Cognitive Therapy and Research (2019) 43:959–970
https://doi.org/10.1007/s10608-019-10024-z

ORIGINAL ARTICLE

Evaluating CBT Clinical Competence with Standardised Role Plays and Patient Therapy Sessions

Sheena Liness[1] · Sarah Beale[1] · Susan Lea[2] · Suzanne Byrne[1] · Colette R. Hirsch[1] · David M. Clark[3]

Published online: 24 May 2019

Abstract

Standardised role-plays (SR) have been proposed as an alternative to recordings of patients' therapy sessions (PTS) to assess therapist competence during CBT training. This study compared the following properties of SR assessments with established PTS assessments: interrater reliability, responsiveness to training, convergent validity of competence ratings, and predictive validity for academic outcomes. SR and PTS were both rated using the Cognitive Therapy Scale Revised (CTS-R) to assess CBT trainees' ($n = 88$) level of competence at the beginning and end of training, and at one-year follow-up. Both methods demonstrated excellent inter-rater reliability between pairs of course tutors (ICC range = .81–.93) and good reliability between tutors and an external assessor (ICC range = .71–.74). CTS-R scores for both SR and PTS increased across training to reach the competence threshold and remained stable at follow-up. However, there was only a weak relationship between the two assessment methods. Further refinement of SR as a CBT assessment method is indicated.

Keywords Cognitive behaviour therapy (CBT) · Competence · Standardised role play · Patient recordings · IAPT

Introduction

Multinational initiatives to increase public access to evidence-based mental health treatment have led to unprecedented growth in therapist training. These initiatives, including Improving Access to Psychological Therapies (IAPT) in the United Kingdom (Clark 2018), the Department of Veterans' Affairs and Beck Community Initiative in the United States (Creed et al. 2016; Rosen et al. 2017; Stirman et al. 2009), and the Programme to Reduce the Treatment Gap (PRIME) in low- and middle-income countries (Lund et al. 2016), aim to train therapists to competently deliver evidence-based interventions. Therapist competence is benchmarked against a pre-determined criterion to verify that trainees have met an agreed standard of performance, with eligibility to apply for professional accreditation on graduation (e.g. Academy of Cognitive Therapy, 2014; British Association for Behavioural and Cognitive Psychotherapies, 2012). The increased demand for trained accredited therapists, and the responsibility afforded to training courses for quality assurance, makes it important to develop and evaluate valid, reliable and feasible competence assessment methods (Fairburn and Cooper 2011; Muse and McManus 2013).

Assessment of Competence in Therapist Training

Assessment of therapist skill is essential for the evaluation and implementation of training programmes and clinical initiatives. Therapist skill is often conceptualised to comprise two constructs: adherence—the implementation of the relevant therapeutic procedures—and competence—the capable delivery of these therapeutic procedures (Blackburn et al. 2001; Sharpless and Barber 2009). Effective delivery of CBT is regarded to depend on these factors (Barber et al. 2003; Fairburn and Cooper 2011; Muse and McManus 2013). Adherence and competence are closely related, as

Colette R. Hirsch and David M. Clark contributed equally to this work.

✉ Sheena Liness
sheena.liness@kcl.ac.uk

[1] Department of Psychology, Institute of Psychiatry, Psychology, and Neuroscience, King's College London, PO77 Henry Wellcome Building, IoPPN/King's College London, De Crespigny Park, London SE5 8AF, UK

[2] University of Hull, Hull, UK

[3] Department of Experimental Psychology, University of Oxford, Oxford, UK

competent delivery implies adherence, but not vice versa. While some researchers have recommended differentiating between these constructs, they appear highly correlated (Barber et al. 2003) and many validated assessment methods incorporate adherence into overall judgments of competence (e.g. Blackburn et al. 2001). Competence assessment is the focus of this study.

Clinical competence during training is commonly assessed by rating audio or video recordings of patient therapy sessions (PTS) on validated therapy assessment scales (Karlin et al. 2012; Liness et al. 2019; McManus et al. 2010). Competence assessed through PTS appears to increase with training, with the majority of trainees achieving benchmarked standards (Creed et al. 2016; Karlin et al. 2012; Liness et al. 2019; McManus et al. 2010). Former trainees appear to maintain these gains following transition into routine clinical practice (Liness et al. 2018; Simons et al. 2010). Competence, or aspects of competence, assessed through PTS may also demonstrate a relationship with patient clinical outcome. Established empirical literature has linked therapist competence with symptom reduction in the treatment of common disorders, such as those characterized by anxiety and depression. Meta-analytic evidence is mixed when overall competence is assessed, but clearer with a focus on specific competences (e.g., therapist skill in reviewing homework and motivational interviewing) (Webb et al. 2010), and on CBT studies (Zarafonitis-Müller et al. 2014). Evidence for this relationship is strongest in CBT for depression (Webb et al. 2010; Zarafonitis-Müller et al. 2014) and when disorder-specific protocols are assessed (Ginzburg et al. 2012).

Whilst observation of therapists' work with patients is acknowledged as informative and necessary (Miller 1990; Roth and Pilling 2008), difficulties and limitations exist with the established PTS assessment method. PTS submissions require student and work-place compliance, adherence to data protection, and the safe transfer of confidential material (Kaslow et al. 2009). Self-selection of patient recordings may also introduce bias since trainees may submit their strongest—rather than representative—sessions.

The use of PTS to assess trainee competence also introduces variability in patient presentations, which creates an unequal and possibly unfair assessment process (Boswell et al. 2013; DeRubeis et al. 2014). Inconsistency in therapist competence ratings has been identified within and across therapy caseloads in studies of Motivational Enhancement Therapy (Imel et al. 2011) and CBT (Keen and Freeston 2008). Evaluation of therapist competence on standardized role-played (SR) patients in combination with patient sessions has been suggested (Fairburn and Cooper 2011; Schmidt et al. 2018). The current study evaluated SR alongside the traditional PTS sessions for assessment of therapy competence.

Standardised Role-Play (SR)

Standardised assessment of clinical competence is common practice in medical training via objective structured clinical examinations (OSCEs; Epstein 2007; Newble 2004). Standardised assessment in therapy training usually entails a role-played clinical scenario with trainees assessed on the same patient presentation by independent observers using a benchmarked criteria or scale. This enables, in theory, a fairer equitable examination process and may be particularly useful to identify whether therapists implement the appropriate intervention (i.e. adherence to the model or protocol) plus are doing it well (i.e. therapist skill or competence). However, OSCEs have been criticised for the resource commitment involved in development and implementation, the time commitment required from assessors (Kaslow et al. 2009), and for creating anxiety amongst students (Johnson et al. 2018; Yap et al. 2012). Concerns have also been raised that they may not reflect authentic patient scenarios (Sharpless and Barber 2009).

Investigation into standardized role-play (SR) as an assessment method in therapy training is limited. While SR has been used to evaluate CBT low intensity training (Branson et al. 2018), the reliability and validity of the SR assessment itself was not investigated. Two studies provide preliminary indication that SR may hold promise as a CBT assessment method. One study (Sholomskas et al. 2005) used SR sessions to assess the CBT skills of substance use counsellors across three training conditions (manual only, manual + website, or manual + seminar + supervision). SR assessments demonstrated expected patterns of responsiveness to training (Karlin et al. 2012; Liness et al. 2019; McManus et al. 2010), with modest improvements in the manual + seminar + supervision condition, providing some indication that SR may successfully measure CBT skill improvement; direct comparison with PTS is indicated. In addition, a telephone-based SR assessment rated on a new standardised competence assessment rating scale (SCARS-CT) demonstrated excellent inter-rater reliability (ICC = .89) for varied clinical scenarios (Schmidt et al. 2018). The same study reported inadequate inter-rater reliability on the CTS (Young and Beck 1980) for PTS assessment (ICC = .41). The low CTS reliability in this study was unusual and therapist numbers were small. The standardised assessment also took more time. Explicit evaluation of the effectiveness of SR assessment is warranted in CBT training.

Standardised Role-Play (SR) versus Real Patient (PTS) Assessment

Studies comparing SR versus PTS sessions to assess therapist competence have been conducted in Motivational Interview (MI) training with mixed results. An evaluation of 91 MI trainees (Decker et al. 2013) found weak associations between SR and PTS sessions on ratings of adherence and competence ($r = .05–.27$) and poor agreement about which therapists achieved the adequate performance criterion. A dissemination trial of 189 MI therapists (Imel et al. 2014) identified an average relationship between SR and PTS sessions of $r = .40$ (range $r = .04–.75$) for therapist adherence. Between-patient differences accounted for substantially less variance in adherence scores for SR than PTS sessions (Imel et al. 2014). These findings indicate that SR sessions may provide a more equitable assessment of therapists' performance. To establish the reliability and validity of SR assessment in the context of CBT training, direct comparison with PTS—the established gold-standard competence assessment—is required.

King's College London Improving Access to Psychological Therapies (KCL IAPT) Training Course

The King's College London Improving Access to Psychological Therapies (KCL IAPT) course is a 1 year full-time training programme in CBT for depression (behaviour activation and cognitive therapy) and anxiety disorders (social anxiety disorder, panic disorder, post-traumatic stress disorder, health anxiety, obsessive compulsive disorder, generalised anxiety disorder; Roth and Pilling 2008). Trainees are employed in IAPT services for 3 days per week during training and carry a caseload of twelve patients at a time. Work environments and patient demographics vary across services. Trainees are required to record eight training cases for supervision and assessment of clinical competence. Clinical governance and data protection agreements are in place for each service. SR assessments of therapist competence were introduced to address practical challenges with data protection legislation, student avoidance of submitting therapy recordings, patient variability across services and equity of assessment. An increase in trainee numbers and demand on course resources, and the responsibility of acting as a gateway to CBT professional accreditation, also made evaluation of this new assessment method timely and important.

Aims and Objectives

This study aimed to evaluate the reliability and validity of SR for the assessment of competence in CBT training through comparison with PTS. SR and PTS sessions were conducted at the beginning and end of training and at twelve-month follow-up. Ratings of therapist competence on the Cognitive Therapy Scale—Revised (CTS-R; Blackburn et al. 2001) and binary ratings of patient complexity were obtained for each session. Objectives were:

- To evaluate inter-rater reliability between pairs of raters both within SR assessments and within PTS assessments during the course.
- To evaluate agreement of competence ratings between SR and PTS sessions at each training-related key time points (baseline, end-of-training, and post-training follow-up).
- To assess whether competence assessed through SR versus PTS demonstrated similar patterns of change across training.
- To assess the variability in patient complexity between SR and PTS assessments.
- To investigate if SR or PTS assessments of competence demonstrated greater predictive validity for course performance, measured through final overall grade including and excluding PTS results.

Method

Ethics Statement

This study was approved by NHS and university research ethics committee.

Participants

Participants were 88 trainees from three academic years (2012–15) of the KCL IAPT CBT training course. The trainee sample comprised 77.27% ($n = 68$) females and 22.73% ($n = 20$) males, and 86.36% ($n = 76$) were white and 13.64% ($n = 12$) black and minority ethnic. Median age at training was 32.00 years ($IQR = 8$ years). Final awards were Merit (19.32%, $n = 17$), Pass (78.41%, $n = 69$), and Fail/Withdrawn (2.28%, $n = 2$). Trainees professions were Psychological Wellbeing Practitioner (48.86%, $n = 43$), clinical psychologist (22.73%, $n = 20$), counselling psychologist (14.77%, $n = 13$), counsellor (5.68%, $n = 5$), psychotherapist (3.41%, $n = 3$), and other (4.55%, $n = 4$).

Design

This study employed an observational design with SR conducted at three key points (baseline—1 week pre-training, end-of-training—1 week after formal teaching, and 12 month follow up). Trainees' demographic and academic details were collected from course data. Patient therapy

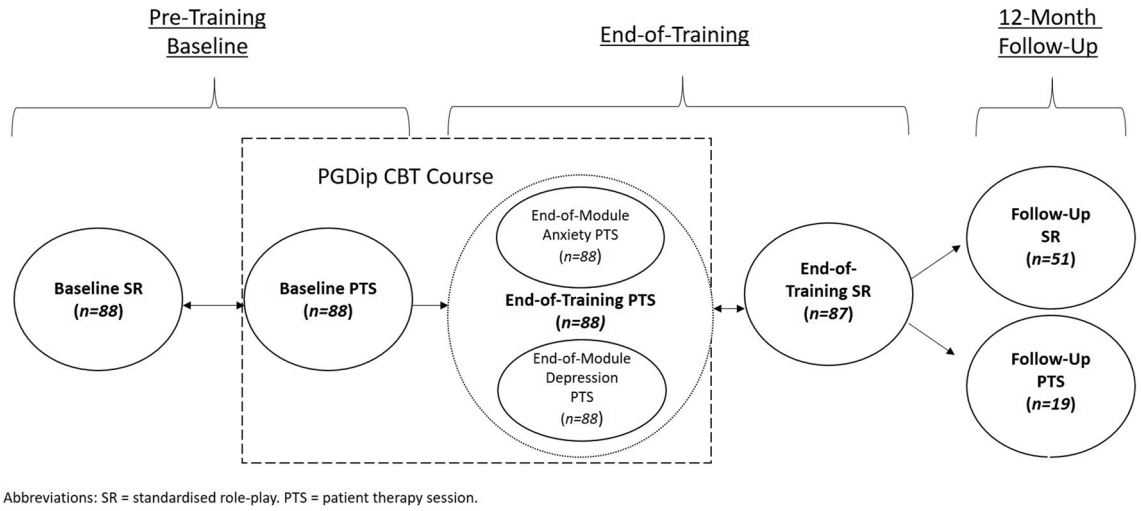

Fig. 1 SR and PTS data collection timeline

recordings (PTS) were collected from course submissions at similar time points. Figure 1 describes the timeline of data collection for SR and PTS.

Measures

Therapeutic Competence

Trainee competence was assessed using the Cognitive Therapy Scale—Revised (CTS-R; Blackburn et al. 2001) for both SR and PTS. The CTS-R comprises 12 items that assess general interpersonal and therapeutic abilities (Items 1–5) and CBT-specific therapeutic skills (Items 6–12). Each item is scored between 0 (non-competent) and 6 (expert), to yield a total CTS-R score between 0–72. The competence threshold, also used as a pass mark on the course in the present study, is set at total score ≥ 36 (Blackburn et al. 2001). Internal consistency is high for the CTS-R (α range = .75–.97; Blackburn et al. 2001; Kazantzis et al. 2018; Reichelt et al. 2003). Inconsistent inter-rater reliability across CTS-R items has been reported; however, good inter-rater reliability for the full scale and for the generic and specific subscales has been achieved using expert and/or trained raters (Kazantzis et al. 2018; Reichelt et al. 2003). All raters in the current study were experienced CBT practitioners and supervisors with accreditation from the British Association of Behavioural and Cognitive Psychotherapy (BABCP). The course trains all new staff on induction to rate the CTS-R by blind marking therapy recordings with experienced markers in order to reach interrater agreement and conducts on-going reviews and reliability monitoring throughout the academic year. All markers in this study had received previous training and acquired extensive experience using the measure for course assessments. The inter-rater reliability of CTS-R ratings between internal course staff and between internal and external markers was assessed for SR and PTS as part of this study and is addressed in the results section.

Patient Complexity

To assess SR actor/research assistant (RA) standardisation and complexity of PTS sessions, a binary rating of patient complexity during the session (complex vs. non-complex) was added by course staff during the second cohort year of this study (2013–2014). Actors/RAs were required to portray a straightforward presentation in SR sessions, in line with predicted complexity of course training cases, i.e. a clear main depression or anxiety presentation with minimal co-morbidity or severe interpersonal or psychosocial problems. To assess reliability of complexity ratings, 83 (20%) of the 418 role-play assessments and submitted therapy tapes were second-rated by a course marker. Overall inter-rater reliability was acceptable (McHugh 2012): $\kappa = .79$. Inter-rater reliability for the 41 second-rated PTS was excellent: $K = 1.00$. It was not possible to calculate Kappa for SR due to lack of variability in complexity ratings; 98% ($n = 41$) of the 42 second-rated SR agreed on non-complexity, while 2% ($n = 1$) were rated complex by one rater and non-complex by the other. This was the only interrater disagreement for the whole sample.

Course Performance

Final overall grade (range = 0–100) based on academic assignments, including a CBT theory essay and written case

reports, was extracted from course records for each student to assess academic performance.

Standardised Role-Play Submission

Baseline SR was conducted 1 week before the course to assess initial CBT competence. End-of-training SR was conducted 1 week following the end of course teaching (month 9 of a 12 month course) to assess post-training competence. Follow-up SR was conducted 12 months after completion of the full course to assess maintenance of CBT competence. SRs were run at the university, and a continuing professional development workshop was incorporated at follow-up to thank participants for their time. The role-play diagnosis (i.e. anxiety or depression) alternated across trainees and time points as competence was required across disorders and to prevent practice effects from repeated assessment. All SRs were rated by course tutors using the CTS-R with feedback returned to trainees within 2 weeks.

Standardised Role-Play

SRs comprised a 30 minute simulated mid-treatment section of a CBT session. Course staff developed the role-play scenarios based on a prototypical client with an acute episode of either depression or panic disorder. Panic disorder was chosen as the representative anxiety presentation for this study due to its high prevalence as a training case and the presence of panic symptoms across multiple anxiety disorders. Diagnostic breakdown for SR at baseline was 68% panic disorder ($n=59$) and 32% depression ($n=28$), at end-of-training was 69% panic disorder ($n=60$) and 31% depression ($n=27$), and at follow-up was 75% panic disorder ($n=38$) and 25% depression ($n=13$). The higher proportion of panic SR is similar to greater representation of anxiety versus depression in course cases, as reported below.[1] Standardised patients were intended to match the complexity of real patients treated by trainees. Patients were portrayed in the SR by either professional actors recruited through a local acting company or by psychology research assistants. Briefing sheets were provided that outlined relevant background information (e.g. diagnosis, the main current difficulties, a recent formulation, work conducted and homework set in the previous session).

All actors/RAs were given the briefing sheet relating to the patient they were representing and attended training which involved a discussion of the training and type of patients seen, observation of a tutor role-playing a typical session, and conducting practise role-plays. Trainees were given the same briefing sheet, sets of recently completed patient outcome measures, a copy of the homework the patients were bringing back with them, and a list of the overall goals of therapy. They had 15 min to make notes and prepare for the role-play, and were able to take notes into the session. The room was set up with therapy resources (e.g. thought records, activity schedules, experiment sheets). All SRs were video-recorded and assessed by a course tutor using the CTS-R.

Course Submissions

As part of course requirements, trainees submitted PTS tapes of representative 50 min mid-treatment sessions for formal assessment. PTS baseline recordings ($n=88$) were submitted within the first month of training to gauge initial CBT skills and comprised a session of treatment for a client with any anxiety disorder or depression. Diagnoses were anxiety disorders (65.52%, $n=57$), and depression (34.48%, $n=30$). Course end-of-training recordings ($n=88$) were an anxiety treatment session submitted at the end of the anxiety module and a depression treatment session submitted at the end of the depression module to formally assess CBT skills in the treatment of both of these disorders. End-of-training competence was operationalised as the mean CTS-R score for the anxiety and depression recordings. Follow-up recordings ($n=19$) were submitted voluntarily by a subset of those who completed the follow-up role-play one-year post-training to assess retention of CBT skills and comprised a recent treatment session. Diagnoses were anxiety disorders (68.42%, $n=13$), and depression (31.58%, $n=6$). Written feedback covering therapist strengths and areas to improve as well as CTS-R overall and item ratings were returned to therapists 2–4 weeks after each submission.

Statistical Methods

Missing Data

SR data were missing for one trainee at the end-of-training. Consequently, $n=87$ trainees were included. There were no other missing data for SR or PTS at baseline or end-of-training, and no missing academic outcome data.

Follow-Up SR Participation

Fifty-one trainees participated in the optional SR at follow-up. A total of 37 therapists did not participate at follow-up due to life events (maternity leave/moved abroad; $n=16$ 43.24%), unavailability on designated date ($n=15$, 40.54%), and non-response ($n=6$, 16.22%). There were no significant differences between follow-up participants

[1] The greater proportion of panic SR occurred due to introduction of depression SR in Year 2 of the study; consequently, course year was controlled in relevant analyses.

and non-participants in gender, age, ethnicity, final award, profession, or baseline CTS-R scores, $p < .05$. There was a significant difference in end-of-training SR scores between follow-up participants ($M = 38.87$, $SD = 3.36$) and non-participants ($M = 36.10$, $SD = 5.12$), $t(85) = -3.05$, $p = .003$, indicating higher post-training competence for participants. SR complexity ratings were available for only 55 tapes at baseline and end-of-training and for 50 tapes at follow-up, as the complexity rating was added during the second cohort year of this study (2013–2014) onwards.

Follow-Up PTS Participation

Follow-up PTS were only available for 19 therapists working in NHS trusts with follow-up ethics approval due to voluntary submission. There was no significant difference in baseline PTS scores between therapists who submitted a follow-up PTS tape and those who did not, $p < .05$, indicating no significant difference in baseline competence. There was a significant difference in end-of-training PTS for trainees who provided a follow-up tape ($M = 39.23$, $SD = 2.42$) and those who did not ($M = 37.29$, $SD = 3.46$), $t(85) = -2.06$, $p = .04$, indicating higher competence at the end of training for participants who submitted a follow-up tape.

Only 15 trainees completed both a follow-up SR and PTS assessment, limiting direct comparison between assessment methods at follow-up to these 15 cases. Where follow-up SR and PTS assessments were not directly compared, all available data were included; thus, reported *n*s at follow-up vary across the results.[2]

Results

Evaluating Inter-Rater Reliability for SR and PTS Assessments

To assess inter-rater reliability, a random selection of ~ 30% ($n = 62$) of the 225 submitted SR tapes were second-marked by course staff blind to the primary rating. Inter-rater reliability between pairs of ten internal markers was excellent for SR: one-way random single-measures intraclass correlation coefficient (ICC) (61, 62) = .81, 95% CI [.70, .88]. For the 193 submitted PTS tapes, internal inter-rater reliability based on a random selection of ~ 30% ($n = 60$) of tapes was also excellent between pairs of the ten internal markers: one-way random single-measures ICC (59, 60) = .93, 95% CI [.88, .96].

[2] The Benjamini–Hochberg Procedure (Benjamini and Hochberg 1995) was applied to findings of all hypothesis tests to correct for multiplicity.

Table 1 CTS-R scores for SR and PTS assessments by time point

	$N_{(SR)}$	SR (*M, SD*)	N_{PTS}	PTS (*M, SD*)
Time point				
Baseline	87	28.49 (5.19)	87	28.28 (4.90)
End-of-training	87	37.72 (4.37)	87	37.62 (3.37)
Follow-up	51	39.09 (3.43)	19	39.74 (4.62)

CTS-R score of 36 or above indicates competence

CTS-R cognitive therapy scale—revised (Blackburn et al. 2001), *SR* standardized patient assessment, *PTS* real patient assessment

To assess inter-rater reliability with external markers, a random selection of ~ 10% ($n = 23$) of SR were marked by an external expert rater blind to trainee identity, other markers' ratings, and session time. Inter-rater reliability between internal markers and external markers was good: one-way random single-measures ICC (22, 23) = .71, 95% CI [.43, .86]. Inter-rater reliability between course markers and external expert markers blinded to trainee identity, internal rating, and session time was also assessed for a random selection of ~ 14% ($n = 27$) of PTS. Inter-rater reliability for course tapes was good between internal and external markers: one-way random single-measures ICC (26, 27) = .74, 95% CI [.51, .87].

Overall, both SR and PTS demonstrated strong interrater reliability with ICCs in the excellent range for reliability between internal raters and ICCs in the good range for reliability between external and internal raters.

Ratings of Therapist Competence Across Training on SR and PTS Assessments

Descriptive statistics for CTS-R outcome at each time point for SR and PTS are presented in Table 1. Means were very similar between SR and PTS at each time point, and were below competence at baseline, and above competence at end-of-training and follow-up.

A mixed analysis of covariance (ANCOVA) was conducted to assess change in CTS-R scores across time (Baseline and End-of-Training) by assessment method (SR vs PTS) controlling for course year. Follow-Up scores were not included in the ANCOVA to prevent loss of data caused by lower numbers at follow-up. There was a significant main effect of time on CTS-R scores: $F_{Time}(1, 171) = 17.93$, $p < .001$, partial $\eta^2 = .09$. A post hoc *t* test found that overall CTS-R scores improved significantly between baseline ($M = 28.38$, $SD = 5.03$) and end-of-training ($M_{Time2} = 37.67$, $SD = 3.89$): $t(173) = -23.12$, $p < .001$, $d = 1.76$. There was no significant main effect of assessment method, $F_{Assessment}(1, 171) = .08$, $p = .78$, partial $\eta^2 = .00$, and no significant interaction, $F_{Time*Assessment}(1, 171) = .02$, $p = .30$, partial $\eta^2 = .01$.

Table 2 Agreement of competence between SR and PTS assessments

Time point	SR competence % (*N*)	PTS competence % (*N*)	Percentage agreement % (*N*)	X^2	*df*	*p*	Φ
Baseline	8.05 (7)	9.20 (8)	82.76 (72)	.77[a]	1	1.0	−.09
End-of-training	70.11 (61)	75.86 (66)	68.97 (60)	4.15	1	.04*	.22
Follow-up	93.33 (14)	80.00 (12)	73.33 (11)	.27[a]	1	1.0	−.13

Only the 15 trainees who submitted a Follow-up SR and PTS assessment are included above for sake of comparison

SR standardized patient assessment, *PTS* real patient assessment

*Sig $p < .05$

[a]Exact X^2

To investigate change in SR CTS-R scores between end-of-training and follow-up for the 51 trainees who participated in follow-up, a paired-samples t-test was conducted. There was no significant change in SR CTS-R scores between end-of-training ($M = 38.87$, $SD = 3.36$) and follow-up ($M = 39.09$, $SD = 3.43$): $t(50) = -.36$, $p = .72$, $d = -.07$. A Wilcoxon signed-rank test was conducted to investigate changes between end-of-training and follow-up for the 19 trainees who submitted a follow-up PTS tape. There was no significant difference between therapy tape CTS-R scores at end-of-training ($Mdn = 38.75$, $IQR = 2.50$) and follow-up ($Mdn = 41.00$, $IQR = 5.50$): $Z = -1.53$, $p = .13$, $r = -.35$.[3]

A Mann–Whitney U-test was conducted to preliminarily assess whether CTS-R scores differed by assessment method at follow-up. For the 15 trainees who had a follow-up SR and PTS session, no significant difference was found between CTS-R scores for follow-up SR ($Mdn = 40.00$, $IQR = 5.00$) versus PTS ($Mdn = 41.00$, $IQR = 5.50$): $U = 393.50$, $p = .28$, $r = -.28$.

In summary, both SR and PTS assessments demonstrated significant improvement between baseline and end-of-training, crossing the competence threshold of the CTS-R. Competence was stable between end-of-training and follow-up and average CTS-R scores remained above the threshold for competence. There was no significant difference between assessment methods at any time point.

Agreement of Competence Ratings for SR and PTS: Association at Each Time Point

Correlations were conducted to assess the relationship between SR and PTS ratings at each time point. At baseline, there was no significant relationship between SR and PTS CTS-R scores: $r(84) = .17$, $p = .11$. There was a significant positive relationship between end-of-training SR and PTS scores: $r(85) = .31$, $p = .004$. There was no significant relationship between follow-up SR and PTS scores: $\rho(13) = .20$, $p = .47$.

Corresponding partial correlations controlling for complexity of PTS patients where these data were available showed similar results: baseline $r(52) = .19$, $p = .17$, end-of-training $r(52) = .35$, $p = .009$), and follow-up $\rho(12) = .10$, $p = .72$.

Agreement of Competence Ratings for SR and PTS: Change Over Time

Correlations were conducted to assess the relationship between SR and PTS change scores (baseline to end-of-training and end-of-training to follow-up). Change scores were investigated as both assessment types were expected to demonstrate comparative levels of improvement between baseline and end-of-training and stability between end-of-training and follow-up. There was no significant relationship between trainees' SR and PTS change scores between baseline and end-of-training, $r(85) = -.06$, $p = .60$, or between end-of-training and follow-up: $\rho(13) = -.40$, $p = .14$. Mean change scores between baseline and end-of-training were 9.24 ($SD = .5.65$) for SR and 9.34 ($SD = .4.96$) for PTS. Mean change for SR between end-of-training and follow-up was .37 ($SD = 4.17$). Median change for PTS between these time points was .50 (IQR = 5.00).

Agreement of Competence Ratings for SR and PTS: Competence Attainment at Each Time Point

Agreement between SR and PTS on overall attainment of competence (CTS-R ≥ 36) was assessed using percentage agreement and Chi square tests. Table 2 reports these findings and the percentage of trainees meeting the competence threshold for SR and PTS at each time point. There was no significant difference in the proportion of SR and PTS rated competent versus non-competent at Baseline and Follow-Up. At End-of-Training, there was a significant difference between SR and PTS, with more trainees achieving competence on the PTS assessment. The majority of tapes

[3] $r = Z/\sqrt{N}$. (Rosenthal, 1991)

Table 3 Complexity ratings for SR and PTS assessments

	SR *n*	Complex SR *%, n*	PTS *n*	Complex PTS *%, n*
Baseline	55	.00% (0)	55	14.55% (8)
End-of-Training	55	1.82% (1)	55	7.27% (4)
Follow-Up	49	.00% (0)	19	26.32% (5)

SR standardized patient assessment. *PTS* real patient assessment

agreed on competence at all time points, with the median absolute difference score between SR and PTS ranging from 2.50–4.00.

The median absolute differences between SR and PTS scores at each time point for the whole sample were: baseline = 4.00 (IQR = 5.50), end-of-training = 2.75 (IQR = 3.75), and follow-up = 2.50 (IQR = 5.50). Median absolute differences between SR and PTS tapes that disagreed on overall level of competence were: baseline = 9.00 (IQR = 9.00), end-of-training = 4.50 (IQR = 4.25), and follow-up = 7.25 (IQR = 8.63). Mann–Whitney U-tests found that cases that disagreed had significantly greater median difference scores than the overall sample at all time points, $p < .006$.

In sum, SR and PTS assessments demonstrated a significant disagreement in classification of competence at end-of-training only; however, percentage agreement was > 69% at all time points and median differences between paired SR and PTS scores were relatively small for the overall cohort at all time points.

Consistency of Patient Presentation Complexity for SR and PTS Assessments

Patient complexity for SR and PTS sessions was assessed to investigate SR standardisation and complexity ratings across assessments. Table 3 reports proportions of complex cases for SR and PTS at each time point. Complexity ratings were available for 55 SR and PTS at baseline and end-of-training, and for 49 SR and 19 PTS at follow-up due to introduction of complexity ratings in the second cohort of the study. SR demonstrated a lower proportion of complexity with only one session rated as complex across the three time points.

A total of 74 SR sessions were conducted by research assistants (27 sessions respectively at baseline and end-of-training, and 20 sessions at follow-up), of which only one session (1.35%)—conducted by an RA—exceeded the complexity threshold. A total of 86 sessions were conducted by actors (28 sessions respectively at baseline and end-of-training, and 30 sessions at follow-up), of which no sessions exceeded the complexity threshold.

Table 4 Correlations between CTS-R scores and final grade by assessment method

	Final grade *r, p*	Final grade (end-of-training tapes excluded) *r, p*
SR		
End-of-training ($n = 87$)	.30, $p = .005$	.18, $p = .09$
PTS		
End-of-training ($n = 87$)	.67, $p < .001$	.47, $p < .001$

CTS-R cognitive therapy scale—revised (Blackburn et al. 2001)

Overall, SR sessions were largely well-standardised. Proportions of PTS sessions exceeding the recommended complexity threshold for course cases varied between 7 and 26% across key time points.

Predictive Validity of SR and PTS Assessments for Final Course Grade

SR and PTS predictive validity for academic outcomes were addressed through Pearson's correlations between end-of-training CTS-R scores—representative of formal competence assessment—and final grade. Relationships were tested including and excluding anxiety and depression end-of-training PTS therapy assessment results. Mean final grade with all assignments included was 56% ($SD = 3\%$), and mean final grade with PTS tapes removed was 57% ($SD = 5\%$). Results are presented in Table 4.

In summary, both SR and PTS exhibited a significant positive relationship with final grade when end-of-training PTS tapes were included. PTS exhibited a relationship with final grade with end-of-training PTS tapes excluded.

Discussion

This is the first study to assess CBT competence with SR and real patient sessions (PTS) before and after CBT training and at twelve-month follow-up. The study has several strengths. Both assessment methods were evaluated at all time points across three training cohorts and inter-rater reliability was examined using the CTS-R (Blackburn et al. 2001), a validated rating scale. The study also investigated whether patient complexity was in-line with course requirements across SR and PTS clinical assessments and looked at the relationship of each assessment method with overall course performance.

Inter-Rater Reliability

Both SR and PTS demonstrated excellent internal inter-rater reliability (ICC = .81 and .93) and good external inter-rater reliability (ICC = .71 and .74). Findings correspond with recent reports of high inter-rater reliability for SR (Schmidt et al. 2018) and PTS assessment (Kazantzis et al. 2018) of cognitive therapy. Favourable inter-rater reliability in this study may be the result of a CBT team of experienced clinicians with considerable experience of training and marking together (Liness et al. 2018; Mortsiefer et al. 2017). Comprehensive in-house training and on-going reliability monitoring is routine practice for all markers. These findings provide further support for the use of the CTS-R as a reliable measure to assess therapist competence when scored by trained and experienced raters (Kazantzis et al. 2018).

CBT Competence

Both SR and PTS sessions evidenced significant increase in CTS-R scores across training, from below competence at baseline to above competence at the end of training. Competence was maintained with no significant change between end-of-training and twelve-month follow-up. Findings corroborate evidence of increased competence with CBT training (Liness et al. 2019; McManus et al. 2010) that is sustained at follow-up (Liness et al. 2018; Simons et al. 2010). CTS-R means (see Table 1) were very similar for SR and PTS, with less than one-point difference at any time point. However, the lack of significant association between CTS-R change scores for SR and PTS sessions raises questions about the validity of SR assessment responsiveness to training.

SR and PTS Agreement

(1) *Time-Points* Little evidence of a robust relationship between CTS-R scores for SR and PTS was found. There was no significant association at baseline or follow-up, while a significant but weak relationship emerged at the end-of-training, the formal examination time-point. Results remained similar when complexity of PTS tapes was controlled, indicating that low agreement on CTS-R scores was not driven by greater complexity in PTS sessions.

Weak associations could be influenced by study methodology, namely small differences in the SR and PTS data collection timeline. The follow-up analysis was likely underpowered to detect a significant small association ($r = .20$) due to only fifteen therapists contributing both a follow-up SR and PTS session.

(2) *Change Over Time* There was no relationship between SR and PTS change scores from baseline to the end-of-training or between end-of-training and follow-up, although median change scores were similar. These findings indicate that, while SR and PTS demonstrate similar CTS-R scores and change scores across the whole sample, agreement between trainees' pairs of SR and PTS scores exhibited no association in terms of the degree of change over time, indicating poor convergent validity.

(3) *Competence Classification* Agreement in ratings of competence (CTS-R ≥ 36) were compared directly between SR and PTS at each time point. While median absolute difference scores between SR and PTS at all time points were relatively low—indicating little difference in scores overall—they were significantly greater for cases where the SR and PTS tape disagreed on competence classification. A small but significant difference emerged at the end of training, with more trainees achieving competence on the PTS assessment—a notable disparity for the formal pass/fail examination. It is possible SR enabled markers to be clearer about adherence to protocol and the quality of therapy demonstrated with the noise of varied patient presentations removed. However, as the optimum content, duration, and delivery of SR roleplay for CBT assessment has not yet been established, it is also possible this SR role play was just too difficult. Previous studies have reported increased anxiety amongst student participants of OSCEs in clinical psychology (Johnson et al. 2018; Yap et al. 2012)—possibly affecting performance in formal exams. Higher competence ratings for end-of-training PTS tapes could also reflect a trainee self-selection bias, with the strongest PTS sessions submitted for formal examination. However, selection of strong PTS may also reflect trainee awareness of appropriate 'good' sessions for submission. Direct assessment of perceived anxiety, varied content and duration of SR clinical scenarios, and investigation into difficulties encountered for both assessment methods during formal examination is recommended for future studies.

Patient Complexity and Standardisation

SR sessions demonstrated good standardisation at all time points, with only one session rated as portraying a complex patient. Both actors and research assistants (RAs) delivered well-standardised sessions, possibly assisted by the SR training and guidance. The course team conducted spot checks for actor adherence during the role-play process, but how much quality checking is required is unclear. A higher proportion of PTS exceeded the complexity threshold, particularly at follow-up, indicating more patient variability in real sessions consistent with the literature (Boswell et al. 2013; Imel et al. 2011). Greater variability in PTS patient

presentation, however, did not appear to influence SR and PTS agreement.

Predictive Validity for Academic Outcomes

End-of-training PTS sessions demonstrated a moderate positive relationship with final grade, indicating predictive validity for academic outcomes. While SR sessions demonstrated a weak positive relationship with final grade including PTS tape results, the relationship became non-significant when final PTS course grade results were removed. SR may therefore not demonstrate additional predictive validity beyond its relationship with PTS ($r = .31$). The strength of the correlations may have been limited by the restricted range of students' final grade.

General Discussion

The pros and cons of implementing SR assessment have been documented in medical training (Frye et al. 1989; Vu and Barrows 1994), and therapy training (Decker et al. 2013; Kaslow et al. 2009). Our study provides further perspective on the benefits and challenges of SR assessment applied to CBT training.

Standardisation and Delivery

An accessible SR that is consistent across all trainees may create a fairer assessment and address logistical issues around patient consent, data protection and session self-selection (Kaslow et al. 2009). Standardisation was achieved in this study and concerns that SR may be excessively simplistic (Kaslow et al. 2009; Muse and McManus 2013; Sharpless and Barber 2009) were not supported. Preparation and organisation of SR, however, were resource and time-intensive on an already over-stretched training course. The challenge of balancing validity plus feasibility of new methods is important (Schoenwald et al. 2011). Whilst PTS is also resource and time intensive, it is an established process for trainers and demonstrated stronger evidence of validity as an assessment method in the current study.

Comparison with PTS Assessment

Mixed outcomes on agreement of overall competence and the weak or non-significant relationships between SR and PTS CTS-R scores and change scores made it difficult to establish that SR and PTS make the same judgements about trainee ability. As such, replacing PTS assessments with SR cannot be recommended from these findings despite logistical benefits. PTS demonstrated a stronger association with final course outcome assessed on a range of academic assignments. Comparison of the predictive validity of SR and PTS competence ratings and trainees' patient clinical outcomes was not feasible in this study and is important for future research, if conclusively valid SR assessments are developed. A combination of SR and PTS assessments may increase accessibility of clinical assessments across a range of patient presentations (Schmidt et al. 2018), enable a more reliable estimate of competence (Imel et al. 2014) and allow additional constructive feedback prior to work with patients (Miller 2010).

Follow-Up Participation

Substantially more trainees returned to participate in the SR session compared to the PTS session at follow-up. Therapists who provided SR and/or PTS follow-up sessions in this study demonstrated significantly higher end-of-training competence, indicating weaker therapists appear less willing to be re-evaluated. Previous difficulties of attrition with PTS recordings have been reported (Liness et al. 2018; Miller et al. 2004). On-going professional accreditation makes follow-up assessment procedures for therapist graduates important. This study indicates SR may be more practicable to engage busy clinicians. SR may also be useful for skill maintenance and evaluation in former trainees with limited access to patients due to changing role requirements.

Limitations

This was a naturalistic study conducted on an established training course with no control condition. Data relied on PTS submissions with trainees self-selecting patient recordings. The assessment of multiple and/or randomly-selected recordings, whilst not feasible in this study, may be beneficial to counter any selection bias. Time points varied slightly for SR and PTS recordings. There was a lack of information to evaluate the relationship of therapy competence to clinical outcome. Patient complexity data was limited as the complexity rating was only added from the second year of the study. Findings at follow-up were limited by small number of tapes so should be viewed with caution. SR sessions focused on two scenarios and future studies would benefit from role-plays of multiple presentations. The brevity of SR session length may have provided insufficient time for reliable scoring of all elements of the CTS-R, as the measure is intended for full-length CBT sessions. Trainees at follow up demonstrated significantly higher competence at the end-of-training than non-participants, possibly indicating self-selection bias in the follow-up sample and thus requiring caution interpreting findings.

Conclusion

This study provides an important contribution to the CBT training literature as the first to explicitly evaluate the use of SR sessions for assessing CBT therapy competence. SR assessment was compared to existing PTS recordings, with both methods examined on the CTS-R (Blackburn et al. 2001). Both assessment methods demonstrated robust inter-rater reliability, responsiveness to improvement with training and maintenance of gains at follow-up. However, the convergent validity of SR with PTS sessions remains unclear. Notably, relatively poor agreement between SR and PTS CTS-R scores did not appear to be impacted by greater PTS complexity. Consequently, implementing SR on a wider scale or replacing PTS assessments is not recommended until a conclusive relationship between SR and PTS is established. When and how to effectively use SR and PTS to assess CBT trainee competence, and which method is the better predictor of patient outcome needs to be explored in future research.

Acknowledgements With many thanks to all the therapists and patients who participated in this study, and the IAPT services who granted access to information. To research assistants Hannah Parker and Steffen Nestler for their help with data collection and management.

Funding This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. Dr. Colette Hirsch receives salary support from the National Institute for Health Research (NIHR), Mental Health Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London.

Compliance with Ethical Standards

Conflicts of Interest Sheena Liness and Suzanne Byrne run the IAPT CBT training at the IoPPN/KCL course that is the subject of this study. David M. Clark is NHS England's Clinical Advisor for the IAPT programme. Susan Lea, Colette Hirsch and Sarah Beale have no conflict of interest with respect to this publication.

Informed Consent Informed consent was obtained from all individual participants included in the study.

Animal Rights No animal studies were carried out by the authors for this article.

References

Academy of Cognitive Therapy. (2014). *Certification: candidate handbook*. Retrieved from https://cdn.ymaws.com/www.academyofct.org/resource/collection/68F74ABA-27C5-47DD-8F0F-E6E8FB3C710D/Candidate_Handbook.pdf.

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.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B, 57,* 289–300.

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

Boswell, J. F., Gallagher, M. W., Sauer-Zavala, S. E., Bullis, J., Gorman, J. M., Shear, M. K., et al. (2013). Patient characteristics and variability in adherence and competence in cognitive-behavioral therapy for panic disorder. *Journal of Consulting and Clinical Psychology, 81*(3), 443–454.

Branson, A., Myles, P., Mahdi, M., & Shafran, R. (2018). The relationship between competence and patient outcome with low-intensity cognitive behavioural interventions. *Behavioural and Cognitive Psychotherapy, 46*(1), 101–114.

British Association for Behavioural and Cognitive Psychotherapies. (2012). *BABCP minimum training standards for the practice of cognitive behavioural therapy*. Retrieved from https://www.babcp.com/files/Accreditation/General/Minimum-Training-Standards-V7-0215.pdf.

Clark, D. M. (2018). Realizing the mass public benefit of evidence-based psychological therapies: the IAPT program. *Annual Review of Clinical Psychology, 14,* 159–183.

Creed, T. A., Frankel, S. A., German, R. E., Green, K. L., Jager-Hyman, S., Taylor, K. P., et al. (2016). Implementation of transdiagnostic cognitive therapy in community behavioral health: the beck community initiative. *Journal of Consulting and Clinical Psychology, 84*(12), 1116.

Decker, S. E., Carroll, K. M., Nich, C., Canning-Ball, M., & Martino, S. (2013). Correspondence of motivational interviewing adherence and competence ratings in real and role-played client sessions. *Psychological Assessment, 25*(1), 306–312.

DeRubeis, R. J., Gelfand, L. A., German, R. E., Fournier, J. C., & Forand, N. R. (2014). Understanding processes of change: how some patients reveal more than others—and some groups of therapists less—about what matters in psychotherapy. *Psychotherapy Research, 24*(3), 419–428.

Epstein, R. M. (2007). Assessment in medical education. *New England Journal of Medicine, 356*(4), 387–396.

Fairburn, C. G., & Cooper, Z. (2011). Therapist competence, therapy quality, and therapist training. *Behaviour Research and Therapy, 49*(6), 373–378.

Frye, A. W., Richards, B. F., Philp, E. B., & Philp, J. R. (1989). Is it worth it? A look at the costs and benefits of an OSCE for second-year medical students. *Medical Teacher, 11*(3–4), 291–293.

Ginzburg, D. M., Bohn, C., Höfling, V., Weck, F., Clark, D. M., & Stangier, U. (2012). Treatment specific competence predicts outcome in cognitive therapy for social anxiety disorder. *Behaviour Research and Therapy*, *50*(12), 747–752.

Imel, Z. E., Baer, J. S., Martino, S., Ball, S. A., & Carroll, K. M. (2011). Mutual influence in therapist competence and adherence to motivational enhancement therapy. *Drug and Alcohol Dependence, 115*(3), 229–236.

Imel, Z. E., Baldwin, S. A., Baer, J. S., Hartzler, B., Dunn, C., Rosengren, D. B., et al. (2014). Evaluating therapist adherence in motivational interviewing by comparing performance with standardized and real patients. *Journal of Consulting and Clinical Psychology, 82*(3), 472–481.

Johnson, H., Mastroyannopoulou, K., Beeson, E., Fisher, P., & Ononaiye, M. (2018). An evaluation of multi-station Objective

Structured Clinical Examination (OSCE) in clinical psychology training. *Clinical Psychology Forum, 301,* 38–43.

Karlin, B. E., Brown, G. K., Trockel, M., Cunning, D., Zeiss, A. M., & Taylor, C. B. (2012). National dissemination of cognitive behavioral therapy for depression in the department of veterans affairs health care system: therapist and patient-level outcomes. *Journal of Consulting and Clinical Psychology, 80*(5), 707–718.

Kaslow, N. J., Grus, C. L., Campbell, L. F., Fouad, N. A., Hatcher, R. L., & Rodolfa, E. R. (2009). Competency assessment toolkit for professional psychology. *Training and Education in Professional Psychology, 3*(4S), S27–S45.

Kazantzis, N., Clayton, X., Cronin, T. J., Farchione, D., Limburg, K., & Dobson, K. S. (2018). The cognitive therapy scale and cognitive therapy scale-revised as measures of therapist competence in cognitive behavior therapy for depression: relations with short and long term outcome. *Cognitive Therapy and Research, 42*(4), 385–397.

Keen, A. J., & Freeston, M. H. (2008). Assessing competence in cognitive-behavioural therapy. *The British Journal of Psychiatry, 193*(1), 60–64.

Liness, S., Beale, S., Lea, S., Byrne, S., Hirsch, C. R., & Clark, D. M. (2018). The sustained effects of cbt training on therapist competence and patient outcomes. *Cognitive Therapy and Research, 43,* 1–11.

Liness, S., Beale, S., Lea, S., Byrne, S., Hirsch, C. R., & Clark, D. M. (2019). Multi-professional IAPT CBT training: clinical competence and patient outcomes. *Behavioural and Cognitive Psychotherapy, 28,* 1–14.

Lund, C., Tomlinson, M., & Patel, V. (2016). Integration of mental health into primary care in low-and middle-income countries: the prime mental healthcare plans. *British Journal of Psychiatry, 208*(56), s1–s3.

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia Medica, 22*(3), 276–282.

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*(11), 1123–1132.

Miller, G. E. (1990). The assessment of clinical skills/competence/performance. *Academic Medicine, 65*(9), S63–S67.

Miller, J. K. (2010). Competency-based training: objective structured clinical exercises (OSCE) in marriage and family therapy. *Journal of Marital and Family Therapy, 36*(3), 320–332.

Miller, W. R., Yahne, C. E., Moyers, T. B., Martinez, J., & Pirritano, M. (2004). A randomized trial of methods to help clinicians learn motivational interviewing. *Journal of Consulting and Clinical Psychology, 72*(6), 1050–1062.

Mortsiefer, A., Karger, A., Rotthoff, T., Raski, B., & Pentzek, M. (2017). Examiner characteristics and interrater reliability in a communication OSCE. *Patient Education and Counseling, 100*(6), 1230–1234.

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

Newble, D. (2004). Techniques for measuring clinical competence: objective structured clinical examinations. *Medical Education, 38*(2), 199–203.

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.

Rosen, R. C., Ruzek, J. I., & Karlin, B. E. (2017). Evidence-based training in the era of evidence-based practice: challenges and opportunities for training of PTSD providers. *Behaviour Research and Therapy, 88,* 37–48.

Rosenthal, R. (1991). *Meta-analytic procedures for social research* (Vol. 6). London: Sage Publishing.

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*(02), 129–147.

Schmidt, I. D., Strunk, D. R., DeRubeis, R. J., Conklin, L. R., & Braun, J. D. (2018). Revisiting how we assess therapist competence in cognitive therapy. *Cognitive Therapy and Research, 42,* 1–16.

Schoenwald, S. K., Garland, A. F., Chapman, J. E., Frazier, S. L., Sheidow, A. J., & Southam-Gerow, M. A. (2011). Toward the effective and efficient measurement of implementation fidelity. *Administration and Policy in Mental Health and Mental Health Services Research, 38*(1), 32–43.

Sharpless, B. A., & Barber, J. P. (2009). A conceptual and empirical review of the meaning, measurement, development, and teaching of intervention competence in clinical psychology. *Clinical Psychology Review, 29*(1), 47–56.

Sholomskas, D. E., Syracuse-Siewert, G., Rounsaville, B. J., Ball, S. A., Nuro, K. F., & Carroll, K. M. (2005). We don't train in vain: a dissemination trial of three strategies of training clinicians in cognitive-behavioral therapy. *Journal of Consulting and Clinical Psychology, 73*(1), 106–115.

Simons, A. D., Padesky, C. A., Montemarano, J., Lewis, C. C., Murakami, J., Lamb, K., et al. (2010). Training and dissemination of cognitive behavior therapy for depression in adults: a preliminary examination of therapist competence and client outcomes. *Journal of Consulting and Clinical Psychology, 78*(5), 751–756.

Stirman, S. W., Buchhofer, R., McLaulin, J. B., Evans, A. C., & Beck, A. T. (2009). Public-academic partnerships: the beck initiative: a partnership to implement cognitive therapy in a community behavioral health system. *Psychiatric Services, 60*(10), 1302–1304.

Vu, N. V., & Barrows, H. S. (1994). Use of standardized patients in clinical assessments: recent developments and measurement findings. *Educational Researcher, 23*(3), 23–30.

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

Yap, K., Bearman, M., Thomas, N., & Hay, M. (2012). Clinical psychology students' experiences of a pilot objective structured clinical examination. *Australian Psychologist, 47*(3), 165–173.

Young, J. E., & Beck, A. T. (1980). *Cognitive therapy scale: rating manual*. Unpublished Manuscript, University of Pennsylvania, Philadelphia.

Zarafonitis-Müller, S., Kuhr, K., & Bechdolf, A. (2014). The relationship between therapist's competence and adherence to outcome in cognitive-behavioural therapy—results of a meta-analysis. *Fortschritte der Neurologie Psychiatrie, 82*(09), 502–510.