). It is important to note that different forms of ICC can give different results. For example, absolute agreement can give smaller ICCs than the "consistency" definition (Koo & Li, 2016). Therefore, our use of the absolute agreement method means that a lower ICC value can be sufficient to describe agreement as "good" than if we had used the consistency method.

Part B of the CBTS-ED. Inter-rater reliability for total mean Part B scores was determined using ICC (as above), and was calculated across all raters combined, expert raters only, and non-expert raters only. As not all Part B items are rated every time, the inter-rater agreement on each item could not be assessed in the same way as Part A. Instead, items were included if they were identified as being relevant to the vignette being observed by the majority of raters across the vignettes. The means, standard deviations, and range of scores for each of the Part B items across the vignettes are presented in the supplementary materials.

Results

The results will be described in two parts, addressing the reliability of each of the two sections of the measure. It should be noted that the mean scores on the two sections were strongly positively correlated (Kendall's tau-b = .867, $n = 6$, $p = .015$). Therefore, levels of competence on the general and the context dependent elements were strongly associated.

Universal CBT-ED competence: part A of the CBTS-ED

Table 2 shows Part A item mean scores and standard deviations on the competence ratings for all raters across the vignettes.

Table 2. Part a item mean scores and standard deviations for all raters across all vignettes ($N = 16$).

CBTS-ED Part A Item	Mean	(SD)
A1. Starts session well	4.60	1.45
A2. Selection and use of core assessment, progress and outcome measures	2.75	1.86
A3. Enhances patient's self-efficacy	4.71	1.40
A4. Engages patient in collaborative work	4.88	1.22
A5. Appropriate pacing and efficient use of time	4.67	1.38
A6. Clarity and honesty in communication style	5.07	1.20
A7. Combines firmness and empathy	4.65	1.47
A8. Weighs the patient collaboratively and openly	2.60	2.03
A9. Stresses 168-hour-per-week therapy	3.88	1.95
A10. Summarises sessions collaboratively	4.08	1.69
Total 10-item Part A mean score (SD)	**4.19**	**1.57**

Internal consistency

Cronbach's alpha and McDonald's omega were high (α and ω = .876), indicating that the scale has strong internal consistency. Removing individual items did not improve the alpha or omega coefficients. The alpha and omega levels were equally strong for expert raters (α = 0.870, ω = .869) and non-experts (α = .884, ω = .887).

Inter-rater reliability for total mean scores

Table 3 shows the ICCs for mean CBTS-ED Part A competence ratings across all raters, expert-raters only, and non-expert raters only. Agreement between all raters was good, with an ICC of 0.87. Agreement between expert raters was excellent, with an ICC of 0.90. Non-expert agreement was moderate, with an ICC of 0.57, suggesting that there are items that non-experts are less likely to agree on. However, it was noted that one non-expert rater consistently rated most of the vignettes lower than the other non-experts. A similar pattern has been found for the CTS-R (Blackburn et al., 2001), where some individuals were identified as outliers, whose scores might inappropriately bias conclusions. Removal of that one non-expert outlier raised that group's inter-rater reliability to a much stronger ICC = 0.71, $p = 0.002$.

Inter-rater reliability for individual items

Table 4 shows ICCs for each individual Part A item of the CBTS-ED across all raters, expert raters only, and non-expert raters only. The ICCs across all raters ranged from 0.64 (A6. *Clarity and honesty in communication style*) to 0.91 (A1. *Starts session well*), indicating moderate to excellent inter-rater reliability. When assessing agreement among expert raters only, ICCs also demonstrated moderate to excellent inter-rater reliability, from 0.65 (Item A6) to 0.91 (Item A1), except for item A8 (*Weighs the patient collaboratively and openly*) which had a low ICC of 0.07, indicating poor reliability (possibly due to it being unclear whether weighing was assumed to have been done, as it was not role-played in the sessions, which were online due to COVID restrictions). Agreement among non-expert raters varied widely from −0.06 (Item A6) to 0.78 (A9. *Stresses 168-hour-per-week therapy*). Non-expert agreement was poor on items A3 (*Enhances patient's self-efficacy*) to A6 but was moderate to good on the remaining Part A items.

Summary

Overall, Part A of the CBTS-ED demonstrates good reliability, with high Cronbach's alpha and McDonald's omega coefficients and moderate to good levels of inter-rater agreement on ratings of competence on most items. Item A6 (*Clarity and honesty in communication style*) demonstrated consistently poorer ICC values across all raters,

Table 3. Intraclass correlation coefficients for total mean CBTS-ED Part A scores.

Rater group	ICC	[95% CI]	Significance
All raters ($n = 16$)	0.87	[0.68–0.98]	<.001
Experts-only ($n = 8$)	0.90	[0.71–0.98]	<.001
Non-experts only ($n = 8$)	0.57	[0.09–0.92]	.009
Non-experts only ($n = 7$)[a]	0.71	[0.28–0.95]	.002

[a]Low scoring non-expert rater removed.

Table 4. Intraclass correlation coefficients for individual CBTS-ED Part A items.

CBTS-ED Part A items	All raters ($n = 16$) ICC [95% CI]	Experts only ($n = 8$) ICC [95% CI]	Non-experts only ($n = 8$) ICC [95% CI]	Non-experts only ($n = 7$)[a] ICC [95% CI]
A1. Starts session well	0.91 [0.77–0.99]***	0.91 [0.74–0.99]***	0.74 [0.32–0.95]**	0.77 [0.37–0.96]**
A2. Appropriate selection of core assessment, progress & outcome measures	0.81 [0.53–0.97]***	0.80 [0.48–0.97]***	0.55 [0.09–0.91]**	0.56 [0.09–0.91]**
A3. Enhances patient's self-efficacy	0.77 [0.43–0.96]***	0.79 [0.41–0.97]**	0.24 [−0.44–0.84]	0.39 [−0.63–0.90]
A4. Engages patient in collaborative work	0.80 [0.51–0.97]***	0.79 [0.45–0.97]***	0.37 [−0.26–0.87]	0.61 [−0.04–0.93]
A5. Appropriate pacing and efficient use of time	0.77 [0.43–0.96]***	0.83 [0.54–0.97]***	0.01 [−0.84–0.79]	0.46 [−0.33–0.91]
A6. Clarity and honesty in communication style	0.64 [0.15–0.94]**	0.75 [0.37–0.96]***	−0.06 [−1.37–0.80]	0.35 [−0.93–0.90]
A7. Combines firmness and empathy	0.83 [0.56–0.97]***	0.65 [0.04–0.94]*	0.74 [0.26–0.96]**	0.87 [0.63–0.98]***
A8. Weighs the patient collaboratively and openly	0.65 [0.28–0.93]***	0.07 [−0.23–0.68]	0.62 [0.19–0.92]***	0.67 [0.24–0.94]***
A9. Stresses 168-hour-per-week therapy	0.89 [0.72–0.98]***	0.88 [0.66–0.98]***	0.78 [0.43–0.96]***	0.79 [0.43–0.96]***
A10. Summarises sessions collaboratively	0.89 [0.70–0.98]***	0.86 [0.60–0.98]***	0.75 [0.32–0.96]**	0.74 [0.29–0.96]**

[a]Low scoring non-expert rater removed; *$p < .025$; **$p < .01$, ***$p < .001$.

expert raters and non-expert raters, suggesting this item requires refinement or removal. The higher level of inter-rater agreement among experts compared to non-experts suggests that experience in the field enhances agreement. However, item A8 (*Weighs the patient collaboratively and openly*) is the exception, with expert agreement being much poorer than non-expert agreement.

Situational CBT-ED competence: part B of the CBTS-ED

To assess the reliability of Part B of the competence measure, the inter-rater reliability for total mean scores was assessed using ICC. As not all Part B items are rated every time, internal consistency and ICC for individual items could not be computed. Instead, the concordance among raters on which items should or should not have been rated was assessed, based on the case or session type.

Inter-rater reliability for total mean scores

Table 5 shows the ICCs for total mean CBTS-ED Part B scores across all raters, expert raters only, and non-expert raters only. Agreement between all raters was good, with an

Table 5. Intraclass correlation coefficients for total mean CBTS-ED Part B scores.

Rater group	ICC	[95% CI]	Significance
All raters ($n = 16$)	0.87	[0.67–0.98]	<.001
Experts-only ($n = 8$)	0.82	[0.52–0.97]	<.001
Non-experts only ($n = 8$)	0.65	[0.10–0.94]	.015
Non-experts only ($n = 7$)[a]	0.78	[0.36–0.96]	.003

[a]Low scoring non-expert rater removed.

ICC of 0.87. Agreement between expert raters was also good, with an ICC of 0.82. Non-expert agreement was moderate, with an ICC of 0.65, suggesting that there are items that non-experts tend not to agree on.

Individual part B items – concordance on which items should be rated

The number of raters who rated each Part B item across the vignettes is shown in Table 6. As Part B items are only rated when relevant to the session being observed, it is important that raters can agree on which items are relevant to the vignette they are observing. Our operational definition of rater agreement is whether most raters (≥80%) agreed on whether an item was relevant to competences that were related to the vignette being observed on at least four out of the six vignettes. Items in bold (Table 6) indicate those where most raters either agreed the item was relevant to the session (≥80% raters rating item), or agreed it was not relevant (≤20% raters rating item). The five items with poor agreement were B1 (*Addresses essential or non-negotiable aspects of treatment*), B5 (*Retains or re-establishes focus on treating issues that maintain the eating disorder*), B9 (*Regular and sufficient eating*), B13 (*Works with different levels of cognition/belief*) and B14 (*Works with emotional states*). This suggests that raters generally agreed on whether the remaining Part B items should or should not have been rated, depending on the vignette and the relevant competences.

Due to how Part B items are rated, ICCs could not be calculated for each item to assess the inter-rater reliability of the scores. Therefore, the means, standard deviations, and range of each item score for all the vignettes can be found in the supplementary materials.

Table 6. Number of raters ($n = 17$) rating each CBTS-ED Part B item for each vignette. Items in bold indicate ≥80% of raters agreeing to either rate or to not rate them on at least 4/6 vignettes. Circled items are those where regardless of the case or session type being observed, they had high numbers of raters (≥80%) agreeing that they should be rated on at least 4/6 vignettes.

	Vignette					
CBTS-ED Part B Item	1	2	3[a]	4[a]	5	6
B1. Addresses essential or non-negotiable aspects of treatment	11	15	6	5	7	16
B2. Focuses on early behavioural change	**17**	**14**	**0**	**7**	**2**	**11**
B3. Elicits and validates patient's emotions	**13**	**15**	**11**	**16**	**16**	**16**
B4. Uses individualised formulation	**17**	**17**	**2**	**14**	**14**	**17**
B5. Retains or re-establishes focus on treating issues that maintain the eating disorder	6	9	5	10	5	11
B6. Provides appropriate psychoeducation	**16**	**16**	**12**	**14**	**14**	**17**
B7. Guidance on necessary dietary change	**14**	**10**	**2**	**3**	**0**	**14**
B8. Monitor eating and other behaviour	**16**	**17**	**8**	**16**	**7**	**15**
B9. Regular and sufficient eating	15	9	4	9	1	17
B10. Prioritise weight regain	**7**	**2**	**1**	**1**	**0**	**17**
B11. Encourages change/do things differently	**17**	**17**	**11**	**12**	**17**	**17**
B12. Behaviour experiments or surveys	**1**	**0**	**16**	**2**	**17**	**9**
B13. Works with different levels of cognition/belief	11	9	6	8	14	11
B14. Works with emotional states	1	8	8	14	8	5
B15. Addresses negative body image	**0**	**3**	**13**	**2**	**15**	**1**
B16. Handling of endings and relapse prevention	**0**	**1**	**15**	**1**	**2**	**1**

[a]Due to missing data, n = 16.

Vignette 1 –addressing early dietary change in a patient with BED; Vignette 2 –addressing restrictive eating in a patient with atypical AN; Vignette 3 –preparing for the end of therapy for a patient with BN; Vignette 4 –addressing emotional triggers in a patient with BN; Vignette 5 –addressing body image distress in an overweight patient with BED; Vignette 6 –addressing compensatory behaviours in a low weight patient.

Table 7. Intraclass correlation coefficients for total mean scores on the refined 15-item Part a of the CBTS-ED, using ratings from the current study.

Rater group	ICC	[95% CI]	Significance
All raters ($n = 16$)	0.87	[0.67–0.98]	<.001
Experts-only ($n = 8$)	0.88	[0.66–0.98]	<.001
Non-experts only ($n = 8$)	0.56	[−0.003–0.91]	.025
Non-experts only ($n = 7$)a	0.71	[0.24–0.95]	.005

[a]Low scoring non-expert rater removed.
Cronbach's alpha, McDonald's omega and individual item ICCs could not be calculated on the five additional Part A items, due to missing data.

Individual part B items – items to be brought into part A

Table 6 indicates that five items are rated relatively consistently over most of the vignettes (B3 – *Elicits and validates patient's emotions*; B4 – *Uses individualised formulation*; B6 – *Provides appropriate psychoeducation*; B8 – *Monitor eating and other behaviour*; B11 – *Encourages change/do things differently*). That is, regardless of the case or session type, these items had high numbers of raters (≥80%) who agreed that the competences should be rated. Vignette 3 was the exception, with fewer raters agreeing on rating these items. However, vignette 3 demonstrated a therapist preparing for the end of therapy. Thus, it is unlikely that a therapist would be expected to demonstrate all the competences outlined in these five Part B items. Therefore, we conclude that these five Part B items (3, 4, 6, 8, 11) should be brought into the 'general' part of the measure (Part A) in future iterations of the measure, rather than being treated as "situation-specific" items. This final version of the measure can be found in. The ICCs for mean item scores on the 15-item Part A are given in Table 7, which shows very similar levels of agreement in each group to that shown in the original 10-item version, with very strong overall inter-rater reliability (ICC = 0.87).

Discussion

The measurement of therapist competence is a complex matter, where concepts need to be more clearly defined and measured in a way that is suited to the specific clinical and training need (Serfaty et al., 2020). Observation-based measures of therapist competence and quality/fidelity have the potential to act as a benchmark for less intensive measures in the future. The current study aimed to develop a new measure to rate therapist competence in CBT-ED delivery across different models—the Cognitive Behavioural Therapy Scale for Eating Disorders (CBTS-ED)—and to pilot its inter-rater reliability. The study also aimed to identify CBTS-ED items that should be retained in the final version of the measure and identify those requiring refinement or removal.

The original Part A of the CBTS-ED, assessing "general" elements of competence that are relevant across all sessions, demonstrated good psychometric properties, with strong internal consistency and moderate to good inter-rater reliability (Tables 3 and 4). Non-expert raters were generally less consistent in their ratings compared to experts, suggesting that experience in the field of eating disorders improves agreement on competence ratings. With regards to Part B, assessing "specific" aspects of competence that are only relevant to some parts of treatment, six out of the 16 Part B items were not consistently rated and should be dropped from the measure. In contrast, five of the Part B items were

rated by the majority of raters across most of the vignettes, and therefore would be better placed in the "general" part of the measure, to be rated every session regardless of the session or case type. These updated 15 items are presented in, for future development and use. This 15-item Part A has strong inter-rater reliability.

The findings show that the disorder-specific CBTS-ED compares well with existing and widely utilised generic measures such as the CTS-R (Blackburn et al., 2001) and the CTACS (Barber et al., 2003) suggesting that the CBTS-ED is a useful specific addition to our ability to work with eating disorders. In particular, its inter-rater reliability indices appear broadly similar to those of the CTS-R (Blackburn et al., 2001), used with anxiety and depression. Individual ICC values are also comparable to those of the CTS-R and the CTACS, with most of the CBTS-ED ICC scores at the upper end of the ranges described in those measures (−0.14 to 0.84 – Blackburn et al., 2001; 0.34 to 0.92 – Barber et al., 2003). The one exception is CBTS-ED item A6, which could not be reliably rated by non-experts.

Whilst this study has provided information on the psychometric properties of the CBTS-ED, it has some limitations. First, the expert raters included those involved with the development of the CBTS-ED, and so they were considerably more familiar with the items than the non-experts, potentially explaining their larger ICC values (Table 4). However, this difference also suggests that training in the use of the CBTS-ED could improve consensus among less experienced raters. Improved reliability with experience has been demonstrated in the CTS-R, with overall agreement among raters on therapist competence significantly improving following training (Reichelt et al., 2003). This need for training does therefore add to the resources required to use such an approach reliably. Nevertheless, despite the lack of training on the CBTS-ED, non-experts still demonstrated acceptable levels of agreement regarding CBT-ED competences, comparable to those using the 13-item CTS-R (Blackburn et al., 2001).

The iterative process of generation of the items in this study was carried out by a set of CBT-ED experts, with experience in delivering a range of evidence-based cognitive-behavioural therapies with eating disorder patients. However, such an approach can result in excessive conformity, and future work of this sort might benefit from a wider set of participants, using a more structured method of generating key items, such as a Delphi study. Furthermore, the development of this and other competence scales would benefit from checking the comprehensibility of expert-generated items among individuals who are non-experts, to ensure the clinical validity of the items.

The current pilot study based its findings on a smaller number of videotaped sessions than other studies (102 videotapes in the CTS-R psychometrics study – Blackburn et al., 2001, p. 134 audiotapes in the CTACS study – Barber et al., 2003). Having only a few sessions might have led to a lack of variation in rater scores, which can influence ICC in determining inter-rater agreement. Higher ICC values can result from a more heterogenous sample compared to a homogenous one, despite similar levels of agreement (Costa-Santos et al., 2011; Müller & Büttner, 1994). This appeared to be the case on several Part A items when assessing agreement among non-experts, where the ICCs were small despite the distribution of rater scores of therapist competence appearing similar across each vignette. Furthermore, the recordings were from online "sessions", and therefore some interventions (e.g. weighing, use of measures) were harder to observe. Overall, it

can be concluded that this preliminary development work on the CBTS-ED needs to be extended to show that the findings are replicable over a substantially larger number of tapes and live sessions, to evaluate the measure comprehensively. The context for such work (e.g. out-patient vs more intensive care) should also be considered, in keeping with Rodriguez-Quintana et al. (2021) suggestion that context might require consideration in assessing competence.

It is noteworthy that several of the CBTS-ED items measure relatively generic competences, which are like those in other measures. This is to be expected, given the similar cognitive behavioural theoretical basis that underpins a range of disorders and treatments. However, it is possible that other CBT-ED skills might be considered for inclusion in future versions. Finally, it would have been beneficial to interview the rater who was considered an outlier, to understand the limitations on the generalisability of the scoring of the CBTS-ED. However, as all raters were anonymous, it was not possible to identify the rater in order to interview them. Future qualitative research into the scoring of the CBTS-ED is needed.

Further verification of the psychometric properties of the revised version of the CBTS-ED (15 general items: 11 specific items) using a larger sample of recorded sessions is required. This should include live sessions across the range of eating disorder presentations, and at various time points in therapy to demonstrate the utility of the measure.

The present findings suggest that a further refined and validated version of the CBTS-ED (e.g. validating skills against training input and against patient outcomes) could be a valuable tool to evaluate clinician competence, and hence the quality of CBT-ED delivered. Such validation research should investigate the construct validity of the CBTS-ED using factor analysis. When factor loadings are available in that way, it is possible that they will demonstrate that some of the competences identified here are more influential than others (e.g. a generic competence such as "Engages the patient in collaborative work" might be more powerful in influencing overall outcomes than "Weighs the patient collaboratively and openly". At that stage, it might be more clinically valid to weight items according to their factor loading, rather than assuming equal importance, as has been done here.

Future research should also consider the utility of the CBTS-ED in demonstrating whether improved competence scores are associated with improved training, supervision, and patient outcomes. The potential of any such measure to enhance clinical outcomes is a critical issue (Power et al., 2022). The CBTS-ED has the potential to identify therapists who are delivering CBT-ED competently and skilfully, which is important in terms of training and supervision. Conversely, the measure could identify those with lower competence, and who would benefit from further support or training, giving patients the best chances of recovery.

Conclusion

This pilot study has described the development and initial reliability of a therapist-rated direct measure of competence in the delivery of CBT-ED—the CBTS-ED. The study presents promising results regarding the inter-rater reliability of the measure and the key items to retain, but also indicates future refinements and directions for clinical research in eating disorders.

Acknowledgments

We thank Rachel Clarke, Victoria Clarke, Melanie Coleman, Tabitha Jackson, P. Evelyna Kambanis, Sharon Keighley, Jessica Kinghorn, Aggie Payne, Marcela Radunz, Setareh Rossman, and Georgia Turnbull for taking part in the development of materials for this study and for the rating of tapes.

Disclosure statement

No potential conflict of interest was reported by the author(s).

ORCID

Jessica Beard https://orcid.org/0000-0002-4330-4536
Zafra Cooper https://orcid.org/0000-0001-7963-656X
Philip Masson https://orcid.org/0009-0003-9878-0438
Victoria A. Mountford https://orcid.org/0000-0002-1932-1723
Rebecca Murphy https://orcid.org/0000-0001-9922-2700
Bronwyn Raykos https://orcid.org/0000-0003-3640-4229
Madeleine Tatham https://orcid.org/0000-0002-6468-6414
Jennifer J. Thomas https://orcid.org/0000-0003-2601-581X
Hannah M. Turner https://orcid.org/0000-0003-4338-5476
Tracey D. Wade https://orcid.org/0000-0003-4402-770X
Glenn Waller https://orcid.org/0000-0001-7794-9546

Data availability statement

CBTS-ED ratings are available for download from the data repository ORDA (DOI: 10.15131/shef.data.21407121). Vignettes are available on request from the corresponding author.

References

Addis, M. E., & Krasnow, A. D. (2000). A national survey of practicing psychologists' attitudes towards psychotherapy treatment manuals. *Journal of Consulting & Clinical Psychology*, *68*, 331–339.

Alfonsson, S., Karvelas, G., Linde, J., & Beckman, M. (2022). A new short version of the cognitive therapy scale revised (CTSR-4): Preliminary psychometric evaluation. *BMC Psychology*, *10*(1), 21. https://doi.org/10.1186/s40359-022-00730-x

Atwood, M. E., & Friedman, A. (2020). A systematic review of enhanced cognitive behavioral therapy (CBT-E) for eating disorders. *International Journal of Eating Disorders*, *53*(3), 311–330. https://doi.org/10.1002/eat.23206

Bailey-Straebler, S., Cooper, Z., Dalle Grave, R., Calugi, S., & Murphy, M. (2022). Development of the CBT-E components checklist: A tool for measuring therapist self-rated adherence to CBT-E. *Italian Journal of Eating Disorders and Obesity*, *4*, 6–10. https://doi.org/10.32044/ijedo.2022.02

Barber, J. P., Gallop, R., Crits-Christoph, P., Barrett, M. S., Klostermann, S., McCarthy, K. S., & Sharpless, B. A. (2008). The role of the alliance and techniques in predicting outcome of supportive-expressive dynamic therapy for cocaine dependence. *Psychoanalytic Psychology*, *25*(3), 461–482. https://doi.org/10.1037/0736-9735.25.3.461

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. https://doi.org/10.1093/ptr/kpg019

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

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

Blackburn, I. M., James, I. A., Milne, D. L., Baker, C., Standart, S., Garland, A., & Reichelt, F. K. (2001). The revised cognitive therapy scale (CTS-R): Psychometric properties. *Behavioural and Cognitive Psychotherapy*, *29*(4), 431. https://doi.org/10.1017/S1352465801004040

Byrne, S. M., Fursland, A., Allen, K. L., & Watson, H. (2011). The effectiveness of enhanced cognitive behavioural therapy for eating disorders: An open trial. *Behaviour Research and Therapy*, *49*(4), 219–226. https://doi.org/10.1016/j.brat.2011.01.006

Cooper, Z., Doll, H., Bailey-Straebler, S., Bohn, K., de Vries, D., Murphy, R., O'Connor, M. E., & Fairburn, C. G. (2017). Assessing therapist competence: Development of a performance-based measure and its comparison with a web-based measure. *JMIR Mental Health*, *4*(4), e7704. https://doi.org/10.2196/mental.7704

Cooper, Z., Doll, H., Bailey-Straebler, S., Kluczniok, D., Murphy, R., O'Connor, M. E., & Fairburn, C. G. (2015). The development of an online measure of therapist competence. *Behaviour Research and Therapy*, *64*, 43–48. https://doi.org/10.1016/j.brat.2014.11.007

Costa-Santos, C., Bernardes, J., Ayres de Campos, D., Costa, A., & Costa, C. (2011). The limits of agreement and the intraclass correlation coefficient may be inconsistent in the interpretation of agreement. *Journal of Clinical Epidemiology*, *64*(3), 264–269. https://doi.org/10.1016/j.jclinepi.2009.11.010

Cowdrey, N. D., & Waller, G. (2015). Are we really delivering evidence-based treatments for eating disorders? How eating-disordered patients describe their experience of cognitive behavioral therapy. *Behaviour Research and Therapy*, *75*, 72–77. https://doi.org/10.1016/j.brat.2015.10.009

Fairburn, C. G. (2008). *Cognitive behavior therapy and eating disorders*. Guilford Press.

Fairburn, C. G., Bailey-Straebler, S., Basden, S., Doll, H. A., Jones, R., Murphy, R., O'Connor, M. E., & Cooper, Z. (2015). A transdiagnostic comparison of enhanced cognitive behaviour therapy (CBT-E) and interpersonal psychotherapy in the treatment of eating disorders. *Behaviour Research and Therapy*, *70*, 64–71. https://doi.org/10.1016/j.brat.2015.04.010

Fairburn, C. G., & Cooper, Z. (2011). Therapist competence, therapy quality, and therapist training. *Behaviour Research and Therapy*, *49*(6–7), 373–378. https://doi.org/10.1016/j.brat.2011.03.005

Fairburn, C. G., Cooper, Z., Doll, H. A., O'Connor, M. E., Bohn, K., Hawker, D. M., Wales, J. A., & Palmer, R. L. (2009). Transdiagnostic cognitive-behavioral therapy for patients with eating disorders: A two-site trial with 60-week follow-up. *The American Journal of Psychiatry*, *166*(3), 311–319. https://doi.org/10.1176/appi.ajp.2008.08040608

Fairburn, C. G., Marcus, M. D., & Wilson, G. T. (1993). Cognitive-behavioral therapy for binge eating and bulimia nervosa: A comprehensive treatment manual. In C. G. Fairburn & G. T. Wilson (Eds.), *Binge eating: Nature, assessment, and treatment* (pp. 361–404). Guilford Press.

Gibbons, C. J., Nich, C., Steinberg, K., Roffman, R. A., Corvino, J., Babor, T. F., & Carroll, K. M. (2010). Treatment process, alliance and outcome in brief versus extended treatments for marijuana dependence. *Addiction*, *105*(10), 1799–1808. https://doi.org/10.1111/j.1360-0443.2010.03047.x

Gutermann, J., Schreiber, F., Matulis, S., Stangier, U., Rosner, R., & Steil, R. (2015). Therapeutic adherence and competence scales for developmentally adapted cognitive processing therapy for adolescents with PTSD. *European Journal of Psychotraumatology*, *6*(1), 26632. https://doi.org/10.3402/ejpt.v6.26632

Haddock, G., Devane, S., Bradshaw, T., McGovern, J., Tarrier, N., Kinderman, P., Baguley, I., Lancashire, S., & Harris, N. (2001). An investigation into the psychometric properties of the cognitive therapy scale for psychosis (CTS-Psy). *Behavioural and Cognitive Psychotherapy*, *29*(2), 221–233. https://doi.org/10.1017/S1352465801002089

Knott, S., Woodward, D., Hoefkens, A., & Limbert, C. (2015). Cognitive behaviour therapy for bulimia nervosa and eating disorders not otherwise specified: Translation from randomized controlled trial to a clinical setting. *Behavioural and Cognitive Psychotherapy*, *43*(6), 641–654. https://doi.org/10.1017/S1352465814000393

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. https://doi.org/10.1016/j.jcm.2016.02.012

Lampard, A. M., & Sharbanee, J. M. (2015). The cognitive-behavioural theory and treatment of bulimia nervosa: An examination of treatment mechanisms and future directions. *Australian Psychologist*, *50*(1), 6–13. https://doi.org/10.1111/ap.12078

Loades, M. E., & Armstrong, P. (2016). The challenge of training supervisors to use direct assessments of clinical competence in CBT consistently: A systematic review and exploratory training study. *The Cognitive Behaviour Therapist*, *9*, e27. https://doi.org/10.1017/S1754470X15000288

Moberg, L. T., Solvang, B., Sæle, R. G., & Myrvang, A. D. (2021). Effects of cognitive-behavioral and psychodynamic-interpersonal treatments for eating disorders: A meta-analytic inquiry into the role of patient characteristics and change in eating disorder-specific and general psychopathology in remission. *Journal of Eating Disorders*, *9*(1), 1–12. https://doi.org/10.1186/s40337-021-00430-8

Mountford, V. A., Allen, K. L., Tchanturia, K., Eilender, C., & Schmidt, U. (2021). Implementing evidence-based individual psychotherapies for adults with eating disorders in a real world clinical setting. *International Journal of Eating Disorders*, *54*(7), 1238–1249. https://doi.org/10.1002/eat.23504

Mulkens, S., de Vos, C., de Graaff, A., & Waller, G. (2018). To deliver or not to deliver cognitive behavioral therapy for eating disorders: Replication and extension of our understanding of why therapists fail to do what they should do. *Behaviour Research and Therapy*, *106*, 57–63. https://doi.org/10.1016/j.brat.2018.05.004

Mulkens, S., & Waller, G. (2021). New developments in cognitive-behavioural therapy for eating disorders (CBT-ED). *Current Opinion in Psychiatry*, *34*(6), 576. https://doi.org/10.1097/YCO.0000000000000745

Müller, R., & Büttner, P. (1994). A critical discussion of intraclass correlation coefficients. *Statistics in Medicine*, *13*(23–24), 2465–2476. https://doi.org/10.1002/sim.4780132310

Muse, K., Kennerley, H., & McManus, F. (2022). The why, what, when, who and how of assessing CBT competence to support lifelong learning. *The Cognitive Behaviour Therapist*, *15*, E57. 10.1017/S1754470X22000502

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

Muse, K., McManus, F., Rakovshik, S., & Thwaites, R. (2017). Development and psychometric evaluation of the assessment of core CBT skills (ACCS): An observation-based tool for assessing cognitive behavioral therapy competence. *Psychological Assessment*, *29*(5), 542–555. https://doi.org/10.1037/pas0000372

National Institute for Health and Care Excellence. (2017). *Eating disorders: Recognition and treatment* (NICE Guideline NG69). https://www.nice.org.uk/guidance/ng69

Power, N., Noble, L., Simmonds-Buckley, M., Kellett, S., Stockton, C., Firth, N., & Delgadillo, J. (2022). Associations between treatment adherence–competence–integrity (ACI) and adult psychotherapy outcomes: A systematic review and meta-analysis. *Journal of Consulting and Clinical Psychology*, *90*(5), 427–445. https://doi.org/10.1037/ccp0000736

Reichelt, F. K., James, I. A., & Blackburn, I. M. (2003). Impact of training on rating competence in cognitive therapy. *Journal of Behavior Therapy and Experimental Psychiatry*, *34*(2), 87–99. https://doi.org/10.1016/S0005-7916(03)00022-3

Rodriguez-Quintana, N., Choi, S. Y., Bilek, E., Koschmann, E., Albrecht, J., Prisbe, M., & Smith, S. (2021). The cognitive behavioral therapy competence scale (CCS): Initial development and validation. *The Cognitive Behaviour Therapist*, *14*, E39. https://doi.org/10.1017/s1754470x21000362

Roth, A. D. (2016). A new scale for the assessment of competences in cognitive and behavioural therapy. *Behavioural and Cognitive Psychotherapy*, *44*(5), 620–624. https://doi.org/10.1017/S1352465816000011

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

Serfaty, M., Shafran, R., Vickerstaff, V., & Aspden, T. (2020). A pragmatic approach to measuring adherence in treatment delivery in psychotherapy. *Cognitive Behaviour Therapy*, *49*(5), 347–360. https://doi.org/10.1080/16506073.2020.1717594

Signorini, R., Sheffield, J., Rhodes, N., Fleming, C., & Ward, W. (2018). The effectiveness of enhanced cognitive behavioural therapy (CBT-E): A naturalistic study within an out-patient eating disorder service. *Behavioural and Cognitive Psychotherapy*, *46*(1), 21–34. https://doi.org/10.1017/S1352465817000352

Södersten, P., Bergh, C., Leon, M., Brodin, U., & Zandian, M. (2017). Cognitive behavior therapy for eating disorders versus normalization of eating behavior. *Physiology & Behavior*, *174*, 178–190. https://doi.org/10.1016/j.physbeh.2017.03.016

Sudak, D. M. (2015). Measures of competence in cognitive behavioral therapy. In D. M. Sudak, R. T. Codd III, J. Ludgate, L. Sokol, M. G. Fox, R. Reiser, & D. L. Milne (Eds.), *Teaching and supervising cognitive behavioral therapy* (pp. 67–84). Wiley.

Turner, H., Marshall, E., Stopa, L., & Waller, G. (2015). Cognitive-behavioural therapy for outpatients with eating disorders: Effectiveness for a transdiagnostic group in a routine clinical setting. *Behaviour Research and Therapy*, *68*, 70–75. https://doi.org/10.1016/j.brat.2015.03.001

von Consbruch, K., 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. https://doi.org/10.1017/S1352465811000622

Wallace, L. M., & von Ranson, K. M. (2011). Treatment manuals: Use in the treatment of bulimia nervosa. *Behaviour Research and Therapy*, *49*(11), 815–820. https://doi.org/10.1016/j.brat.2011.09.002

Waller, G. (2009). Evidence-based treatment and therapist drift. *Behaviour Research and Therapy*, *47*, 119–127.

Waller, G. (2016). Recent advances in psychological therapies for eating disorders. *F1000Research*, *5*, 702. https://doi.org/10.12688/f1000research.7618.1

Waller, G., Cordery, H., Corstorphine, E., Hinrichsen, H., Lawson, R., Mountford, V., & Russell, K. (2007). *Cognitive behavioral therapy for eating disorders: A comprehensive treatment guide*. Cambridge University Press.

Waller, G., Stringer, H., & Meyer, C. (2012). What cognitive-behavioral techniques do therapists report using when delivering cognitive behavioral therapy for the eating disorders? *Journal of Consulting and Clinical Psychology*, *80*(1), 171–175. https://doi.org/10.1037/a0026559

Waller, G., Tatham, M., Turner, H., Mountford, V. A., Bennetts, A., Bramwell, K., Dodd, J., & Ingram, L. (2018). A 10-session cognitive-behavioral therapy (CBT-T) for eating disorders: Outcomes from a case series of nonunderweight adult patients. *International Journal of Eating Disorders*, *51*(3), 262–269. https://doi.org/10.1002/eat.22837

Waller, G., & Turner, H. (2016). Therapist drift redux: Why well-meaning clinicians fail to deliver evidence-based therapy, and how to get back on track. *Behaviour Research and Therapy*, *77*, 129–137.

Waltz, J., Addis, M. E., Koerner, K., & Jacobson, N. S. (1993). Testing the integrity of a psychotherapy protocol: Assessment of adherence and competence. *Journal of Consulting and Clinical Psychology*, *61*(4), 620–630. https://doi.org/10.1037/0022-006X.61.4.620