Behavioural and Cognitive Psychotherapy, 2021, page 1 of 17
doi:10.1017/S1352465821000242

MAIN

The quality of research exploring in-session measures of CBT competence: a systematic review

Kathryn Rayson, Louise Waddington* and Dougal Julian Hare

South Wales Doctoral Programme in Clinical Psychology, 11th Floor, Tower Building, Park Place, Cardiff CF10 3AT, UK
*Corresponding author. Email: cav_psychology.training@wales.nhs.uk

(Received 10 January 2019; revised 5 May 2021; accepted 6 May 2021)

Abstract

Background: Cognitive behavioural therapy (CBT) is in high demand due to its strong evidence base and cost effectiveness. To ensure CBT is delivered as intended in research, training and practice, fidelity assessment is needed. Fidelity is commonly measured by assessors rating treatment sessions, using CBT competence scales (CCSs).

Aims: The current review assessed the quality of the literature examining the measurement properties of CCSs and makes recommendations for future research, training and practice.

Method: Medline, PsychINFO, Scopus and Web of Science databases were systematically searched to identify relevant peer-reviewed, English language studies from 1980 onwards. Relevant studies were those that were primarily examining the measurement properties of CCSs used to assess adult 1:1 CBT treatment sessions. The quality of studies was assessed using a novel tool created for this study, following which a narrative synthesis is presented.

Results: Ten studies met inclusion criteria, most of which were assessed as being ‘fair’ methodological quality, primarily due to small sample sizes. Construct validity and responsiveness definitions were applied inconsistently in the studies, leading to confusion over what was being measured.

Conclusions: Although CCSs are widely used, we need to pay careful attention to the quality of research exploring their measurement properties. Consistent definitions of measurement properties, consensus about adequate sample sizes and improved reporting of individual properties are required to ensure the quality of future research.

Keywords: CBT; clinical skill; cognitive behavioural therapy; competence; training

Introduction

Treatment fidelity or integrity refers to the extent to which a psychological treatment is implemented as intended (Fairburn and Cooper, 2011), and consists of both adherence and competence. Adherence is the extent to which a therapist delivers a therapy in accordance with the therapy model or manual. Competence is the skill with which a therapist delivers the therapy. Adherence and competence have been shown to be highly correlated (Barber *et al*., 2003), with a complex hierarchical relationship. Adherence is necessary but not sufficient for therapist competence, and competence is not sufficient without adherence (Waltz *et al*., 1993). Competence in therapy consists of adherence to the therapy, ability to engage a client and skilful use of treatment change strategies; as well as knowledge of when and when not to apply these strategies (Yeaton and Sechrest, 1981).

A copy of the Checklist for the Appraisal of Therapy Competence Scale Studies (CATCS) is in the supplementary material of the online version of this article. Further information regarding the development of the CATCS is available by contacting the author.

Measuring fidelity or integrity in cognitive behavioural therapy (CBT) is necessary for outcomes research to be meaningful. CBT has become one of the most prominent psychological therapies worldwide (Hofmann *et al*., 2012) due to its strong performance in outcome studies. Clearly, it is important to know that the therapy offered in these studies is actually CBT. Reliable and valid measures of competence in CBT are needed to establish treatment fidelity (Shafran *et al*., 2009) and these must be used with care: for example, a systematic review found that inter-rater reliability of the Cognitive Therapy Scale (CTS) or its revised version (CTS-R) is not often reported and when it is, the results are variable (Loades and Armstrong, 2016).

As the demand for CBT increases, commissioners, services, trainers and researchers all need effective methods to ensure that CBT is delivered with fidelity to the evidence base. CBT is recommended in the UK for many psychological difficulties [National Institute for Clinical Excellence, 2004, 2007, 2011, 2014a, 2014b; *Scottish Psychological Therapy Matrix* (National Health Service Education for Scotland, 2015); *Matrics Cymru: Delivering Evidence-Based Psychological Therapy in Wales* (National Psychological Therapies Management Committee, 2017)]. In England there has been a firm commitment for services to deliver CBT through the roll-out of the Improving Access to Psychological Therapies (IAPT) initiative (Clark, 2011). Commissioners and trainers responsible for disseminating CBT skills need effective methods to assess the impact of training on practitioner competence and to ensure the quality of CBT treatment in everyday practice (Kazantzis, 2003). Studies linking CBT competence to patient outcomes have not given consistent results (Branson *et al*., 2015; Dobson and Kazantzis, 2003; Jacobson and Gortner, 2000). One explanation for this could be the poor reliability of tools used to assess competence (Crits-Christoph *et al*., 1991).

The core competences needed to deliver effective CBT have been incorporated into a broad framework consisting of five domains: (1) generic therapeutic competences; (2) basic CBT competences; (3) specific behavioural and cognitive therapy competences; (4) problem specific competences; and (5) meta-competences (Roth and Pilling, 2007). This framework gives a comprehensive definition of CBT competence, but the authors acknowledge that it is not a measure of competence and advocate the use of competence measures that assess a subset of core competencies (Roth and Pilling, 2008).

A previous systematic review (Muse and McManus, 2013) presented a helpful framework by which different levels of CBT competence can be demonstrated and assessed (see Fig. 1). The framework is based on Miller's (1990) proposal that there are four levels of assessment of competence: (a) the clinician *knows* or has the *knowledge*; (b) the clinician *knows how* to use this knowledge; (c) the clinician *can show how* to do a skill; and (d) the clinician *displays this skill in practice*. Thus, in this hierarchical framework the highest level of competence is evidenced by the therapist using a skill in practice, which can be assessed by rating treatment sessions (assessor or self), supervisory assessments and patient surveys.

The framework suggests that assessor ratings of therapist in-session performance are considered the 'gold standard' in assessing competency (Muse and McManus, 2016). To carry out these ratings, assessors use CBT Competence Scales (CCSs) consisting of a list of domains in which the level of competence observed is rated on an analogue scale. Ratings from each domain can be combined to create an overall competence score and a cut-off point agreed at which a therapist has met a satisfactory level of competence. Crucially, CCSs can be used by independent assessors to avoid bias (Rozek *et al*., 2018).

One of the first CCSs developed was the Cognitive Therapy for Depression Checklist (CCCT: Beck *et al*., 1979); later developed into the Cognitive Therapy Scale (CTS: Dobson *et al*., 1985; Vallis *et al*., 1986). The CTS has been further revised (CTS-R: Blackburn *et al*., 2001), disorder-specific versions developed around the CTS/CTS-R framework (e.g. Competence Rating Scale for PTSD: Dittman *et al*., 2017; CTS for Psychosis: Haddock *et al*., 2001; Cognitive Therapy Competence Scale for Social Phobia: von Consbruch *et al*., 2012) and other global CCSs developed (e.g. Assessment of Core CBT Skills: Muse *et al*., 2017).

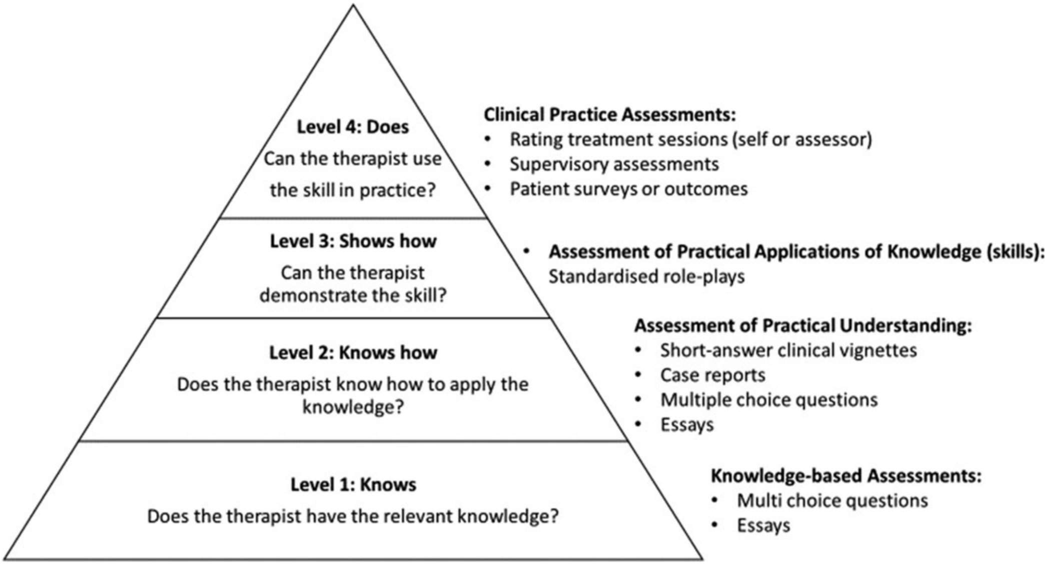
Level 4: Does
Can the therapist use the skill in practice?
Level 3: Shows how
Can the therapist demonstrate the skill?
Level 2: Knows how
Does the therapist know how to apply the knowledge?
Level 1: Knows
Does the therapist have the relevant knowledge?
Clinical Practice Assessments:
• Rating treatment sessions (self or assessor)
• Supervisory assessments
• Patient surveys or outcomes
• Assessment of Practical Applications of Knowledge (skills): Standardised role-plays
Assessment of Practical Understanding:
• Short-answer clinical vignettes
• Case reports
• Multiple choice questions
• Essays
Knowledge-based Assessments:
• Multi choice questions
• Essays

Figure 1. A framework for CBT therapist competence measures, based on Miller's (1990) clinical skills hierarchy (Muse and McManus, 2013).

Due to the widespread use of CCSs in training, development and research, and the consensus that they are the 'gold standard' of competency assessment (Muse and McManus, 2013, 2016), it is essential that the measurement properties of these tools are assessed. A previous review of CBT competence found the reliability and validity of existing CCSs to be mixed (Kazantzis, 2003). A further systematic review examining the assessment of CBT competence found there was still a lack of empirically evaluated CCSs with adequate reliability and validity (Muse and McManus, 2013). Further research is needed to either refine existing measures or develop new scales (Muse and McManus, 2016), but to do so there must be a better understanding of the problems within the existing research.

In the field of psychometric research, it is important to distinguish between the study outcomes and the study design. As there are no published reviews that have assessed the quality of research examining the measurement properties of CCSs, the present review aims to fill this gap. Due to the complex relationship between adherence and competence, the present review will consider measures that assess either competence or a combination of adherence and competence. The specific research questions addressed within the review are:

(1) What is the quality of the research examining CCSs?
(2) How can research into the measurement properties of CCSs be improved?
(3) What are the implications for training and clinical practice in CBT?

Method

Search strategy

Studies were identified through an electronic search of relevant databases: MEDLINE, PsychINFO, Scopus and Web of Science, on 12 February 2018. The following general search strategy was used (see online Supplementary material for individual database search strategies):

(1) ('therap* competen*' OR 'clinical competen*' OR 'therap* skill' OR 'assess* competen*' OR 'competen* assess*' OR 'therap* quality' OR 'intervention competen*' OR 'intervention quality' OR 'clinical expertise') AND ('cognitive therapy' OR 'behav* therapy' OR 'cognitive-behavio*' OR 'cognitive behavio*' OR 'CBT')

OR

(2) ('cognitive therapy scale' OR 'revised cognitive therapy scale' OR 'CTS-R').

A further 10 studies were identified through snowballing methods by cross checking reference lists, key author searches and consultation with a CBT expert.

Inclusion/exclusion criteria

The following inclusion criteria was used to assess eligibility:

(1) Studies published in English from 1980 to the present day.
(2) Studies where the primary aim was the investigation of a CCS based on adult, individual, face-to-face CBT.
(3) Studies published in peer-reviewed journals.
(4) Studies that included competence or mixed adherence and competence scales.

Randomised control trials (RCTs) that use a CCS to assess treatment fidelity were excluded, because their primary focus is not investigating the validity, reliability or responsiveness of a CCS (Terwee *et al.*, 2011). Although the CCCT (Beck *et al.*, 1979) was the first known attempt at a CCS, the psychometrics were not reported until its development into the CTS (Dobson *et al.*, 1985; Vallis *et al.*, 1986). Therefore, the date 1980 was used to ensure any scales reported before 1985 were identified. Article selection was conducted by the lead author (K.R.), in regular discussion with author L.W.

Quality assessment

A thorough literature search was conducted to identify a suitable tool to appraise the quality of the selected papers. As there are to date only a very few tools suitable for assessing the methodological quality of studies on measurement properties, an international Delphi study developed the COnsensus-based Standards for the selection of health Measurement INstruments checklist (COSMIN) (Mokkink *et al.*, 2010a, 2010b). Although there is much overlap between the measurement properties of health status measurement instruments (HSMIs) and CCSs, there are some distinct differences. The COSMIN checklist was considered for use within the present review but was too broad in scope and lacked specificity in relation to studies reporting the measurement properties of CCSs. A new tool, The Checklist for the Appraisal of Therapy Competence Scale Studies (CATCS), was therefore developed for this purpose, based on the criteria in COMSIN, its accompanying definitions of measurement properties, scoring guidelines (when no other published information was available to inform the scoring) and information from a precursor to COSMIN proposing quality criteria (Terwee *et al.*, 2007): the CATCS checklist consists of 17 items relating to: (a) generalisability; (b) reliability: inter-rater reliability, test–re-test reliability, measurement error, internal consistency; (c) validity: structural validity, hypothesis testing, criterion validity, content validity; and (d) responsiveness (a copy of the CATCS can be found in the online Supplementary material). Each item is rated on a

Table 1. Definitions of measurement properties adapted from COSMIN (Mokkink *et al.*, 2010c)

Domain	Measurement property	Aspect of measurement property	Definition
Generalisability			The degree studies have provided sufficient information that one can assign qualitative meaning to an instrument's quantitative scores or change in scores
Reliability			The ability of an instrument to score performance that has not changed, the same way for repeated measures, under several conditions
	Inter-rater reliability		Different raters scoring the same treatment session the same way
	Test–re-test reliability		Scoring the same treatment session, the same way on different occasions
	Measurement error		The difference between the obtained score and its theoretical true score
	Internal consistency		The degree of the interrelatedness among the items
Validity			The degree to which an instrument truly measures the construct(s) it purports to measure
	Construct validity		The degree to which the scores of an instrument are consistent with hypotheses (for instance about internal relationships, relationships to scores of other instruments, or differences between relevant groups) based on the assumption that the instrument validly measures the construct to be measured
		Structural validity	The degree to which the scores of an instrument are an adequate reflection of the dimensionality of the construct to be measured
		Hypothesis testing	The extent to which scores on a questionnaire relate to other measures in a manner that is consistent with theoretically derived hypotheses concerning the concepts that are being measured. Convergent validity tests whether constructs on a scale that should be related are related. Discriminant validity tests whether constructs on a scale that are not supposed to be related, are actually unrelated; e.g. detecting difference between novice and expert therapists
	Criterion validity		The extent to which scores on an instrument are an adequate reflection of a 'gold standard'
	Content validity		The degree to which an instrument includes all the necessary items to represent the concepts to be measured
		Face validity	The degree to which (the items of) an instrument indeed looks as though they are an adequate reflection of the construct to be measured
Responsiveness			The ability of an instrument to detect important change over time in the construct to be measured

scale from 0 to 2 (0 = poor, 1 = fair and 2 = excellent) based on either the design and/or reporting. There is no assumption that these areas are equally weighted and therefore total scores for each paper were not calculated. Definitions of the measurement properties included can be found in Table 1. The results of the quality assessment were synthesised narratively due to the heterogeneity of the studies. Existing critical appraisal tools recognise the importance of generalisability but are unable to capture features that are important for CCSs. Poor reporting in this area undermines reporting of other properties (Terwee *et al.*, 2007): for example, excellent inter-rater reliability cannot be meaningfully generalised if a study does not provide adequate information about the patient population, therapists and raters. A total score for generalisability is reported, with ≥10 deemed to be acceptable.

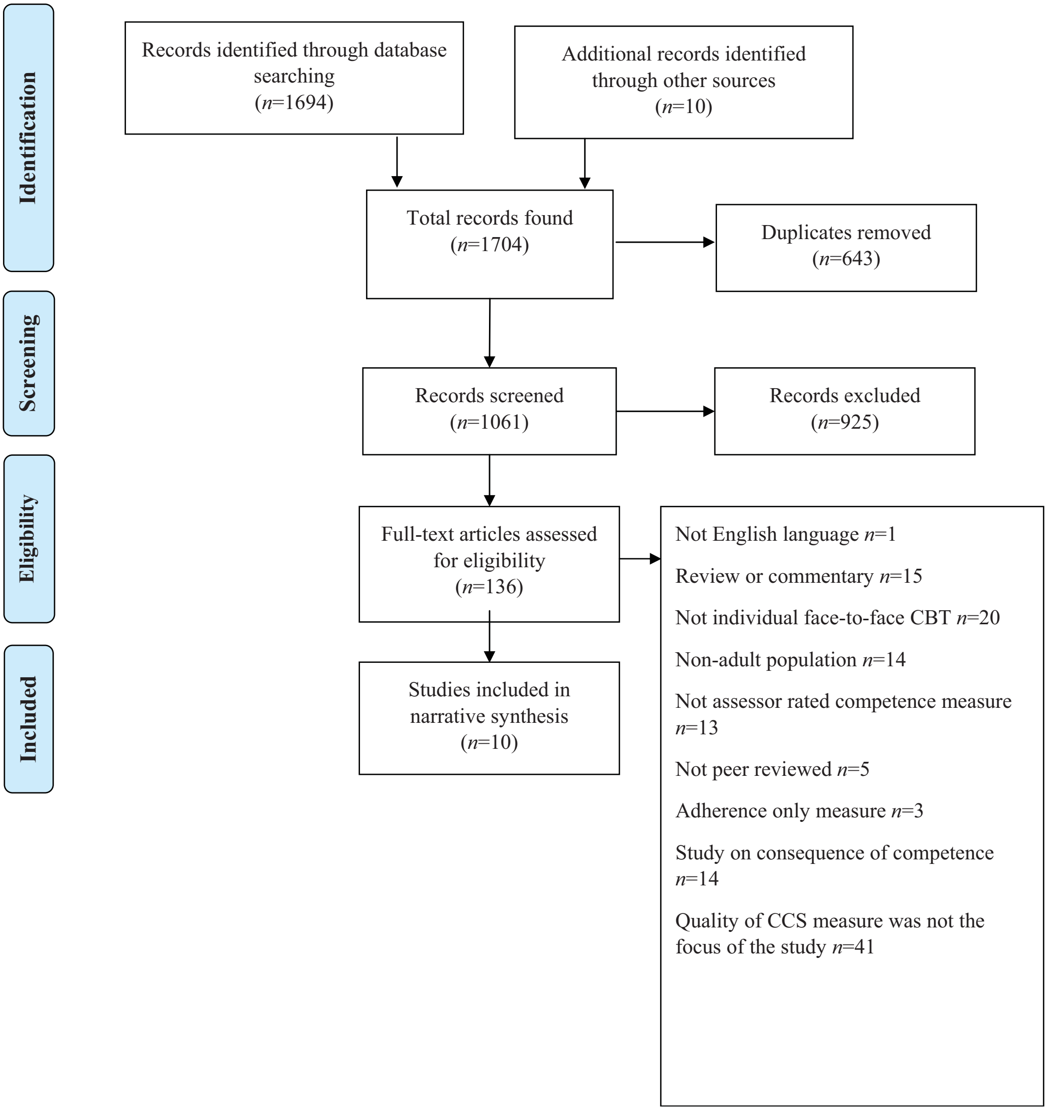
Identification
Screening
Eligibility
Included
Records identified through database searching (n=1694)
Additional records identified through other sources (n=10)
Total records found (n=1704)
Duplicates removed (n=643)
Records screened (n=1061)
Records excluded (n=925)
Full-text articles assessed for eligibility (n=136)
Not English language n=1
Review or commentary n=15
Not individual face-to-face CBT n=20
Non-adult population n=14
Not assessor rated competence measure n=13
Not peer reviewed n=5
Adherence only measure n=3
Study on consequence of competence n=14
Quality of CCS measure was not the focus of the study n=41
Studies included in narrative synthesis (n=10)

Figure 2. Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA; Moher *et al.*, 2009) study flow diagram.

Results

Study selection

After 642 duplicates were removed, the remaining articles were assessed for inclusion by title or abstract and 925 excluded as clearly irrelevant. Full checks of the remaining 136 articles were then conducted, which led to 10 final papers that met inclusion criteria (see Fig. 2 for study flow diagram). Reference lists of the selected papers were also checked to identify any further potential studies.

Study characteristics

Three studies reported the measurement properties for the CTS (Dobson *et al.*, 1985; Dittman *et al.*, 2017; Vallis *et al.*, 1986) and two for the CTS-R (Blackburn *et al.*, 2001; Gordon, 2006).

Table 2. Overview of studies included in the review in chronological order

Study	CCS examined in the study	Purpose of scale/s
Dobson *et al.* (1985)	Cognitive Therapy Scale: CTS	The CTS is a rating scale to assess the quality of cognitive therapy. Originally developed for assessing the quality of cognitive therapy for depression by Young and Beck (1980)
Vallis *et al.* (1986)	Cognitive Therapy Scale-Revised: CTS-R	Developed as a transdiagnostic measure of adherence and competence of cognitive therapy
Carroll *et al.* (2000)	Yale Adherence and Competence Scale: YACS	Developed to rate therapist adherence and competence in delivering behavioural treatments for substance use disorders
Blackburn *et al.* (2001)	Cognitive Therapy Scale-Revised: CTS-R	As above
Haddock *et al.* (2001)	Cognitive Therapy Scale- Psychosis: CTS-PSY	The CTS-PSY was developed to assess the quality of CBT with patients experiencing psychosis. It was adapted from the CTS
Barber *et al.* (2003)	Cognitive Therapy Adherence and Competence Scale: CTACS	Developed to measure adherence and competence of cognitive therapists treating cocaine-dependent patients, but authors report it can also be used on non-drug-dependent patients
Gordon (2006)	Cognitive Therapy Scale- Psychosis: CTS-PSY and Cognitive Therapy Scale-Revised: CTS-R	As above
von Consbruch *et al.* (2012)	Cognitive Therapy Competence Scale for Social Phobia: CTCS-SP	Adapted from the CTS to measure therapist competence in delivering cognitive therapy for social phobia
Dittman *et al.* (2017)	Competence Rating Scale for Cognitive Processing Therapy: CRS-CPT, Competence Rating Scale for PTSD: CRS-PTSD and Cognitive Therapy Scale: CTS	The CRS-CPT was developed as a treatment and disorder specific competence rating scale for treating PTSD with Cognitive Processing Therapy. The CRS-PTSD was developed as a disorder specific competence rating scale for the treatment of PTSD
Muse *et al.* (2017)	Assessment of Core CBT Skills: ACCS	The ACCS aims to assess therapist competence in core general therapeutic and CBT-specific skills, that reflect the current evidence base for the presenting problem

Adapted disorder specific versions of the CTS/R were reported in four papers (Dittman *et al.*, 2017; Gordon, 2006; Haddock *et al.*, 2001; von Consbruch *et al.*, 2012). Two studies reported a competence subscale within a scale that also examined adherence (Barber *et al.*, 2003; Carroll *et al.*, 2000). One scale was a report of a newly developed global measure of CBT competence (Muse *et al.*, 2017). See Table 2 for an overview of the studies.

Findings of the quality assessment

The quality assessment of the studies in the current review was conducted by the lead author (K.R.) and a sample of 50% of the studies was assessed by a colleague independent from the review. Inter-rater reliability was assessed using a linear weighted kappa (Cohen, 1968) and was found to be good: $\kappa = 0.76$ (95% confidence interval, 0.70 to 0.88), $p<0.0005$ (Altman, 1991). A small number of differences were identified and discussed between the two raters and resolved for the final ratings. Results of the quality assessment using the CATCS are reported in Table 3 for generalisability and Table 4 for quality of measurement property methodology and reporting.

Generalisability

All but one study (Dittman *et al.*, 2017) had a total score ≥10 for generalisability. Therefore, most of the studies provided sufficient information that the results can be meaningfully interpreted

Table 3. Critical appraisal results for generalisability using the CATCS

Study and CBT competence scale examined	Study purpose	Protocol for scale	Therapy/ patients/ setting	Recordings	No. of raters	Raters	No. of therapists	Therapists	Total (max. 16)
Barber *et al.* (2003) CTACS	2	2	2	1	0	2	2	2	13
Blackburn *et al.* (2001) CTS-R	2	2	2	2	1	0	2	2	13
Carroll *et al.* (2000) YACS	2	2	2	1	1	2	2	0	12
Dittman *et al.* (2017) CRS-CPT, CRS-PTSD and CTS	2	1	2	1	0	0	1	2	9
Dobson *et al.* (1985) CTS	2	2	2	0	1	2	2	2	13
Gordon (2006) CTS-PSY and CTS-R	2	2	2	0	2	2	2	2	14
Haddock *et al.* (2001) CTS-PSY	2	2	2	0	1	1	2	2	12
Muse *et al.* (2017) ACCS	2	2	2	2	2	2	2	2	16
Vallis *et al.* (1986) CTS-R	2	2	2	1	2	2	2	2	15
von Consbruch *et al.* (2012) CTCS-SP	2	2	2	1	2	2	2	1	14

Scoring criteria for quality: 0, poor; 1, fair; 2, excellent. ACCS, Assessment of Core CBT Skills; CTS, Cognitive Therapy Scale: CTS; CTS-R, Cognitive Therapy Scale-Revised; CTACS, Cognitive Therapy Adherence and Competence Scale; YACS, Yale Adherence and Competence Scale; CRS-CPT, Competence Rating Scale for Cognitive Processing Therapy; CRS-PTSD, Competence Rating Scale for PTSD; CTCS-SP, Cognitive Therapy Competence Scale for Social Phobia; CTS-PSY, Cognitive Therapy Scale- Psychosis.

within the specific contexts of each study. All the studies provided clear information about the purpose of the study, a protocol for the CCS and about the types of patients treated.

Three of the studies (Barber *et al.*, 2003; Blackburn *et al.*, 2001; Dittman *et al.*, 2017) received a poor rating for number of raters used or types of raters used. Risk of bias was increased in three studies that employed raters who were not independent of the study. The seven studies that received an excellent rating for 'Raters' employed at least one rater that was truly independent and provided a clear explanation of the training the raters underwent. Four studies were assessed as excellent both for the number of raters used and the characteristics of the raters (Gordon, 2006; Muse *et al.*, 2017; Vallis *et al.*, 1986; von Consbruch *et al.*, 2012).

All studies reported using an acceptable number of different therapists, with no studies receiving a poor rating in this domain. Similarly, most of studies provided an excellent description of therapists and their training, with only one study receiving a poor rating (YACS: Carroll *et al.*, 2000).

Reliability

Inter-rater reliability

Inter-rater reliability of competence scales can be measured for the total scales and individual items. Often the inter-rater reliability of total scales is found to be good but lower for individual items (e.g. Barber *et al.*, 2003; Blackburn *et al.*, 2001; Dobson *et al.*, 1985; von Consbruch *et al.*, 2011). All but one study (Vallis *et al.*, 1986) reported both total scale and individual item correlations using appropriate statistical analysis.

Table 4. Reliability and validity methodology ratings using the CATCS

CBT competence scale	Inter-rater reliability	Test–re-test reliability	Measurement error	Internal consistency	Structural validity	Hypothesis testing	Criterion validity	Content validity	Responsiveness
Barber *et al.* (2003) CTACS	1 (92)			1 (?)	1 (?)	1 (92)		2	
Blackburn *et al.* (2001) CTS-R	2 (102)			1 (?)				2	0 (22)
Carroll *et al.* (2000) YACS	0 (19)				1 (83)	1 (79–576)			
Dittman *et al.* (2017) CRS-CPT, CRS-PTSD and CTS	0 (21)			0 (21)				2	
Dobson *et al.* (1985) CTS	1 (30)			1 (30)		1 (30)		2	
Gordon, 2006 CTS-PSY and CTS-R	1 (20–26)		1 (20–26)			0 (20–26)		2	
Haddock *et al.* (2001) CTS-PSY	0 (5)					0 (24)		1	
Muse *et al.* (2017) ACCS	1 (55)			2 (111)		1 (68–76)		2	0 (17)
Vallis *et al.* (1986) CTS	0 (10)			1 (90)	1 (90)	1 (53)		1	
von Consbruch *et al.* (2011) CTCS-SP	2 (161)	1 (15)		1 (161)					

Scoring criteria for quality: 0, poor; 1, fair; 2, excellent. Blank boxes indicate the domain was not reported on in the study. Items in brackets denote the sample size the analysis was performed on; '?' indicates that the sample size was not clear. ACCS, Assessment of Core CBT Skills; CTS, Cognitive Therapy Scale: CTS; CTS-R, Cognitive Therapy Scale-Revised; CTACS, Cognitive Therapy Adherence and Competence Scale; YACS, Yale Adherence and Competence Scale; CRS-CPT, Competence Rating Scale for Cognitive Processing Therapy; CRS-PTSD, Competence Rating Scale for PTSD; CTCS-SP, Cognitive Therapy Competence Scale for Social Phobia; CTS-PSY, Cognitive Therapy Scale-Psychosis.

Test–re-test reliability

Only one study (von Consbruch *et al.*, 2012) conducted a test–re-test reliability assessment for a CCS. Some researchers have suggested that the re-test method should not be used to estimate reliability and advocate the use of internal consistency (Nunnally and Bernstein, 1994). Reasons cited include the stability of the attribute being measured and carry-over effects in the second rating (Polit, 2015). Carry-over effects in relation to rating CCSs could include the rater recalling their previous ratings or wanting to appear consistent. von Consbruch and colleagues (2012) attempted to reduce the influence of carry-over effects by ensuring there was 18 to 24 months between each rating. Despite this, the sample size used for the analysis was only 15 tapes, so test–re-test reliability methodology received a 'poor' rating.

Measurement error

Measurement error was reported in only one study (Gordon, 2006). Although the analysis used was appropriate, the sample size was <30 and so received a poor rating.

Internal consistency

Internal consistency was reported in seven studies, with varying quality and sample sizes. Only one study received an excellent rating (ACCS: Muse *et al.*, 2017). Studies that received a fair rating did so because they did not calculate factor analysis per dimension (Barber *et al.*, 2003; von Consbruch *et al.*, 2011) or they had small sample sizes. Caution should be exercised in interpreting the internal consistency results of the CRS-CPT and CRS-PTSD, as this domain received a 'poor' quality rating due to the small sample size in the study (Dittman *et al.*, 2017). There are no studies that have examined the internal consistency of the CTS-PSY (Gordon, 2006: Haddock *et al.*, 2001) and YACS (Carroll *et al.*, 2000).

Validity

Criterion validity

Barber *et al.* (2003) was the only study to report criterion validity but the description provided is more congruent with the definition of discriminant validity and so was considered as such. Criterion validity was included in the quality assessment tool as initially it appeared that some studies reported this construct, but an actual 'gold standard' for CCSs may not currently exist.

Content validity

Content validity is less relevant in studies of instruments that have been adapted from an original scale or if the scale has already demonstrated content validity reported elsewhere. Only two studies did not report content validity. It is not clear why it was not reported for the YACS (Carroll *et al.*, 2000), as this was a novel instrument. von Consbruch *et al.* (2011) may not have reported on this for the CTCS-SP as it was adapted from the CTS, but they should then have reported this for the new items. Overall, the quality of reporting of content validity was excellent in six studies; but two studies scored 'fair' as they did not cover the domain in sufficient detail (Haddock *et al.*, 2001; Vallis *et al.*, 1986).

Construct validity

Construct validity includes structural validity, hypothesis testing and cross-cultural validity. Cross-cultural validity refers to how well an instrument has been adapted for different cultures or languages but was not included in this review as the search strategy found no such studies published in English that met inclusion criteria.

Structural validity. Structural validity was reported in only three papers, with each achieving a 'fair' score. In two papers the method of analysis was good, but the sample size was not high enough to achieve an 'excellent' rating (Carroll *et al.*, 2000, $n = 83$; Vallis *et al.*, 1986, $n = 90$), and the sample size was not clear in another (Barber *et al.*, 2003).

Hypothesis testing. Hypothesis testing was included in seven of the papers. Five studies were rated 'fair' due to sample sizes of 30–99 or not making explicit hypotheses *a priori*. The two studies examining the CTS-PSY both received a 'poor' rating for hypothesis testing, again due to small sample sizes (Gordon, 2006, $n = 20$–26; Haddock *et al.*, 2001, $n = 24$).

Responsiveness

For the purposes of clarity, this review adopted the COSMIN definition of responsiveness as '*the ability of a scale to detect changes longitudinally*'. In this case, it refers to a CCS detecting changes in competence over time, perhaps because of experience or training. If studies assessed the same therapists at different time points and calculated their change in scores, then this was considered a measure of responsiveness of the scale. If, however, this was done using a cross-sectional design, where change was not calculated for each individual therapist, then it was considered discriminant validity, as the aim was to assess if the scale can discriminate between different groups: e.g. expert versus novice.

Two studies (Blackburn *et al.*, 2001; Muse *et al.*, 2013) reported discriminant validity, but the description they used fits the definition of responsiveness. Unfortunately, both studies received a 'poor' rating for the methodology due to small sample sizes.

Discussion

This study is the first attempt to assess the quality of research reporting the psychometric properties of CCSs. This study also sought to make recommendations for improving the quality of research examining CCSs and consider implications for training and clinical practice in CBT.

What is the quality of the research examining CCSs?

This review found that overall, the quality of the studies was very mixed, and no studies demonstrated 'excellent' quality throughout. The quality was significantly affected by small sample sizes. A sample size of $n \leq 30$ is defined as 'poor' based on the COSMIN guidelines for assessing the quality of patient HSMIs (Mokkink *et al.*, 2010a, 2010b, 2012). The COSMIN benchmark was used due to a lack of any other guidelines, around sample sizes for measurement property research in competence scales, but using this benchmark requires caution. For example, the minimum sample size for performing confirmatory factor analysis is usually quoted as $n = 200$–300 (Polit, 2015). This may be unachievable for studies examining the properties of CCSs and further guidance is required as to minimum sample sizes for assessing the structural validity of CCSs. There are high costs involved in rating CBT treatment sessions due to the need for expert raters (Weck *et al.*, 2011). Further consensus is needed to clarify the minimum number of recordings of rated sessions needed for each measurement property, as this may vary per dimension.

From this detailed analysis, some inferences about overall quality of the studies in the current review can be made. The methodologies of the studies examining the measurement properties of the CTS (Dittman *et al.*, 2017; Dobson *et al.*, 1985; Vallis *et al.*, 1986), CTS-R (Blackburn *et al.*, 2001; Gordon, 2006) and CTS-PSY (Gordon, 2006; Haddock *et al.*, 2001) were assessed as being of 'poor' to 'fair' quality. The exceptions to this were content validity reporting for most studies of the

CTS and inter-rater reliability methodology in one CTS-R study (Blackburn *et al.*, 2001), which were all rated as 'excellent'. The findings of this review suggest the quality of these studies examining the CTS and CTS-R are not robust enough, and that conclusions about their reliability and validity need to be held tentatively.

The study examining the CTCS-SP (von Consbruch *et al.*, 2011) had quality ratings between 'fair' and 'excellent'. The ACCS (Muse *et al.*, 2017) had quality ratings from 'poor' to 'excellent'. 'Fair' and 'poor' scores were awarded due to small sample sizes, but methodology was appropriate otherwise. In some ways, these more recent studies have addressed some of the previous methodological problems in previous studies on the CTS and CTS-R but were still not of consistent high quality. For example, the study examining the CRS-CPT and CRS-PTSD (Dittman *et al.*, 2017) was the only study to not receive an acceptable score for generalisability and was awarded 'poor' for methodology in all domains, except the 'content validity' domain, which received an excellent rating.

Overall, the quality of the assessment of inter-rater reliability was affected by the variation in samples sizes and particular caution should be exercised when interpreting the inter-rater reliability results of the YACS (Carroll *et al.*, 2000), CRS-CPT (Dittman *et al.*, 2017) and CPT-PTSD (Dittman *et al.*, 2017) due to small sample sizes of $n = 19$ and $n = 21$, respectively. This is also true for the CTS, which was used with small sample sizes in three studies (Vallis *et al.*, 1986, $n = 10$; Dittman *et al.*, 2017, $n = 30$; Dobson *et al.*, 1985, $n = 30$).

Assessing measurement error appears to be a neglected area of CCS measurement property evaluation, despite its evident importance. For example, health measurements such as physiological markers of disease are reasonably stable characteristics but measuring competence could be influenced by other components that are not the subject of measurement (Rosenkoetter and Tate, 2018). This means that when measuring competence by a single assessor, a degree of error may exist between the true theoretical score and the actual given score. If a CCS has a cut-off score for competence (e.g. CTS-R), then calculating the measurement error can provide a confidence interval of the estimate of the score (Gordon, 2006).

How can research into the measurement properties of CCSs be improved?

Researchers need to identify a 'gold standard' in order to assess the 'criterion validity' of measures. Although the CTS-R (Blackburn *et al.*, 2001) is used extensively in research and training to assess CBT competence, there is no empirical evidence that it is the 'gold standard' or that another exists. In future research examining the measurement properties of CCSs, 'criterion validity' should not be assessed unless the study presents good evidence that the comparative measure is a 'gold standard'. This 'gold standard' is unlikely to be just one measure of competence, as multiple measures from different sources are more reliable (Muse and McManus, 2013). Instead, 'convergent validity', which measures whether constructs on a scale that should be related are related, might be more appropriate.

Test–re-test reliability was a neglected area of measurement with only one study reporting it (von Consbruch *et al.*, 2012). Future research should consider examining test–re-test reliability as scoring the same treatment session, the same way on different occasions, should be an important feature of CCSs. It is, however, understandable that some studies are not able to run for sufficient time to provide conditions to assess test–re-test reliability, which would enable a reduction in carry-over effects. In these cases, calculating internal consistency only requires the rating of competence at one time point (Polit, 2015), which might be preferable given the time and costs involved in rating the same session twice.

Finally, there is difficulty in interpreting responsiveness in CCS studies, due to the property it is measured against. The discrepancy found between the reporting of discriminant validity and responsiveness in Blackburn *et al.* (2001) and Muse *et al.* (2013) is perhaps understandable given some psychometricians have argued that responsiveness does not require its own label, describing it as longitudinal construct validity (Streiner *et al.*, 2015; Terwee *et al.*, 2003). This may also be why no CCSs, in this review, have examined responsiveness explicitly. To know that a scale can detect change longitudinally you would need to have some way of ensuring that the competence level had in fact changed. There is research supporting the view that competence increases because of training (James *et al.*, 2001; McManus *et al.*, 2010), but these studies only measure competence using CCS scores, when multiple methods would strengthen their designs (Alberts and Edelstein, 1990). Although the COSMIN group reached consensus that responsiveness should be its own distinct domain for health measurement instruments, further clarity is needed to understand if and how this should be applied to CCS research. Regardless of whether responsiveness or discriminant validity is assessed, the research would benefit from including multiple measures of competence from which CCSs could be measured against.

The current review sought to define the properties clearly to assign quality scores, but there is difficulty in defining these constructs. As mentioned, some researchers assert that responsiveness is in fact a version of (longitudinal) construct validity (Streiner *et al.*, 2015). Similarly, criterion validity and convergent validity were often confused, as they may be evaluating the same construct. The COSMIN panel identified the similarity between responsiveness, construct validity and criterion validity (Mokkink *et al.*, 2010c). Although the COSMIN team have attempted to define these properties for HSMIs, further consensus is again needed to specify these terms in relations to therapy competence scales.

What are the implications for training and clinical practice in CBT?

Overall, the quality of the studies means conclusions about the validity and reliability of all CCSs should be held tentatively particularly regarding criterion validity, measurement error, test–re-test reliability, responsiveness and criterion validity.

Researchers, trainers and supervisors should exercise caution if using single assessors to rate therapist's competence based on a suggested cut-off score, as such a score may be subject to measurement error. Best practice would therefore be to ensure that a given session is rated by two different assessors, as is standard practice on British Association for Behavioural and Cognitive Psychotherapies accredited training programmes with all CTSs and CTS-Rs being marked and moderated by two assessors. Further examination of measurement error in research is need for CCSs that have a suggested cut-off score, especially if it is known that the CCSs are used by single assessors.

Limitations

As only the lead author conducted the paper selection, this increased the possibility of researcher bias. The lead author (K.R.) discussed the paper selection at every stage with author L.W., but the review would have been strengthened if paper selection was conducted fully by at least two authors.

The review developed a novel tool (CATCS) to assess the quality of studies examining the measurement properties of CCSs. The CATCS may also have utility for assessing the quality of other competence measures in psychotherapy. A recent systematic review found that 44 to 88% of papers that create novel measures do not report supporting information about the tool's reliability and validity (Flake *et al.*, 2017). The CATCS is the first published tool developed to assess the quality of competence measures research.

Although the inter-rater reliability of the CATCS was found to be good, the analysis was only conducted on half of the studies, which limits the reliability of the measure. Further assessment of the CATCS's measurement properties is required to ascertain its reliability and validity. The CATCS could be further improved by more detailed instructions for scoring criteria to ensure results can be replicated between researchers: expert opinion from psychometrics, HSMI fields and psychotherapy may add further clarity to the construct definitions and adequate samples sizes.

The exclusion criteria were intentionally narrow to focus on research that specifically aims to examine the measurement properties of CCSs. Additional studies reporting some of the measurement properties of these tools were not included, e.g. RCTs using a CSS to assess fidelity. Furthermore, the present review did not have the resources to include non-English language studies which led to the review not examining cross-cultural validity, as it is only appropriate for translated instruments (Mokkink *et al.*, 2010c). The CTS/R has been translated into other languages and further research could include a review of the measurement properties of these scales, including cross-cultural validity.

Conclusions

This systematic review presents the first attempt to assess the quality of the research examining measurement properties of CCSs. The review found only ten studies that met inclusion criteria and overall quality was assessed as 'poor' to 'fair', mostly due to sample sizes. Given the widespread use of CCS to assess competence in research, practice and training, it is important to note that the quality of the research reporting the properties of CCS was not better. The review makes recommendations to improve future research into the measurement properties of CCS, including clarity and consensus regarding definitions of measurement properties and adequate samples sizes.

Acknowledgements. None.

Financial support. This research received no specific grant from any funding agency, commercial or not-for-profit sectors and was undertaken as part of Kathryn Rayson's DClinPsy training.

Conflicts of interest. Kathryn Rayson, Dougal Hare and Louise Waddington have no conflicts of interest with respect to this publication.

Ethical statement. No ethical approval was required as this is a review.

Data availability statement. The authors confirm that the data supporting the findings of this study are available with the article (and/or its supplementary material).

Supplementary material. To view supplementary material for this article, please visit: https://doi.org/10.1017/S1352465821000242

References

Alberts, G., & Edelstein, B. (1990). Therapist training: a critical review of skill training studies. *Clinical Psychology Review*, *10*, 497–511. https://doi.org/10.1016/0272-7358(90)90094-Q

Altman, D. G. (1991). *Practical Statistics for Medical Research*. London, UK: Chapman and Hall. http://doi.org/10.1002/sim.4780101015

Barber, J. P., Liese, B. S., & Abrams, M. J. (2003). Development of the cognitive therapy adherence and competence scale. *Psychotherapy Research*, *13*, 205–221. http://doi.org/10.1093/ptr/kpg019

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

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*, 431–446. http://doi.org/10.1017/S1352465801004040

Branson, A., Shafran, R., & Myles, P. (2015). Investigating the relationship between competence and patient outcome with CBT. *Behaviour Research and Therapy*, *68*, 19–26. https://doi.org/10.1016/j.brat.2015.03.002

Carroll, K. M., Nich, C., Sifry, R. L., Nuro, K. F., Frankforter, T. L., Ball, S. A., … & Rounsaville, B. J. (2000). A general system for evaluating therapist adherence and competence in psychotherapy research in the addictions. *Drug and Alcohol Dependence*, *57*, 225–238. https://doi.org/10.1016/S0376-8716(99)00049-6

Clark, D. M. (2011). Implementing NICE guidelines for the psychological treatment of depression and anxiety disorders: the IAPT experience. *International Review of Psychiatry*, *23*, 318–327. https://doi.org/10.3109/09540261.2011.606803

Cohen, J. (1968). Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, *70*, 213–220. http://doi.org/10.1037/h0026256

Crits-Christoph, P., Baranackie, K., Kurcias, J., Beck, A., Carroll, K., Perry, K., … & Gallagher, D. (1991). Meta-analysis of therapist effects in psychotherapy outcome studies. *Psychotherapy Research*, *1*, 81–91. https://doi.org/10.1080/10503309112331335511

Dittmann, C., Müller-Engelmann, M., Stangier, U., Priebe, K., Fydrich, T., Görg, N., … & Steil, R. (2017). Disorder- and treatment-specific therapeutic competence scales for posttraumatic stress disorder intervention: development and psychometric properties. *Journal of Traumatic Stress*, *30*, 614–625. https://doi.org/10.1002/jts.22236

Dobson, K. S., & Kazantzis, N. (2003). The therapist in cognitive-behavioral therapy: Introduction to a special section. *Psychotherapy Research*, *13*, 131–134. https://doi.org/10.1080/713869635

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*, 295–300. http://doi.org/10.1111/j.2044-8260.1985.tb00662

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

Flake, J. K., Pek, J., & Hehman, E. (2017). Construct validation in social and personality research: durrent practice and recommendations. *Social Psychological and Personality Science*, *8*, 370–378. https://doi.org/10.1177/1948550617693063

Gordon, P. K. (2006). A comparison of two versions of the Cognitive Therapy Scale. *Behavioural and Cognitive Psychotherapy*, *35*, 343–353. https://doi.org/10.1017/S1352465806003390

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

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

Jacobson, N. S., & Gortner, E. T. (2000). Can depression be de-medicalized in the 21st century: scientific revolutions, counter-revolutions and the magnetic field of normal science. *Behaviour Research and Therapy*, *38*, 103–117. https://doi.org/10.1016/S0005-7967(99)00029-7

James, I. A., Blackburn, I. M., Milne, D. L., & Reichfelt, F. K. (2001). Moderators of trainee therapists' competence in cognitive therapy. *British Journal of Clinical Psychology*, *40*, 131–141. https://doi.org/10.1348/014466501163580

Kazantzis, N. (2003). Therapist competence in cognitive-behavioural therapies: review of the contemporary empirical evidence. *Behaviour Change*, *20*, 1–12. https://doi.org/10.1375/bech.20.1.1.24845

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

McManus, F., Westbrook, D., Vazquez-Montes, M., Fennell, M., & Kennerley, H. (2010). An evaluation of the effectiveness of diploma-level training in cognitive behaviour therapy. *Behaviour Research and Therapy*, *48*, 1123–1132. https://doi.org/10.1016/j.brat.2010.08.002

Miller, G. E. (1990). The assessment of clinical skills/competence/performance. *Academic Medicine*, *65*, S63–67. Available at: http://winbev.pbworks.com/f/Assessment.pdf (accessed 15 December 2017).

Moher D., Liberati A., Tetzlaff J., Altman D.G. and **The PRISMA Group** (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: the PRISMA statement. *PLOS Medicine*, *6*, e1000097. https://doi.org/10.1371/journal.pmed.1000097

Mokkink, L. B., Terwee, C. B., Patrick, D. L., Alonso, J., Stratford, P. W., Knol, D. L., Bouter, L. M. & de Vet, H. C. W. (2010a). The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. *Quality of Life Research*, *19*, 539–549. https://doi.org/10.1007/s11136-010-9606-8

Mokkink, L. B., Terwee, C. B., Knol, D. L., Stratford, P. W., Alonso, J., Patrick, D. L., Bouter, L. M. & de Vet, H. C. W. (2010b). The COSMIN checklist for evaluating the methodological quality of studies on measurement properties: a clarification of its content. *BMC Medical Research Methodology*, *10*, 22. https://doi.org/10.1186/1471-2288-10-22

Mokkink, L. B., Terwee, C. B., Patrick, D. L., Alonso, J., Stratford, P. W., Knol, D. L., … & de Vet, H. C. (2010c). The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *Journal of Clinical Epidemiology*, *63*, 737–745. https://doi.org/10.1016/j.jclinepi.2010.02.006

Mokkink, L. B., Terwee, C. B., Patrick, D. L., Alonso, J., Stratford, P. W., Knol, D. L., … & de Vet, H. C. (2012). *COSMIN Checklist Manual.* Amsterdam, The Netherlands: COSMIN. Available at: http://www.cosmin.nl/images/upload/files/COSMIN%20checklist%20manual%20v9.pdf (accessed 9 February 2018).

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

Muse, K., & McManus, F. (2016). Expert insight into the assessment of competence in cognitive-behavioural therapy: a qualitative exploration of experts' experiences, opinions and recommendations. *Clinical Psychology & Psychotherapy*, *23*, 246–259. https://doi.org/10.1002/cpp.1952

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*, 542–555. http://doi.org/10.1037/pas0000372

National Health Service Education for Scotland (2015). *The Matrix: A Guide to Delivering Evidence-Based Psychological Therapies in Scotland.* Edinburgh, UK: NHS Education for Scotland.

National Institute for Clinical Excellence (2004). Eating disorders. Clinical Guideline 9. London, UK: NICE.

National Institute for Health and Clinical Excellence (2007). Drug misuse: psychosocial intervention. Clinical Guideline 51. London, UK: NICE.

National Institute for Health and Clinical Excellence (2011). Common mental health problems: identification and pathways to care. Clinical Guideline 123. London, UK: NICE.

National Institute for Health and Clinical Excellence (2014a). Psychosis and schizophrenia in adults: treatment and management. Clinical Guideline 178. London, UK: NICE.

National Institute for Health and Clinical Excellence (2014b). Bipolar disorder: the management and assessment of bipolar disorder in adults, children and young people in primary and secondary care. Clinical Guideline 185. London, UK: NICE.

National Psychological Therapies Management Committee (2017). *Matrics Cymru: Guidance for Delivering Evidence-Based Psychological Therapy in Wales.* NPTMC, Public Health Wales.

Nunnally, J., & Bernstein, I. H. (1994). *Psychometric Theory* (3rd edn). New York, USA: McGraw Hill.

Polit, D. F. (2015). Assessing measurement in health: beyond reliability and validity. *International Journal of Nursing Studies*, *52*, 1746–1753. https://doi.org/10.1016/j.ijnurstu.2015.07.002

Rosenkoetter, U., & Tate, R. L. (2018). Assessing features of psychometric assessment instruments: a comparison of the COSMIN checklist with other critical appraisal tools. *Brain Impairment*, *19*, 103–118. https://doi.org/10.1017/BrImp.2017.29

Roth, A. D., & Pilling, S. (2007). The competences required to deliver effective cognitive and behavioural therapy for people with depression and with anxiety disorders. Available at: https://www.ucl.ac.uk/pals/sites/pals/files/migrated-files/CBT_competences_-_map_Sept_2015.pdf (accessed 23 January 2018).

Roth, A. D., & Pilling, S. (2008). Using an evidence-based methodology to identify the competences required to deliver effective cognitive and behavioural therapy for depression and anxiety disorders. *Behavioural and Cognitive Psychotherapy*, *36*, 129–147. https://doi.org/10.1017/S1352465808004141

Rozek, D. C., Serrano, J. L., Marriott, B. R., Scott, K. S., Hickman, L. B., Brothers, B. M., … & Simons, A. D. (2018). Cognitive behavioural therapy competency: pilot data from a comparison of multiple perspectives. *Behavioural and Cognitive Psychotherapy*, *46*, 244–250. https://doi.org/10.1017/S1352465817000662

Shafran, R., Clark, D. M., Fairburn, C. G., Arntz, A., Barlow, D. H., Ehlers, A., … & Salkovskis, P. M. (2009). Mind the gap: improving the dissemination of CBT. *Behaviour Research and Therapy*, *47*, 902–909. https://doi.org/10.1016/j.brat.2009.07.003.

Streiner, D. L., Norman, G. R. & Cairney, J. (2015). *Health Measurement Scales: A Practical Guide to their Development and Use* (5th edn). USA: Oxford University Press.

Terwee, C. B., Bot, S. D., de Boer, M. R., van der Windt, D. A., Knol, D. L., Dekker, J., … & de Vet, H. C. (2007). Quality criteria were proposed for measurement properties of health status questionnaires. *Journal of Clinical Epidemiology*, *60*, 34–42. https://doi.org/10.1016/j.jclinepi.2006.03.012

Terwee, C. B., Dekker, F. W., Wiersinga, W. M., Prummel, M. F., & Bossuyt, P. M. M. (2003). On assessing responsiveness of health-related quality of life instruments: guidelines for instrument evaluation. *Quality of Life Research*, *12*, 349–362. https://doi.org/10.2307/4038855

Terwee, C. B., de Vet, H. C. W., Prinsen, C. A. C, & Mokkink, L. B. (2011). Protocol for systematic reviews of measurement properties. *COSMIN: Knowledgecenter Measurement Instruments.* Available at: http://www.cosmin.nl/images/upload/files/Protocol%20klinimetrische%20review%20version%20nov%202011(1).pdf (accessed 9 February 2018).

Vallis, T. M., Shaw, B. F., & Dobson, K. S. (1986). The Cognitive Therapy Scale: psychometric properties. *Journal of Consulting and Clinical Psychology*, *54*, 381–385. https://doi.org/10.1037/0022-006x.54.3.381

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*, 149–161. https://doi.org/10.1017/S1352465811000622

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

Weck, F., Hilling, C., Schermelleh-Engel, K., Rudari, V., & Stangier, U. (2011). Reliability of adherence and competence assessment in cognitive behavioral therapy: influence of clinical experience. *Journal of Nervous and Mental Disease*, *199*, 276–279. https://doi.org/10.1097/NMD.0b013e3182124617

Yeaton, W. H., & 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. http://doi.org/10.1037/0022-006X.49.2.156

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

Cite this article: Rayson K, Waddington L, and Hare DJ. The quality of research exploring in-session measures of CBT competence: a systematic review. *Behavioural and Cognitive Psychotherapy*. https://doi.org/10.1017/S1352465821000242