Psychometric evaluation of therapist competency rating scales

Lucy Hughes

Submitted for the award of Doctorate in Clinical Psychology

Clinical Psychology Unit Department of Psychology

The University of Sheffield

November 2017

This page is intentionally blank

Access to Thesis Form

Declaration

This thesis has been submitted for the award of Doctorate in Clinical Psychology at The University of Sheffield. It has not been submitted for any other qualification or to any other academic institution.

Word Count

Literature review	7,974
Including references	10,143
Research report	10,553
Including references	12, 177
Appendices	6,663
Total word count	28,983
Excluding references and appendices	18,527

Abstract

Literature Review

A systematic review of the psychometric properties and quality of scales measuring therapist competency in delivering psychotherapy to adults was conducted. Thirteen studies met the a priori criteria and were included in the final analysis. The results showed seven therapist competency rating scales had good reliability and validity. All studies tested the interrater reliability of scales, but limited evidence was provided for validity. The psychometric methodology between studies was inconsistent. Most scales were applicable to high-intensity CBT practice, or for specific treatment with drug-dependent patients. Further research is needed to develop psychometrically valid and reliable therapist competency rating scales for a range of theoretical therapeutic approaches and mental health conditions.

Research Report

The research report provided a psychometric evaluation of the Psychological Wellbeing Practitioner Competency rating Scale for Assessment (PWPCS-A) and Treatment (PWPCS-T). The scales measure practitioner competency in delivering low-intensity CBT treatments for patients with mild to moderate anxiety or depression. Data was utilised from PWPCS-A and PWPCS-T ratings from 176 expert, qualified, and novice psychological wellbeing practitioners (PWPs). Further analysis of reliability, and validity was determined from data collected from 114 PWP trainees' Observed Structured Clinical Examinations. The PWPCS-A showed excellent reliability and validity, and the PWPCS-T demonstrated acceptable results. The research provides support for the use of the PWP competency scales for PWP training. Limitations, clinical implications, and future research are discussed.

Acknowledgements

I would like to take this opportunity to express my thanks and gratitude for the people who have helped in this process. Firstly, thank you to Steve Kellett, my research supervisor, for his patience and guidance in producing this thesis. I would like to thank all the PWPs who agreed to be part of this study, and all the trainers, the actors, and all practitioners involved. I also appreciate the contributions made by Jennie Hague, Ellie Hutchinson, Sally Dawson, Mel Simmonds-Buckley, and Emma Limon.

I would also like to express my gratitude to all my family and friends for all their support and encouragement throughout this process. A special thank you goes to my husband, Sam, and my beautiful children.

Table of Contents

Table of Contents (continued)

Section Two: Research Report

A psychometric evaluation of the Psychological Wellbeing Practitioner Competency Rating Scale (PWPCS).

Appendices

Section One: Literature Review

Psychometric Properties of Therapist Competency Rating Scales: A Systematic Review

Abstract

Purpose

Ensuring therapist competency is crucial in providing safe, quality and appropriate treatment for people with mental health concerns. There is currently no evaluation of the psychometric quality of assessments of therapist competency. The purpose of this review was to critically appraise and evaluate the psychometric properties and methodological quality of rating scales used to assess therapist competency in delivering psychotherapy (regardless of theoretical approach).

Method

A systematic review of the literature on the psychometric properties of scales that aim to measure therapist competence was performed using Medline, Scopus, Web of Science, and PsychINFO databases. The psychometric quality was determined using the COSMIN checklist (Terwee et al., 2011).

Results

Thirteen studies met the a priori criteria and were included in the final analysis. All measures showed evidence of interrater reliability, though variability in acceptability of results. The results of studies evaluating validity were limited in number and quality. Most scales were applicable to high-intensity CBT, or for the treatment of drug use. There was a disparity in methods used to determine psychometric quality.

Conclusion

Overall, there is a lack of consistency in the psychometric methodological quality of therapist competency rating scales.

Practitioner Points

- The review provides an overview of therapist competency rating scales and their psychometric properties.

- Therapist competency scales should be psychometrically evaluated, and include analyses of reliability and validity.
- There should be consistency in the methods of psychometric assessment of therapist competency rating scales.
- Scales need to be developed for a range of therapeutic approaches, and mental health conditions.
- The definition and interpretation of therapist competency needs further clarification.

Introduction

Competence as a Construct

Therapist competence is defined as an attribute based on knowledge and skill in delivering therapy to a standard that is effective (Bennett & Parry, 2004; Fairburn & Cooper, 2011). The literature on therapist competence identifies two types: global and limited-domain (Barber, Sharpless, Klostermann & McCarthy, 2007). Global competence refers to skills independent to the therapeutic intervention model and includes the ability to promote a strong alliance and collaboration with the patient (Southam-Gerow & McLeod, 2013). Limited-domain competence refers to the ability to deliver appropriate specific therapy components (Barber et al., 2007).

Norman (1985) described five domains of professional competencies needed for psychotherapeutic practice. These include ensuring a therapist has: knowledge and understanding; technical skills; clinical skills; clinical judgment and problem solving skills; and personal attributes. Roth and Pilling (2007) developed a framework for the Centre for Outcomes, Research and Effectiveness (CORE) of essential competencies for Cognitive Behavioural Therapy (CBT). These included five domains: basic CBT competencies; specific behavioural competencies; problem specific competencies; global competencies; and meta-competencies. Sperry (2010) stated there are six core competencies used in psychotherapy, which are skills in: conceptual foundations; culturally and ethnically sensitive practice; intervention planning; relationship building and maintenance; intervention implementation; and evaluation and termination.

Therapist Competence and Patient Outcomes

The results of studies on therapist competence and patient outcomes are variable, with some showing therapist competency significantly impacted on patient-rated change (O'Malley et al., 1988; Davidson et al., 2004; Strunk, Brotman, DeRubeis, & Hollon, 2010), and others showing limited support for the relationship between competence and

outcomes (Shaw et al., 1999; Branson, Shafran, & Myles, 2015; Hogue et al., 2008). A meta-analytic review was conducted by Webb, DeRubeis and Barbers (2010) on the effect of both therapist adherence and competence on patient outcomes. The results of 17 included studies showed there was no significant effect (from weighted means) for competence. However, the sample size was small and was limited by the paucity of assessment methods to measure therapist competency, thus highlighting the need for valid and reliable assessments of psychotherapeutic competence to allow more in-depth investigation of the process mechanisms that could influence patient success in treatment (Bennett & Parry, 2004).

Assessment of Therapist Competence

Plumb and Vilardaga (2010) state that an assessment of competency should measure whether a therapist can address client need, show responsiveness to treatment targets, and apply therapeutic procedures. It should include an assessment of knowledge of treatment and ability to apply such knowledge skillfully (Cooper et al., 2017). Methods should include a way of incorporating an assessment of a range of both global and specific competencies to demonstrate therapist ability to deliver therapeutic treatment to an acceptable standard (Barber et al., 2007; Bennett & Parry, 2004; Fairburn & Cooper, 2011).

Assessing competence plays an important role in the recognition and development of therapists' ability to deliver psychological treatments (Fairburn & Cooper, 2011). Therapists should be trained to a competent level in order to deliver evidence-based psychological therapy and patient care that is appropriate and helpful. Ensuring that treatment is given in a competent manner is a professional and ethical responsibility when working with people with mental health concerns (Sharpness & Barber, 2009).

Kohrt et al. (2015) stated that a lack of valid and reliable measures of competency is a barrier to ensuring therapists can deliver evidence-based psychological therapy. Competence measures are crucial in evaluating outcomes of treatment efficacy, developing and refining training and supervision models, as well as disseminating psychological therapy interventions in a real life context (Kohrt et al.). Research validity in therapy would be questionable if interventions were not delivered competently (Bennett & Parry, 2004; Fairburn & Cooper, 2011; Muse & McManus, 2013).

Methods of Competence Assessment

A range of methods for determining therapist competence have been suggested and utilised in training and clinical practice. These include patient evaluation of the session, therapist self-assessment, standardised role play (e.g. Objective Structured Clinical Examinations, OSCEs); or clinical practice assessments using rating scales (Fairburn & Cooper, 2011). Using patient evaluations may identify what was helpful (or unhelpful) during therapy and how this impacts of treatment efficacy, however, they neglect the influence of patient related factors, such as problem severity (Rakovshik & McManus, 2010). Brosan, Reynolds and Moore (2008) found that therapists' self-assessment of competence was often overly optimistic and not a true representation of capability, and this was particularly prevalent in less competent therapists.

Competency rating assessment of either OSCEs or clinical practice provides an effective overview of treatment delivery (Fairburn & Cooper, 2011). Several rating scales have been developed to assess therapist competency in delivering a range of psychotherapeutic interventions for different mental health concerns.

Limitations of Current Assessment Methods Measuring Therapist Competency

Fairburn and Cooper (2011) explain that there is very little research on the assessment methods of therapeutic competence and state the need to evaluate the content, reliability, validity, and operationalisation of these measures. Further

psychometric evaluation of rating scales is needed to determine how best to assess therapist competency (Muse and McManus, 2013). To date there has not been a systematic review on the psychometric quality of therapist competency rating scales.

Study Aim

The aim of this review was to critically appraise and evaluate the psychometric properties and methodological quality of rating scales used to assess therapist competency in delivering psychotherapy to adults (regardless of theoretical approach).

Method

Search Process

The PRISMA statement checklist contains a total of 27 essential item areas for transparent reporting of systematic reviews (Liberati, 2009). The checklist was utilised throughout this review and the PRISMA diagram is shown in Figure 1.

Inclusion Criteria

Studies were included in the review if the studies contained: (i) a psychometric evaluation of a rating scale; (ii) an investigation into the competence of therapists (or trainee therapists) during psychotherapy sessions; (iii) an inclusion of a quantifiable competency rating scale; (iv) an assessment of competence that had been videotaped, audiotaped, or observation of therapy sessions rated by trained or expert raters, rather than by patients or therapists; (v) ratings by at least two assessors.

Exclusion Criteria

Studies were excluded if studies: (i) did not explicitly measure therapist competency; (ii) did not distinguish between adherence and competency; (iii) were trials examining the impact of interventions, unless they also reported a psychometric evaluation of a rating scale; (iv) did not specify a theoretical psychotherapeutic approach to treatment intervention; (v) related to scales for therapists treating children and young people; (vi) were dissertation abstracts, articles from non-peer reviewed journals, or unpublished studies.

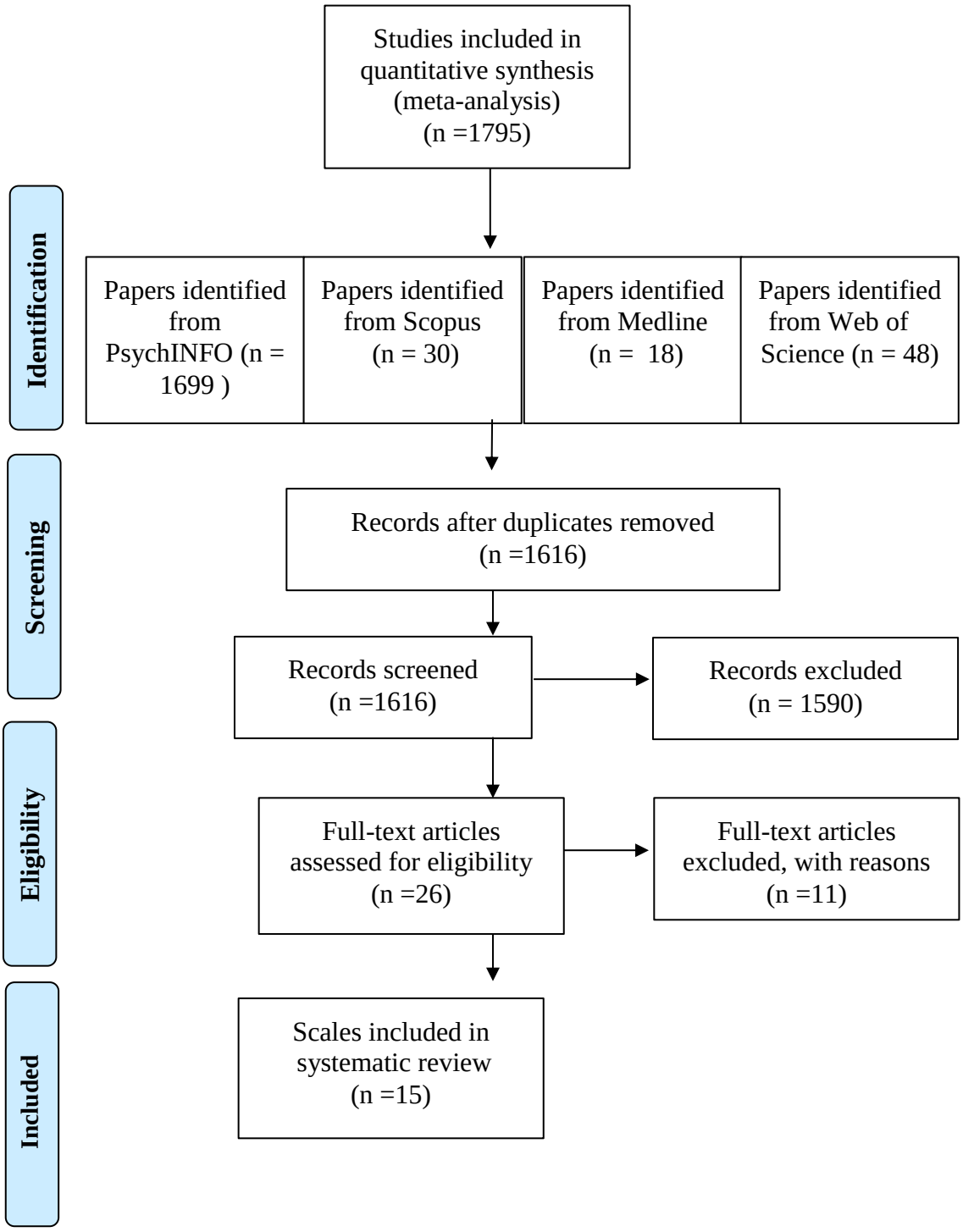


Figure 1: PRISMA flow chart

Search Strategy

The following electronic databases were searched in March 2017: PsychInfo (via OvidSP) 1806 to 2017, Web of Science (via OvidSP) 1864 to March 2017, Scopus, and Medline. The search terms used were 'Therapist', 'Competenc*', 'Scale', and 'Psychometrics'. The terms within each subject were combined using the Boolean operator 'AND'. The keywords were searched anywhere within research papers (title, abstract, text). In addition, reference lists and citations of included articles were considered and further inclusions of studies were made. The search strategy included English language studies only.

Duplicates were removed and the remaining articles were screened using an adapted criteria from Moher, Liberati, Tetzlaff, and Altman (2009). After removal of duplicates, 1616 papers were rated against the inclusion and exclusion criteria. Following a screening and eligibility process 15 studies were included in analysis in this review.

Procedure

Each study was examined and psychometric properties were considered. The methodology of determining the reliability and validity of scales was then evaluated.

Data Analysis of the Methodological Quality

The methodological quality of the studies was collated and assessed through a quality assurance checklist. No consensus criteria exist for psychometric evaluation studies of rating scales, therefore the quality of the studies was determined using relevant items from the consensus-based standards for the selection of health status measurement instruments (COSMIN) checklist (Terwee, Mokkink, Knol, Ostelo, Bouter & de Vet, 2012).

Six items from the COSMIN checklist were used to evaluate the appropriate methodological quality of studies in relation to the psychometric analysis (see Appendix

A). These domains were: internal consistency; reliability; content validity; structural validity; hypothesis testing; and responsiveness (see Table 1).

Criteria for the Quality of Measurement Properties

Reliability. Psychometric properties relate to reliability and validity of the measures. Reliability is defined as the extent to which a tool performs consistently over repeated use and is an accurate measurement of the construct under investigation (Abell, Springer, Kanata, 2009). Kirk and Miller (1986) identified three types of reliability: the stability of a measure over time; the similarity of measurements within a given time period; and the consistency of measurements over repeated use. Within this review, studies were assessed for evidence of internal consistency and interrater reliability of scales.

Internal consistency is the degree of relatedness among items. Cronbach's alpha was considered an appropriate measure of internal consistency and scores above .7 were deemed acceptable (Terwee et al., 2007).

Intraclass correlation coefficients (ICC) and weighted Kappa were also considered acceptable measure of interrater reliability with scores about .70 considered adequate (Terwee et al., 2007).

Validity. Validity refers to the extent to which scores derived from a measure are interpretable and meaningful. Validity cannot be conclusively determined for an outcome measure, rather evidence is gathered in support of validity (Foster & Cone, 1995). This can be assessed by analysing the content of the measure, the construct, and the criterion validity. Content validity accounts for the degree to which the content of the scale is an adequate representation of the construct being measured (Mokkink et al., 2010). This was scored dependent on information provided regarding a process of evaluation in the development of the study, such as using the content validity measure.

Construct validity is divided into structural validity, hypothesis testing, and cross-cultural validity. Structural validity refers to the extent to which the scale ratings are an adequate reflection of construct being measured (Mokkink et al., 2010). This was demonstrated if studies included factor analysis whereby all factors explained greater than 50% total variance.

Hypothesis testing assessed whether studies provided a comparative analysis with a measure of a similar construct, and whether a clear hypothesis was stated as to the expected relationship and direction were stated. Pearson's correlation coefficients were considered an appropriate method of analysis with scores above .5 and showing significance deemed acceptable (Mokkink et al., 2010).

Cross-cultural validity was not assessed as none of the included studies provided information regarding translated or cultural adaptations for scales. Criterion validity was also not evaluated as no gold standard exists for therapist competency rating scales.

Responsiveness. Responsiveness refers to the ability of a scale to detect change over time in the construct being measured (Mokkink et al., 2010). Results over three time periods were assessed to determine if they were in accordance with a priori defined hypotheses, and calculated using either analysis of variance (ANOVA) or t-test to.

Table 1.

Description of COSMIN items and statistic methods of psychometric analysis.

COSMIN item	COSMIN definition	Statistical methods
Internal consistency	The degree of the interrelatedness among the items	Cronbach's alpha
Reliability	The proportion of the total variance in the measurements which is due to 'true' differences between patients	ICC
Content validity	The degree to which the content of scale is an adequate reflection of the construct to be measured	Appropriate analysis of scale items
Structural validity	The degree to which the scores of scale are an adequate reflection of the dimensionality of the construct to be measured	Exploratory or confirmatory factor analysis
Hypothesis testing	The degree to which the scores of scale are consistent with hypotheses based on the assumption that the scales validly measures the construct to be measured	Statistical comparison with other measure (or subscale)
Responsiveness	The ability of scale to detect change over time in the construct to be measured	Appropriate analysis of discriminant validity (ANOVA, t-test)

Terwee et al. (2012) developed a four-point rating scale per item (poor, fair, good and excellent). A total score, using the COSMIN checklist, was determined using a scoring system proposed by Cordier et al. (2015).

$$\text{Total score for psychometric quality} = \frac{\text{(Total score obtained - minimum score possible)}}{\text{(Max score possible - minimum score possible)}} \times 100$$

Using these criteria the results were presented as a percentage and were rated poor (0-25%), fair (26-50%), good (51-75%), or excellent (76-100%). To ensure consistency of COSMIN checklist ratings, all studies were scored by the first author and a sample (n=5) were randomly rated by an independent assessor. An intraclass correlation coefficient (ICC) was calculated to check reliability (ICC= .77) and was found to be within the *good* range (Koo & Li, 2015).

Results

The literature search identified 15 scales used to evaluate therapist competency. The descriptive information for each scale is presented and discussed below.

Overview of Measures

The Cognitive Therapy Adherence and Competence Scale (CTACS; Barber, Liese & Abrams, 2003). The Cognitive Therapy Adherence and Competence Scale (CTACS) was developed by reviewing items from cognitive therapy (CT) manuals, the Collaborative Study Psychotherapy Rating Scale (CSPRS), and Cognitive Therapy Scale (CTS) to assess therapists working with cocaine-dependent patients. The scale has 25-items in five sections*: cognitive therapy structure; development of a collaborative relationship; case conceptualisation; cognitive and behavioural techniques;* and *overall performance*. Items are rated on a 7-point Likert scale, one score for adherence and one for competence (only competence was evaluated for this study).

The Cognitive Therapy Scale- Revised (CTS-R; Blackburn et al., 2001; Reichelt, James & Blackburn, 2003). The Cognitive Therapy Scale- Revised (CTS-R) is an up-dated version of the Young and Beck's (1988) Cognitive Therapy Scale (CTS). It is a 14-item scale (rated on a 7-point Likert scale). Changes to the CTS include three additional items (*facilitation of emotional expression, charisma,* and *non-verbal behaviour*) and incorporation of three existing items on the CTS into one.

The Manual Assisted Cognitive Therapy Rating Scale (MACT; Davidson et al., 2004). The MACT Rating Scale includes 11-items used to evaluate therapist competency in applying techniques, interpersonal effectiveness, and adherence to the therapy model. The scale is used to assess competency in delivering manualised cognitive therapy specifically for patients who self –harm. Ratings are made on a 7-point Likert scale.

The Cognitive Therapy Scale (CTS; Dobson, Shaw & Vallis, 1985; Gordon, 2006; Vallis et al., 1986, Young & Beck, 1988). The Cognitive Therapy Scale (CTS) was developed to evaluate therapist competency in delivering CT for depression. It is an observer rating scale with 11 items (rated on a 7-point Likert scale) divided into two subscales. The general skill subscale includes items assessing: *agenda setting; obtaining feedback; therapist understanding; interpersonal skills; collaboration;* and *pacing of the session*. The specific skills subscale evaluates the therapist's ability to: *assess empiricism; focus on key cognition and behaviours; apply a change strategy; use appropriate cognitive-behavioural techniques;* and *assign homework.*

The Cognitive Therapy Scale for Psychosis (CTS-Psy; Gordon, 2006; Haddock et al., 2001). The CTS-Psy is a modified version of the CTS used specifically when treating patients with psychosis. It includes two subscales (general skills and technical skills) and has 13 items (rated on a 7-point Likert scale).

The Assessment of Core CBT Skills* (*ACCS; Muse et al., 2017). The Assessment of Core CBT Skills (ACCS) was developed to evaluate a therapist's core and CBT-specific competencies in delivering treatment for various conditions. The scale has 22 items organised into eight competency domains (rated on a 4-point scale): *agenda settings; formulation; CBT intervention; homework; effective communication; forming a therapeutic relationship; timing*; and *assessing change.*

The University College of London (UCL) scale for Structured Observation (Roth, 2016). This scale was developed as part of the IAPT programme and includes an evaluation of therapist competence in delivering CBT specific interventions (26 items) and core and generic therapist skills (13 items). Ratings are made on a 5-point Likert scale.

The Cognitive Therapy Competence Scale for Social Phobia (CTCS-SP; von Consbruch, Clark & Stangier, 2011). The scale was adapted from the CTS (Young &

Beck, 1988) to assess therapist's delivery of cognitive therapy specifically for social phobia. The Cognitive Therapy Competence Scale for Social Phobia (CTCS-SP) has 16 items (rated on a 7-point Likert scale). In addition to each item rating observers also provide an overall score of competency, and the degree of difficulty associated with working with the particular client.

Scales used to assess competency in other therapeutic models.

The Adherence/ competence scale for Individual Drug Counselling (ACS-IDCCD; Barber, Mercer, Krakauer & Calvo, 1996). The Adherence/ competence scale for Individual Drug Counselling (IDC) for cocaine dependence (ACS-IDCCD) is comprised of 43 items. Each item is rated on a 7-point Likert scale and is scored for frequency (adherence) and quality (competence). The competency ratings were used within this study. The scale has five subscales: monitoring drug use behaviour; encouraging abstinence; use of the 12-step model; relapse prevention; and providing education.

The competency in Cognitive Analytic Therapy scale (CCAT; Bennett & Parry, 2004). The competency in Cognitive Analytic Therapy scale (CCAT) measures the therapist competence when using cognitive analytic therapy. The CCAT competencies are based on three areas: *assessment and producing a formulation of client difficulties; establishing a therapeutic relationship;* and *developing, planning and evaluating therapeutic practice* (Bennett & Parry, 2004). There are 10 domains and 77 items which are rated using a 5-point Likert scale.

The Yale Adherence and Competence Scale (YACS; Carroll et al., 2010). The Yale Adherence and Competence Scale (YACS) was developed as a multi-model rating scale for the treatment of patients with drug use disorders. The scale was designed to assess treatment using either CBT, clinical management, or the twelve step facilitation. It has 55-items assessing general and model specific competence over six domains

(three general and three specific). Ratings are scored on a 5-point Likert scale for the quantity (adherence) and quality (competence).

The Mindfulness-Based Relapse Prevention Adherence and Competence scale (MBRP-AC; Chawla et al., 2010). The Mindfulness-Based Relapse Prevention Adherence and Competence scale (MBRP-AC) contains two sections each with two subscales. The first is the adherence section which provides an observer rating scale to assess therapist adherence to the model (this part of the scale will not be considered in this study). The second is a competency section that contains two subscales, one to evaluate the therapist style and approach within therapy, which assesses the therapist ability to provide timely, appropriate and empathetic response to patients. The second subscale is used to assess overall therapist performance and is designed to capture the rater's impression of the therapist's competence over the session. Each subscale has four items, each measured with on a 5-point Likert scale. The therapist is assessed on competency in delivering group treatment.

Mentalisation-Based Treatment Adherence and Competence Scale (MBT-ACS; Karterud et al., 2012). The 17-item Mentalisation-Based Treatment Adherence and Competence Scale (MBT-ACS) is used to rate therapist treating patients with borderline personality disorder (BPD). Each item requires a score from the rater for adherence to the treatment model, and a score for therapist competency (this was examined in this study). Scores are given on a 7-point Likert scale.

The Interpretive and Supportive Technique Scale* (*ISTS; Ogrodniczuk & Piper, 1999). The Interpretive and Supportive Technique Scale (ISTS) is used to assess therapist competence when using different forms of dynamically oriented psychotherapy. The scales consists of 14 items and assess the therapist's ability to be competent in a number of therapeutic techniques, such as *providing praise and to gratify the patient, make interpretations, engage in problem solving,* and *focus on the*

patient/therapist relationship. The scale has two subscales: Interpretive and Supportive, and each item is rated on a 5-point Likert scale.

The Behavioural Family Management Therapist Competency and Adherence Scale (BFM-TCAS; Weisman et al. 1998). The BFM-TCAS is used to evaluate the competency and adherence of a therapist delivering Behavioural Family Management (BFM) with patients with bipolar disorder. The scale has 13 items rated on a seven point Likert scale and also includes a measure of overall family difficulty and family expressed emotion status.

Results Summary

The search process highlighted a total of 15 scales, from seven different theoretical therapeutic intervention models. These included: eight from CBT (CTACS, CTS-R, CTS, CTS-Psy, ACCS, CTCS-SP, UCL scale, MACT); one from Individual Drug Counselling (ACS-IDCCD); the YACS could be used with either CBT, clinical management, or Twelve Step Facilitation (TSF); one from Cognitive Analytic Therapy (CCAT); one from behavioural family management (BFM-TCAS); two studies detailing third wave CBT approaches (MBRP-AC; MBT-ACS); and one from dynamic psychotherapy (ISTS). The review included 11 scales which were disorder specific: four scales specific for patients with drug dependency (ACS-IDCCD, CTACS, MBRP-AC, YACS), one for psychosis patients (CTS-Psy), one for borderline personality disorder (BPD) (MBT-ACS), one for social phobia (CTCS-SP), one for bipolar disorder (BFM-TCAS), one for patients who self-harm (MACT), and the CTS and CTS-R are specific for depression and anxiety. Four scales (ACCS, UCL scale, CCAT, CTS-R, ISTS) were transdiagnostic. Fourteen studies were identified that evaluated therapist competence in delivering one to one therapy, and one study involved rated therapist competence in delivering group treatment (MBRP-AC).

From the identified 15 scales, the results of the literature review showed 13 studies had been conducted to evaluate the psychometrically quality of twelve of the scales. No research evidence was found for the validity or reliability of the UCL scale, BFM-TCAS, or the MACT. Table two shows a summary of the 13 psychometric studies.

Table 2.

Descriptive properties of included psychometric studies.

Authors	Therapist Rating Scale	Therapy type	Patient condition	No. Items	Training/ Manual	cut-off	No of sessions rated (method)	No. of Raters	No. of therapists	Type of therapist	No. of patients
Barber, Liese & Abrams (2003)	CTACS	CBT	Drug use	21	-	-	129 (audio)	2	40	Qualified/ Trainees	129
Blackburn et al. (2001)	CTS-R	CBT	Depression and anxiety	13/14	Manual	-	102	4	20	Trainees	34
Gordon (2006)	CTS-R/ CTS- Psy	CBT	Various/ Psychosis	12/ 10	yes	yes	26 (audio)	9	26	Trainees	-
Haddock et al. (2001)	CTS- Psy	CBT	Psychosis	13	-	-	5 (reliability) 24 (validity)	4	21	Trainees	-
Muse et al. (2017)	ACCS	CBT	Various	22	Manual	-	76 (video)	76	76	Qualified/ Trainees	-
Vallis et al. (1986)	CTS	CBT	Depression	11	yes	-	10/53 (video)	5/7	9	Trainees	-

Authors	Therapist Rating Scale	Therapy type	Patient condition	No. Items	Training/ Manual	cut-off	No of sessions rated (method)	No. of Raters	No. of therapists	Type of therapist	No. of patients
von Consbruch, Clark & Stangier (2011)	CTCS-SP	CBT	Social phobia	16	Manual	yes	161	7	51	Trainees	98
Barber et al. (1996)	ACS-IDCCD	IDC	Drug use	43	-	-	41 (audio)	4	18	Qualified	40
Bennett & Parry (2004)	CCAT	CAT	Various	10	-	-	27 (audio)	3	12	Qualified	-
Carroll et al. (2000)	YACS	Various	Drug use	6	Manual	-	19 (reliability) 576 (validity) (video)	5	-	Qualified	576
Chawla et al. (2010)	MBRP-AC	MBRP	Drug use	8	Manual	-	44	5	10	Qualified	93
Karterud et al. (2012)	MBT-ACS	MBT	borderline personality disorder	17	Manual	yes	18	7	9	Qualified	18

Authors	Therapist Rating Scale	Therapy type	Patient condition	No. Items	Training/ Manual	cut-off	No of sessions rated (method)	No. of Raters	No. of therapists	Type of therapist	No. of patients
Ogrodniczuk & Piper (1999)	ISTS	Dyn	Various	14	Manual	yes	50 (audio)	2	18	Qualified	50

Note. blank sections given when no information provided in study paper. CBT = cognitive behavioural therapy, IDC = individual drug counselling, CAT = cognitive analytic therapy, MBRP = mindfulness based relapse prevention, MBT = mentalisation based treatment, Dyn = psychodynamic therapy.

Psychometric Appraisal of Competency Rating Scales

Details regarding the psychometric properties of included studies are summarised in Table 3. Eight studies reported the internal consistency of scales, all these were adequate ($a > .70$). All 13 studies provided evidence for the reliability of scales with scores for inter rater reliability, with 10 using Intraclass Correlation (ICC; Shrout & Fleiss, 1979), Bennett & Parry (2004) used Cohen's Kappa, and Haddock et al. (2001) using Pearson's correlation coefficients. Von Consbruch et al.'s (2011) study was the one that provided results for test re-test reliability. All but three presented a test for validity, these were either an analysis of convergent validity (comparing scale with another measure of similar construct) or responsiveness to change over time.

Table 3.

Psychometric properties of included studies of competency rating scales.

		Reliability			Validity	
Author (year)	Therapist rating scale	Internal consistency	Interrater reliability	Test re-test	Convergent	Responsiveness
Barber, Liese & Abrams (2003)	CTACS	*a*= .93	ICC= .73	-	*r*=.97 (competence and adherence)	-
Blackburn et al. (2001)	CTS-R	>.70	ICC= .63 (13 item) ICC= .57 (14 item)	-	-	*t*= 4.43** (improved over course)
Gordon (2006)	CTS-R/ CTS-Psy	-	ICC= .38 (CTS-R) ICC= .28 (CTS-Psy) ICC= .76 (CTS-R) ICC= .28 (CTS-Psy) (after training)	-	*r*=.79 ** (CTS-R and CTS-Psy)	

		Reliability			Validity	
Author (year)	Therapist rating scale	Internal consistency	Interrater reliability	Test re-test	Convergent	Responsiveness
Haddock et al. (2001)	CTS- Psy	-	r=.94 (overall score) *r*= .95 (general subscale) *r*= .80 (technical subscale)	-	-	*F*= 10.5 ** (improved over course)
Muse et al. (2017)	ACCS	*a*= .90/.94 (two study groups)	ICC= .74/.73 (two study groups)	-	*r*= .65** (CTS-R)	*F*= 5.50 ** (improved over course)
Vallis et al. (1986)	CTS	-	ICC = .59/ .74/ .84 (number of raters)	-	*r*= .85** (subscales)	-
von Consruch, Clark & Stangier (2011)	CTCS-SP	*a*=. 82- .92 (dependent on raters)	ICC= .73-.88 (pairs of raters)	*r*= .92 ICC= .55- .96	-	-

		Reliability			Validity	
Author (year)	Therapist rating scale	Internal consistency	Interrater reliability	Test re-test	Convergent	Responsiveness
Barber et al. (1996)	ACS-IDCCD	*a*= .83- .95 (items)	ICC= .65-.89 (items)	-	-	-
Bennett & Parry (2004)	CCAT	a=.98	*K*=.67/.64/.63 (Each pair)	-	*r*=.74 ** (TIC) *r*=.72 ** (WAI-O)	-
Carroll et al. (2000)	YACS	-	ICC= .71- .97 (items)		*r*= .12 -.54 * (intercorrelation) Various (WAI, VTAS, Penn, CALPAS) *r*= .21**- .62** (competence and adherence)	
Chawla et al. (2010)	MBRP- AC	*a*= .86/ .82 (subscales)	ICC= .53 - .76	-	no correlation (WAI)	-

		Reliability			Validity	
Author (year)	Therapist rating scale	Internal consistency	Interrater reliability	Test re-test	Convergent	Responsiveness
Karterud et al. (2012)	MBT-ACS	-	ICC= .88 ICC= .68 (number of raters)	-	-	-
Ogrodniczuk & Piper (1999)	ISTS	*a*= .92/ .95	ICC= .95/ .95 (two studies)	-	*r*= .73 ** (TIRS) *r*= .70 ** (PTS)	-

Note. *= *p*>.05 **= *p* >.01 (CTACS and CTCS-SP significance was not reported).

Cognitive Therapy Scales

CTACS. Two expert cognitive therapists rated a total of 129 audio recorded cognitive therapy, supportive-expressive dynamic therapy or individual counselling sessions with cocaine-dependent patients. The inter-rater reliability of CTACS was determined by calculating the Intraclass Correlation Coefficient (ICC; Shrout & Fleiss, 1979) and showed varied results for competency items (ICC= .22 to .94, average ICC= .73). The CTACS had good internal consistency (a= .93) and positive correlation between the adherence and competency subclass (r=.97). Criterion validity was determined by comparing CT scores with supportive expressive dynamic therapy and counselling scores. The results showed significant differences. The CTACS showed acceptable levels of interrater reliability and criterion validity.

CTS-R. Four expert raters assessed 102 tapes from three different stages of therapy from 20 mental health professionals undergoing cognitive therapy training. Sessions were with patients with either anxiety or depression. The results of the analysis of reliability for CTS-R total scores showed adequate moderate inter-rater reliability (13 items ICC= .63/ 14 items ICC= .57). Inter-rater reliability for individual items showed variability (ICC = -.14 to .84). Discriminant validity and scale responsiveness of the CTS-R was determined by evaluating whether trainee competency improved, as expected, over the course of training. Paired *t*-test results showed significant improvement (t= 4.43, *df* 10, p <.001). The results did not show the CTS-R to have adequate reliability but did show scale responsiveness.

CTS-R and CTS-Psy. The study by Gordon (2006) compared the psychometric qualities of the CTS-R and the CTS-Psy. Data was collected from 26 audiotaped sessions rated by two independent assessors using both scales to measure therapist competence. The results showed *poor* inter-rater reliability for both measures (ICC= .38

for the CTS-R/ ICC= .28 for the CTS-Psy). There was an increase in the rater agreement for the CTS-R (ICC= .76) after raters had attended recent specific training, but no increase for the CTS-Psy. There was strong inter-scale agreement between both scales (r= .79, p<..00). Neither the CTS-R nor the CTS-Psy showed good interrater reliability.

CTS-Psy. The reliability of the CTS-Psy was determined by analysing the inter-rater reliability using correlation coefficient of five rated therapy sessions assessed by four expert raters scores. The results showed high inter-rater reliability for the overall scores (r= .94) and the total subscale scores (general r= .95/ technical r= .80). The correlation between raters for individual items showed mostly good inter-rater reliability. The discriminant validity of the CTS-Psy was determined by comparing therapists (n=24) scores who had received psychosis training with those who had not (n=17). Sessions were rated by four expert raters using the CTS-Psy. The results showed highly significant differences in means scores between groups (F(1,21) =10.5, p= .004). The results showed that CTS-Psy showed excellent interrater reliability and good validity.

ACCS. The evaluation recruited therapists from a university CBT training course and an IAPT service. A total of 76 sessions were assessor rated using ACCS and CTS-R, 20 of which were double marked. The results of the psychometric evaluation of the ACCS showed excellent internal consistency (.90 /.94 for two study groups) and good inter-rater reliability for overall total scores (ICC= .74 /.73) The ICC scores showed variability in agreement for individual items (ICC= .27- .83). The results to determine the discriminant validity showed that trainee participants (study one) significantly increased their ACC scores over time during the training course (F(3, 48) = 5.50, p< .01). An analysis of the comparative validity showed a strong positive relationship between the ACCS and the CTS-R (r= .65, p>.00). Comparisons between

the ACCS and the CTS-R showed strong positive correlation (r=.65, p<.01). Overall, the study showed that ACCS is a valid and reliable measure of CBT competence.

CTS. The intraclass correlations were calculated using data collected from 10 videotaped sessions and rated by five experts and showed moderate reliability (ICC= .59) for one rater. An analysis of the ratings of individual item was within poor to moderate range (ICC= .27 - .59). Examining the results of the ICC for two raters the inter-rater reliability increased to show a good correlation (ICC= .77). Fifty three tapes were rated on acceptability and means between acceptable and unacceptable competency ratings were compared and showed significant difference (F= 7.90, p<.00). The correlation between the two subscales of the CTS was high (r= .85, p<.00). The CTS showed poor interrater reliability but more acceptable when rater numbers increased.

CTCS-SP. Ratings from 161 video recorded sessions were collected from qualified therapist involved in a multi-centre trial. Sessions were doubled marked by two of seven raters. The results of the statistical analysis of the psychometric qualities of the CTCS-SP showed good internal consistency (a= .82- .92) and high inter-rater reliability for the total score (ICC= .73- .88). For individual items the inter-rater reliability ranged from low to high (ICC= -.06 to .98). The test re-test reliability was determined by comparing the scores of 15 sessions with ratings made on the same sessions after an 18-24 month period. The results showed substantial correlation (r= .92) between two sessions on therapist training course. The results showed acceptable reliability and validity for the CTCS-SP.

Other therapeutic models scales.

ACS-IDCCD. Three independent raters assessed 41 audiotaped sessions of individual drug counselling (IDC), 11 of cognitive therapy (CT), and 10 of supportive expressive therapy (SE) with patients with cocaine dependency. The results of the

analysis of the psychometric qualities of the ACS-IDCCD showed good internal consistency of each item for the competency ratings (*a*= .83- .95) and moderate to good inter-rater reliability (ICC= .65- .89) between 3 raters for CT, SE and IDC therapists. The ACS-IDCCD showed good interrater reliability, but validity was not evaluated.

CCAT. The psychometric qualities of the CCAT, a therapist rating scale for cognitive analytic therapy (CAT), were evaluated. Three rater pairs scored a total of 27 sessions across NHS and university counselling services. The results showed good internal consistency (*a*= .96 for early sessions and *a*= .98 for later sessions). The inter-rater agreement was calculated using Cohen's Kappa (Fleiss, 1971) and showed good reliability (*K*= .67, .64 and .63 for three rater pairs). The CCAT showed highly significant correlation with the TIC-O ($r = .59$, $p < .001$) and WAI ($r = .61$, $p < .001$). The results showed excellent interrater reliability and good validity for the CCAT.

YACS. The interrater reliability for the YACS was determined from 19 randomly selected tapes from a clinical trial assessing IDC, CT, and SE with cocaine-dependent drug users. Assessments were made five raters. The results showed that total scale scores were within the moderate to excellent range (ICC= .71- .97) and within poor to good range for individual items (ICC= .06- .81). An intercorrelation between competency dimensions showed significant positive results (*r*= .12- .54). The scale was assessed for validity by comparing a total of 576 session YACS ratings with scores from measures of similar construct. Four comparative measures were used: The Working Alliance Inventory (WAI; Horvart & Greenberg, 1986); the California Psychotherapy Alliance Scale (CALPAS; Marmar et al., 1986); the Vanderbilt Therapeutic Alliance Scale (VTAS; Hartley & Strupp, 1983); and the Penn helping alliance rating scale (Penn; Luborsky et al. 1983). The results showed variable results of Pearson correlation coefficients (ranging from -.34 to .57). The relationship between adherence and competence ratings showed significant positive correlations (*r*= .21- .62,

p=.001). Overall, YACS showed excellent reliability and good comparative and discriminant validity.

MBRP-AC. Five expert raters assessed 44 randomly selected audio recorded group sessions of MBRP for patients who drug use. The reliability and validity of the measure's competency subscale was analysed by determining ICC and by evaluating the relationship between MBRP-AC ratings with the results of the Working Alliance Inventory (WAI-S; Horvath & Greenberg, 1989; Tracey & Kokotovic, 1989). For the subscale two components the results showed good internal consistency for the Therapist (*a*= .86) and the Overall Therapist Performance (*a*=.82). The analysis of the inter-rater reliability showed high levels of agreement for the total summary scores for competency. The individual items scored within the good and excellent range (ICC= .53- .76). The correlation between the MBRP-AC (competency subscale) and the WAI did not show any relationship for either component. The MBRP-AC showed good reliability but was unable to show comparative validity.

MBT-ACS. The results of the analysis of the psychometric qualities of the MBT-ACS showed good correlation between seven raters assessed 18 therapy sessions (ICC= .88), however, this declined when rater numbers reduced (ICC= .68). The item correlations were variable (ICC= .49-.90). The scale showed to be a reliable measure of MBT, validity was assessed.

ISTS. The results of the psychometric analysis of the ISTS were split into two studies. The first included scores from 50 audio recorded interpretive and support therapy sessions rated by two expert assessors. The results of study one showed high inter-rater correlation between two raters for total scores (ICC= .95) and for each subscale (ICC= .93 for supportive subscale and ICC= .88 for interpretive subscale). ICC correlations for individual items were within moderate to good range (average ICC= .74), with the exception of one item (ICC= .35). In Study two, the inter-rater reliability

between two different raters (assessing 50 sessions) showed similar results for the full scale (ICC= .95) and the interpretive subscale (ICC=.84), but was lower for the supportive subscale (ICC=.69). Individual items were in the moderate to high range (average ICC=. 54) with the lowest item being 'personal information' (ICC=.28). The ISTS was reported to have high internal consistency for the full scale (*a*= .92/ .95 for each rater), for the supportive subscale (*a*= .92/ .94), and the interpretive subscale (*a*=.86/ .88). The results of the analysis of convergent validity showed that the ISTS highly correlated with two other measures of psychodynamic techniques, the Therapist Intervention Rating System (TIRS; Piper et al., 1987) (*r*=.73, *p* <.00) and the Perception of Technique Scale (PTS; Piper et al., 1993) (*r*=70, *p*<.00). The results show the ISTS to be a valid and reliable measure.

Psychometric Properties and Methodological Quality

Details regarding the methodological quality are presented in table 4. Studies' percentage scores for each criterion are provided and show variability in study quality. All included studies provided an analysis of interrater reliability for scales, yet studies were inconsistent in the extent to which validity was evaluated. The results show that none of the studies provided evidence for every methodological quality domain on the COSMIN checklist.

Table 4.

Item and total percentages for the COSMIN checklist for good methodological quality.

Rating Scales	**Internal consistency**	**Reliability**	**Content validity**	**structural validity**	**Hypothesis testing**	**Responsiveness**
CTACS	22 (poor)	71 (good)	71 (good)	-	36 (fair)	-
CTS-R	56 (good)	76 (excellent)	86 (excellent)	-	-	52 (good)
CTS-R /CTS-psy	-	52 (good)	86 (excellent)	-	40 (fair)	-
CTS-PSY	-	48 (fair)	86 (excellent)	-	-	53 (good)
ACCS	56 (good)	71 (good)	86 (excellent)	-	56 (good)	71 (good)
CTS	-	57 (good)	86 (excellent)	42 (fair)	28 (fair)	-
CTCS-SP	22 (poor)	67 (good)	86 (excellent)	-	-	-

Rating Scales	Internal consistency	Reliability	Content validity	structural validity	Hypothesis testing	Responsiveness
ACS-IDCCD	11 (poor)	62 (good)	57 (good)	-	-	-
CCAT	44 (fair)	62 (good)	62 (good)	-	56 (good)	-
YACS	-	67 (good)	79 (excellent)	58 (good)	44 (fair)	-
MBRP-AC	39 (fair)	67 (good)	86 (excellent)	-	44 (fair)	-
MBT-ACS	-	48 (fair)	-	-	-	-
ISTS	83 (excellent)	86 (excellent)	71 (good)	75 (good)	72 (good)	-

Eight of the 13 studies provided results of internal consistency analysis, all were within acceptable range. All included studies analysed the interrater reliability of scales, thought the results showed only six scales were within consistently within acceptable range (ICC >.70) (CTACS; CTS-Psy; ACCS; CTCS-SP; YACS; ISTS).

All studies assessed content validity, except Karterud et al.'s (2012; MBT-ACS) study which provided no information regarding scale development. Only three studies provided information regarding scale structural validity and included factor analysis (Vallis et al., 1986 ; Carroll et al., 2000 ; Ogrodniczuk & Piper, 1999). Scale responsiveness was evaluated in eight studies. Two studies compared measure subscales (Barber et al., 2003; Vallis et al., 1986) and five compared scales with measures of similar construct, either CTS-R or of therapeutic alliance (Gordon, 2006 ; Muse et al., 2017; Bennett & Parry, 2004 ; Chawla et al., 2010; Ogrodniczuk & Piper, 1999). Carroll et al.'s (2000) study compared YACS with subscales and therapeutic alliance measures. Scores were generally acceptable, except for MBRP-AC (Chawla et al., 2010) which showed no correlation with WAI. The quality of convergent validity analyses for studies was good to fair, as studies did not provide clear hypothesis of expected outcomes of results.

Responsiveness to change over time was evaluated in only three studies (Blackburn et al., 2001; Haddock et al., 2001; Muse et al., 2017). The results showed that all scales showed responsiveness to change as trainee therapists progressed through a training course.

Discussion

This review systematically appraised and critiqued psychometric studies of rating scales which assess therapist competency in delivering psychotherapy to adults. Fifteen scales were identified, with thirteen papers provided evidence of psychometric quality of a scale. Three scales did not have any related research on their reliability or validity (UCL scale; BFM-TCAS; MACT).

The results of the psychometric studies showed that eight scales showed good reliability and validity, two showed only good reliability (ACS-IDCCD; MBRP-AC), and the CTS-Psy showed conflicting results across two studies. The CTS and the CTS-R showed the weakest psychometric results. All included a methodologically robust evaluation of interrater reliability. However the review demonstrated variability in the inclusion and quality of tests for scale validity. None of the studies were consistent in their method of assessment or analysis of reliability and validity.

Three scales did not include any evaluation of psychometric properties (UCL scale, BFM-TCAS, MACT) highlighting that some scales have been developed without evidence as to whether they are reliable measure sof therapist competency or can appropriately evaluate the competency construct. The results showed a paucity of therapist competency scales available (15 in total) and that scale development should include an evaluation of psychometric quality. The variety of outcomes from the 13 studies showed a range of evidence, which highlighted differences in reliability and validity. For the three studies without psychometric evidence the scale quality cannot be determined.

Reliability

The results showed that only eight of the 13 studies provided evidence of internal consistency using Cronbach's alpha. All studies included an analysis of interrater reliability, though results varied and only six studies provided adequate

agreement between raters. The methods of data collection for interrater reliability differed considerably between studies, with some utilising scores from two raters who observed large numbers of therapist sessions (Barber, Liese & Adams, 2003) and other studies collecting data from equal numbers of raters and therapists (Muse et al., 2017). Karterud et al. (2012) note the disparity between analyses of reliability for competency rating scales, and go on to state that some studies may violate the random requirement needed for ICC statistical analysis, potentially making results and conclusions invalid. Differences in methods of determining reliability make comparisons and interpretations of results between studies challenging to assess as methods differ significantly.

Studies provided information regarding interrater agreement of individual items within competency scales. The results showed disparities between item ICC scores, demonstrating that there were higher levels of agreement between some competence items than others, suggesting, therefore, discrepancies in how raters perceive different aspects of competence. Each study provided various levels of training and information regarding rating scales. Barber et al. (2007) state there have been persistent issues regarding the extent of training needed for raters to achieve quality scoring and good interrater reliability on competency scales.

The review results showed differences in the number of items included in competence scales, demonstrating discrepancies in how competency characteristics were defined in scales. The YAC (Carroll et al., 2000) has only six items, whereas the ACS-IDCCD (Barber et al., 1996) has 43. The scales used a range of definitions and assessment criteria to determine therapist competency which differed across theoretical approaches and patient diagnosis. This highlights that there is currently no standard definition of therapist competence. However, setting a generic, transdiagnostic criterion for therapist competency across theoretical models is unlikely to be feasible or applicable for the use in clinical practice (Piper & Ogrodniczuk, 1999).

Validity

Convergent validity was either determined through correlation analyses between competence and adherence subscales, between other competency rating scales, or with measures of therapeutic alliance. Gordon (2006) highlights the risk of using scales of poor psychometric quality as comparative measures. Ratings on the ACCS were compared with ratings on CTS-R (Muse et al., 2017), yet the results of the psychometric evaluation of the CTS-R (Blackburn et al., 2001; Gordon, 2006) show only poor to moderate interrater reliability therefore, as it is a questionable comparable measure of validity.

Responsiveness

Three studies evaluated validity by determining responsiveness of scales and therefore their ability to detect change over time (Blackburn et al.,2001; Haddock et al., 2001; Muse et al., 2017). The evaluation studies of the CTS-R, CTS-Psy, and ACCS collected data over different time periods to determine whether trainees improved on a CT training course. The results showed significant differences in ratings, concluding that scales showed an increase in scores during course progression. However, von Consbruch et al.'s (2011) study also measured the relationship between ratings at two time periods of trainees during a CT course. Yet in their study this was described as test re-test reliability and showed a significant correlation (rather than difference) between rating scores, showing ratings were similar during course duration. These results highlight differences in definitions and methods of analysis of validity, and discrepancies in interpretation of results to provide supporting evidence for psychometric quality. A further limitation in using retest test reliability to determine scales responsiveness to change was that the results may have shown the scale to be reliable (shows an expected difference) yet could not evaluate whether it is correctly measuring the appropriate construct. Hays and Hadom (1992) state that responsiveness

to change can only be considered a validity measurement when various methods of scale validity are used to determine whether the scale is measuring the identified construct. The psychometric studies for the CTS-R (Blackburn et al., 2001) and the CTS-Psy (Haddock et al., 2001) used only responsiveness to change to determine the validity, therefore as there are no other measures of validity, the results were inconclusive as to whether the scales accurately evaluated the therapist competency construct.

Interpretability

Interpretability of measures is considered an important characteristic of psychometric evaluation (Mokkink et al., 2010). Only four studies provided a cut-off score for scales which determined a level of adequate competence for therapists (Gordon, 2006; von Consbruch et al., 2011; Karterud et al., 2012; Ogrodniczuk & Piper, 1999). For the remaining nine scales it would be difficult to determine any qualitative meaning regarding competency from the quantitative ratings or change in ratings on scales.

Five studies collected data from trainee therapists and six with qualified therapists. The validity of these scales was limited by the evaluation context (a training course or one service), potential rater bias (trainer on the course or supervisor in service), and provide only the psychometric quality of scales within one context (Haddock et al., 2001). Two studies incorporated both trainee and qualified therapists (Barber et al., 2003; Muse et al., 2017) and were able to demonstrate the applicability of scales in both training and clinical practice.

Kazantzis (2003) state that therapist competency measures for CBT practice currently lead in comparison to other therapeutic approaches. This was evident in the review results, with seven of the 13 studies applicable to CBT. In terms of diagnosis there were more scales for drug use than any other mental health condition. All studies

related to one to one therapy, expect one (Chawla et al., 2010) which assessed therapist competence in running a treatment group. The review highlighted the paucity of therapist competency measuring delivery of therapy for different theoretical approaches, mental health conditions, and group treatment.

With the exception of the CTS (Vallis et al., 1986) and the CTS-Psy (Gordon, 2006; Haddock et al., 2001) all studies within the review were developed and psychometrically evaluated by the same authors. This introduces potential bias in the interpretation of results, and highlights the need for further evaluation and research into existing therapist competency scales.

Limitations of Review

There are several limitations of this review. The lack of clear definition of therapist competence (Wampold, 2015) meant that selecting studies for the inclusion of this review was challenging. Exclusions were made if studies did not explicitly state that the scale was measuring therapist 'competence'. Studies with scales that rated specific therapist qualities, such as empathy, were not included when it could be argued that these attributes are part of the presentation of a competent therapist. The literature on the definition of competence is broad and is open to interpretation. It is also likely to differ with alternative psychotherapeutic models.

Some studies were excluded from analysis if they did not distinguish between adherence and competency. Carroll et al. (2010) argue that treatment adherence and therapist competency are intrinsically linked. Furthermore, some included scales may be both constructs (such as the CTS-R).

None of the scale authors were contacted during the process of data collection for this literature review to determine whether psychometric evaluation studies had been conducted or were due to be published. This could have yielded further results for the

three scales (UCL scales, BFM-TCAS, MACT) that did not have reliability or validity evidence, or provided further psychometric evidence for the other included scales.

A further limitation was that the review utilised the COSMIN checklist to determine psychometric methodological quality. Use of this tool as an interpretation of the methodological quality is likely to be subject to assessor bias. Without a 'gold standard' method it was unclear how validity and reliability should be defined, assessed, and interpreted and therefore scoring was subjective.

Conclusion

The aim of this systematic review was to critically appraise and evaluate the psychometric properties and methodological quality of rating scales used to assess therapist competency in delivering psychotherapy to adults with mental health conditions (regardless of theoretic approach). The results showed that eight of the 13 studies assessed provided evidence to suggest scales with good reliability and validity. However, there were discrepancies in the methodological quality of included studies, presenting a lack of consistency in how psychometric properties were assessed.

Future Research

Clear areas of focus for future research have emerged from this review.

Ensuring therapist competence in delivering psychotherapy is crucial in providing quality, safe care for patients. The review highlighted paucity in available competency assessment scales. Therefore, further development and research is needed to provide competency measures for a range of psychotherapeutic approaches and mental health conditions, so that therapist competency is assured in training and clinical practice.

Developed competency rating scales must undergo clearly defined, rigorous psychometric evaluation to determine the reliability as well as validity of measures.

Psychometric evaluations should include more than one method of analysis of reliability and validity. Developed scales would benefit from further evaluation.

Clinical Implications

This review provides an overview of current literature on therapist competency rating scales, and an appraisal of scale psychometric properties and methodology for each study. Scales have been developed for the use in training and clinical practice. Therefore, this review may be helpful for trainers and clinicians in selecting appropriate rating scales for the use in practice.

This review highlights the lack of therapist competency scales of good methodological quality, as well as a lack of diversity in the number of scales available. Therefore promoting the development of new scales to assess therapist competency in psychotherapy.

References

Abell, N., Springer, D. W., & Kamata, A. (2009). *Reliability in developing and validating rapid assessment instruments.* Oxford, UK: Oxford Scholarship Online.

Ackerman, S. J., & Hilsenroth, M. J. (2003) A review of therapist characteristics and techniques positively impacting the therapeutic alliance. *Clinical Psychology Review, 23*(1), 1-33. Doi: 10.1016/S0272-7358(02)00146-0.

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

Barber, J.P., Liese, B.S., & Abrams, M.J. (2003) Development of the Cognitive Therapy Adherence and Competence Scale. *Psychotherapy Research, 13,* 205-221. Doi:10.1093/ptr/kpg019

Barber, J. P., Sharpless, B. A., Klostermann, S., & McCarthy, K. S. (2007). Assessing intervention competence and its relation to therapy outcome: A selected review derived from the outcome literature. *Professional Psychology: Research and Practice, 38,* 493-500. Doi: 10.1037/0735-7028.38.5.493

Bennett, D. & Parry, G. (2004) A measure of psychotherapeutic competence derived from cognitive analytic therapy, *Psychotherapy Research, 14,* 176-192. Doi: 10.1093/ptr/kph016

Bennett, D., Parry, G. and Ryle, A. (1999). *Development of a measure of therapist competence in resolving transference enactments which threaten the therapeutic alliance*. Unpublished report, Mental Health Foundation.

Bjaastad, J. F., Haugland, B. S. M., Fjermestad, K. W., Torsheim, T., Havik, O. E., Heiervang, E. R., & Öst, L.-G. (2016). Competence and Adherence Scale for Cognitive Behavioral Therapy (CAS-CBT) for anxiety disorders in youth:

Psychometric properties. *Psychological Assessment, 28,* 908-916. Doi: 10.1037/pas0000230.

Blackburn, I.M, James, I.A., Milne, D.L., Baker, C., Standart, S., Garland, A., & Reichelt, F. K. (2001) The revised cognitive therapy scale (CT-R): psychometric properties. *Behavioural and Cognitive Psychotherapy, 29,* 431-447. Doi: 10.1017/S1352465801004040.

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

Brosan, L., Reynolds, S., & Moore, R. G. (2008). Self-evaluation of cognitive therapy performance: Do therapists know how competent they are? *Behavioural Cognitive Psychotherapy, 36,* 581-587. Doi: 10.1017/S1352465808004438

Carroll, K, M., Nich, C., Sifry, R. L., Nuro, K. F., Frankforter, T. L., Ball, S. A., Fenton, L., & 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. Doi: 10.1016/S0376-8716(99)00049-6.

Chawla, N., Collins, S., Bowen, S., Hsu, S., Grow, J., Douglas, A., & Marlatt, G. A. (2010). The Mindfulness-Based Relapse Prevention Adherence and Competence Scale: Development, Interrater Reliability and Validity. *Psychotherapy Research, 20,* 388–397. Doi: 10.1080/10503300903544257

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, 6,* 284-290. Doi:10.1037//1040-3590.6.4.284

Clarke, V., & Braun, V. (2014). Thematic Analysis. *Encyclopedia of Critical Psychology,* 1947-1952. Doi:10.1007/978-1-4614-5583-7_311

Cooper, Z., Doll, H., Bailey-Straebler, S., Bohn, K., de Vries, D., Murphy, R., O'Connor, M. E., & Fairburn, C. G. (2017) Assessing Therapist Competence: Development of a Performance-based measure and its comparison with a web-based measure. *JMIR, 4,* 51. Doi: 10.1296/mental.7704.

Cordier R, Speyer R, Chen Y-W, Wilkes-Gillan S, Brown T, Bourke-Taylor H, Doma, K., & Leicht, A. (2015) Evaluating the Psychometric Quality of Social Skills Measures: A Systematic Review. *PLoS One, 10*(7), 1-32. Doi: 10.1371/journal.pone.0132299

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16,* 297-334.

Davidson, K., Scott, J., Schmidt, U., Tata, P., Thornton, S., & Tyrer, P. (2004). Therapist competence and clinical outcome in the Prevention of Parasuicide by Manual Assisted Cognitive Behaviour Therapy Trial: The POPMACT study. *Psychological Medicine, 34,* 855-863. Doi:10.1017/S0033291703001855

Fairburn, C. G. & Cooper, Z. (2011) Therapist competence, therapy quality, and therapist training. *Behaviour research and therapy, 49,* 373-378. Doi: 10.1016/j.brat.2011.03.005

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin, 76,* 378-382. Doi: 10.1037/h0031619

Foster, S. L. & Cone, J. D. (1995). Validity issues in clinical assessment. *Psychological Assessment,7,* 248- 260. Doi: 10.1037/1040-3590.7.3.248

Ginzburg, D. M., Bohn, C., Hofling, 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,* 747–752. Doi: 10.1016/j.brat.2012.09.001

Glasziou, P., Irwig, L., Bain, C., & Colditz, G. (2001). *Systematic Reviews in Health*

Care. 1st edition.. Cambridge, UK: Cambridge University Press

Gordon, P. K. (2006). A comparison of two versions of the Cognitive Therapy Scale. *Behavioural and Cognitive Psychotherapy 35*, 343. Doi: 10.1037/pas0000372

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.

Hays, R. D., & Hadorn, D. (1992). Responsiveness to change: an aspect of validity, not a separate dimension. *Quality of Life Research,* 1, 73-75. Doi:10.1007/BF00435438.

Hogue, A., Henderson, C. E., Dauber, S., Barajas, P. C., Fried, A., & Liddle, H. A. (2008). Treatment adherence, competence, and outcome in individual and family therapy for adolescent behavior problems. *Journal of Consulting and Clinical Psychology, 76*, 544-555. Doi: 10.1037/0022-006X.76.4.544

Horvath, A. O., & Greenberg, L. S. (1989). Development and validation of the Working Alliance Inventory. *Journal of Counseling Psychology, 36*, 223-233. Doi: 10.1037/0022-0167.36.2.223

Horvath, A. O., & Symonds, B. D. (1991). Relation between working alliance and outcome in psychotherapy: A meta-analysis. *Journal of Counseling Psychology, 38*, 139-149. Doi: 10.1037/0022-0167.38.2.139

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

Karterud, S., Pedersen, G., Engen, M., Johansen, M. S., Johansson, P. N., Schluter, C., & Bateman, A. W. (2013) The MBT Adherence and Competence Scale (MBT-ACS): Development, structure and reliability. *Psychotherapy Research: Journal*

of the Society for Psychotherapy Research, 23, 705–717. Doi: 10.1080/10503307.2012.708795

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,* S27-S45. Doi: 10.1037/a0015833

Kazantzis, N. (2003). Therapist competence in cognitive-behavioural Therapies: Review of the contemporary empirical evidence. *Behaviour Change, 20,* 1-12. Doi:10.1375/bech.20.1.1.24845

Keen, A, J., & Freeston, M, H. (2008). Assessing competence in cognitive-behavioural therapy. *British Journal of Psychiatry,* 193, 60–64. Doi: 10.1192/bjp.bp.107.038588

Keijsers, G., Schaap, C., & Hoogduin, C. (2000). The impact of interpersonal patient and therapist behavior on outcome in cognitive-behavior therapy. *Behavior Modification, 24,* 264-297. Doi:10.1177/0145445500242006

Kirk, J., & Miller, M. L. (1986). *Reliability and validity in qualitative research*. Beverly Hills, US:Sage Publications.

Kohrt, B. A., Jordans, M. J., Rai, S., Shrestha, P., Luitel, N. P., Ramaiya, M. K., . . . Patel, V. (2015). Therapist competence in global mental health: Development of the ENhancing Assessment of Common Therapeutic factors (ENACT) rating scale. *Behaviour Research and Therapy, 69,* 11-21. Doi:10.1016/j.brat.2015.03.009

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine, 15,* 155–163. Doi: 10.1016/j.jcm.2016.02.012

Lambert, M. J., & Barley, D. E. (2001). Research summary on the therapeutic

relationship and psychotherapy outcome. *Psychotherapy: Theory, Research, Practice, Training, 38*, 357-361. Doi: 10.1037/0033-3204.38.4.357

Liberati, A. (2009). The PRISMA statement for reporting systematic reviews and meta-analyses of studies That evaluate health care interventions: Explanation and elaboration. *Annals of Internal Medicine,* 151. Doi:10.7326/0003-4819-151-4-200908180-00136

Martin, D. J., Garske, J. P., & Davis, M. K. (2000). Relation of the therapeutic alliance with outcome and other variables: A meta-analytic review. *Journal of Consulting and Clinical Psychology, 68*, 438-450. Doi: 10.1037/0022-006X.68.3.438

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D.G, (2009). The PRISMA Group (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. *PLoS Med 6: e1000097*. Doi:10.1371/journal.pmed1000097

Mokkink, L. B., Terwee, C, B., Patrick, D. L., Alonso, J., Stratford, P.W., Knol, D. L.… & de Vet, H. C. W. (2010). 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.

Mokkink, L. B., Terwee, C, B., Patrick, D. L., Alonso, J., Stratford, P.W., Knol, D. L.… & de Vet, H. C. W. (2010). International consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes: results of the COSMIN study. *Journal of Clinical Epidemiology, 63*,737-745.

Mcleod, B. D., Southam-Gerow, M. A., Rodríguez, A., Quinoy, A. M., Arnold, C. C.,

Kendall, P. C., & Weisz, J. R. (2016). Development and Initial Psychometrics for a Therapist Competence Instrument for CBT for Youth Anxiety. *Journal of Clinical Child & Adolescent Psychology,* 1-14. Doi:10.1080/15374416.2016.1253018

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

Muse, K., Mcmanus, F., Rakovshik, S., & Thwaites, R. (2017). Development and psychometric evaluation of the Assessment of Core CBT Skills (ACCS): An observation-based tool for assessing cognitive behavioral therapy competence. *Psychological Assessment, 29*, 542-555. Doi:10.1037/pas0000372

Norman, G. (1985). Defining competence: A methodological review. In V.Neufeld & G.Norman (Eds.). *Assessing clinical competence*. New York NY: Springer.

Ogrodniczuk, J. S., & Piper, W. E. (1999). Measuring Therapist Technique in Psychodynamic Psychotherapies: Development and Use of a New Scale. *The Journal of Psychotherapy Practice and Research, 8*, 142–154.

O'Malley, S.S., Foley, S, H., Rounsaville, B. J., Watkins, J. T., Sotsky, S. M., Imber, S. D., & Elkin, I. (1988). Therapist competence and patient outcome in interpersonal psychotherapy of depression. *Journal of Consulting and Clinical Psychology, 56,* 496–501. Doi: 10.1037/0022-006X.56.4.496

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

Piper, W. E., & Ogrodniczuk, J. S. (1999). Therapy manuals and the dilemma of dynamically oriented therapists and researchers. *American Journal of Psychotherapy, 53*, 467-82

Plumb, C. J., & Vilardaga, R. (2010). Assessing treatment integrity in acceptance and commitment therapy: Strategies and suggestions. *International Journal of Behavioral Consultation and Therapy. 6*. 263-. Doi: 10.1037/h0100912.

Rakovshik S.G., & McManus F. (2010) Establishing evidence-based training in cognitive behavioral therapy: a review of current empirical findings and theoretical guidance. *Clinical Psychology Review. 30,* 496–516. Doi: 10.1016/j.cpr.2010.03.004

Reichelt, F., James, I. A., & Blackburn, I. (2003). Impact of training on rating competence in cognitive therapy. *Journal of Behavior Therapy and Experimental Psychiatry, 34,* 87-99. Doi:10.1016/s0005-7916(03)00022-3

Roe, R. A. (2002). What makes a competent psychologist? *European Psychologist, 7,* 192-202. Doi: 10.1027//1016-9040.7.3.192

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

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

Schwarz, N., Knauper, B., Hippler, H., Noelle-Neumann, E., & Clark, L. (1991). Rating Scales: Numeric Values May Change the Meaning of Scale Labels. *Public Opinion Quarterly, 55,* 570. Doi:10.1086/269282

Southam- Gerow, M. A., & McLeod, B. D. (2013) Advances in applying treatment integrity research for dissemination and implementation science. *Clinical Psychology science and practice, 20,* 1-13. Doi: 10.1111/cpsp.12019.

Sharpless, B. A., & Barber, J. P. (2009). The Examination for Professional Practice in

Psychology (EPPP) in the era of evidence-based practice. *Professional Psychology: Research and Practice, 40,* 333-340. Doi: 10.1037/a0013983.

Shaw, B. F., Elkin, I., Yamaguchi, J., Olmsted, M., Vallis, T. M., Dobson, K. S., . . . Imber, S. D. (1999). Therapist competence ratings in relation to clinical outcome in cognitive therapy of depression. *Journal of Consulting and Clinical Psychology, 67,* 837-846. Doi: 10.1037/002-006X.67.6.837

Sheen, J., McGillivray, J., Gurtman, C. and Boyd, L. (2015), Assessing the clinical competence of psychology students through Objective Structured Clinical Examinations (OSCEs): Student and staff views. *Australian Psychologist, 50,* 51–59. Doi:10.1111/ap.12086

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86,* 420-428. Doi:10.1037//0033-2909.86.2.420

Sperry, L. (2010). *Core competencies in counseling and psychotherapy: Becoming a highly competent and effective therapist*. New York, NY: Routledge.

Streiner, D. L. (2003) Starting at the beginning: An introduction to coefficient alpha and internal consistency. *Journal Personality Assessment, 80,* 99-103. Doi: 10.1207/S15327752JPA8001_18

Strunk, D. R., Brotman, M. A., DeRubeis, R. J., & Hollon, S. D. (2010). Therapist competence in cognitive therapy for depression: Predicting subsequent symptom change. *Journal of Consulting and Clinical Psychology, 78,* 429–437. Doi: 10.1037/a0019631

Svartberg, M. (1999). Therapist competence: Its temporal course, temporal stability, and determinants in short-term anxiety-provoking psychotherapy. *Journal of Clinical Psychology, 55,* 1313-1319. Doi: 10.1002/(SICI)1097-4679(199910)55:10<1313::AID-JCLP12>3.0.CO;2-F

Terwee, C.B., Mokkink, L.B., Knol, D.L., Ostelo, R. W. J. G., Bouter, L. M., & de Vet, H. C. W. (2012) Rating the methodological quality in systematic reviews of studies on measurement properties: a scoring system for the COSMIN checklist. *Quality of Life Research,* 21, 651. Doi: 10.1007/s11136-011-9960-1

Tracey, T. J., & Kokotovic, A. M. (1989). Factor structure of the Working Alliance Inventory. *Psychological Assessment: A Journal of Consulting and Clinical Psychology, 1,* 207-210. Doi: 10.1037/1040-3590.1.3.207

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. Doi: 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. Doi: 10.1017/S1352465811000622

Wampold, B. E. (2015). How important are the common factors in psychotherapy? An update. *World Psychiatry*, *14,* 270–277. Doi:10.1002/wps.20238

Webb, C. A., DeRubeis, R. J., & Barber, J. P. (2010). Therapist adherence/competence and treatment outcome: A meta-analytic review. *Journal of Consulting and Clinical Psychology, 78,* 200-211. Doi: 10.1037/a0018912.

Weisman, A. G., Okazaki, S., Gregory, J., Goldstein, M. J., Tompson, M. C., Rea, M., & Miklowitz, D. J. (1998), Evaluating Therapist Competency and Adherence to Behavioral Family Management with Bipolar Patients. *Family Process*, 37, 107–121. Doi:10.1111/j.1545-5300.1998.00107.x

Wu, S. M., Whiteside, U., & Neighbors, C. (2007). Differences in inter-rater reliability

and accuracy for a treatment adherence scale. *Cognitive Behaviour Therapy, 36,* 230-239. Doi:10.1080/16506070701584367

Yap, K., Bearman, M., Thomas, N. and Hay, M. (2012), Clinical psychology students' experiences of a pilot Objective Structured Clinical Examination. *Australian Psychologist, 47,* 165–173. Doi:10.1111/j.1742-9544.2012.00078.x

Appendices

Appendix A- COSMIN checklist

Box A. Internal consistency					
		excellent	good	fair	poor
1	Does the scale consist of effect indicators, i.e. is it based on a reflective model?				
Design requirements					
2	Was the percentage of missing items given?	Percentage of missing items described	Percentage of missing items NOT described		
3	Was there a description of how missing items were handled?	Described how missing items were handled	Not described but it can be deduced how missing items were handled	Not clear how missing items were handled	
4	Was the sample size included in the internal consistency analysis adequate?	Adequate sample size (≥100)	Good sample size (50-99)	Moderate sample size (30-49)	Small sample size (<30)
5	Was the unidimensionality of the scale checked? I.e. was factor analysis or IRT model applied?	Factor analysis performed in the study population	Authors refer to another study in which factor analysis was performed in a similar study population	Authors refer to another study in which factor analysis was performed, but not in a similar study population	Factor analysis NOT performed and no reference to another study
6	Was the sample size included in the unidimensionality analysis adequate?	7* #items and ≥100	5* #items and ≥100 OR 6-7* #items but <100	5* #items but <100	<5* #items
7	Was an internal consistency statistic calculated for each (unidimensional) (sub)scale separately?	Internal consistency statistic calculated for each subscale separately			Internal consistency statistic NOT calculated for each subscale separately
8	Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study		Other minor methodological flaws in the design or execution of the study	Other important methodological flaws in the design or execution of the study
Statistical methods					
9	for Classical Test Theory (CTT), continuous scores: Was Cronbach's alpha calculated?	Cronbach's alpha calculated		Only item-total correlations calculated	No Cronbach's alpha and no item-total correlations calculated
10	for CTT, dichotomous scores: Was Cronbach's alpha or KR-20 calculated?	Cronbach's alpha or KR-20 calculated		Only item-total correlations calculated	No Cronbach's alpha or KR-20 and no item-total correlations calculated
11	for IRT: Was a goodness of fit statistic at a global level calculated? E.g. χ^2, reliability coefficient of estimated latent trait value (index of (subject or item) separation)	Goodness of fit statistic at a global level calculated			Goodness of fit statistic at a global level NOT calculated

NB. Item 1 is used to determine whether internal consistency is relevant for the instrument under study. It is not used to rate the quality of the study.

Box B. Reliability: relative measures (including test-retest reliability, inter-rater reliability and intra-rater reliability)				
	excellent	**good**	**fair**	**poor**
Design requirements				
1 Was the percentage of missing items given?	Percentage of missing items described	Percentage of missing items NOT described		
2 Was there a description of how missing items were handled?	Described how missing items were handled	Not described but it can be deduced how missing items were handled	Not clear how missing items were handled	
3 Was the sample size included in the analysis adequate?	Adequate sample size (≥100)	Good sample size (50-99)	Moderate sample size (30-49)	Small sample size (<30)
4 Were at least two measurements available?	At least two measurements			Only one measurement
5 Were the administrations independent?	Independent measurements	Assumable that the measurements were independent	Doubtful whether the measurements were independent	measurements NOT independent
6 Was the time interval stated?	Time interval stated		Time interval NOT stated	
7 Were patients stable in the interim period on the construct to be measured?	Patients were stable (evidence provided)	Assumable that patients were stable	Unclear if patients were stable	Patients were NOT stable
8 Was the time interval appropriate?	Time interval appropriate		Doubtful whether time interval was appropriate	Time interval NOT appropriate
9 Were the test conditions similar for both measurements? e.g. type of administration, environment, instructions	Test conditions were similar (evidence provided)	Assumable that test conditions were similar	Unclear if test conditions were similar	Test conditions were NOT similar
10 Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study		Other minor methodological flaws in the design or execution of the study	Other important methodological flaws in the design or execution of the study
Statistical methods				
11 for continuous scores: Was an intraclass correlation coefficient (ICC) calculated?	ICC calculated and model or formula of the ICC is described	ICC calculated but model or formula of the ICC not described or not optimal. Pearson or Spearman correlation coefficient calculated with evidence provided that no systematic change has occurred	Pearson or Spearman correlation coefficient calculated WITHOUT evidence provided that no systematic change has occurred or WITH evidence that systematic change has occurred	No ICC or Pearson or Spearman correlations calculated
12 for dichotomous/nominal/ordinal scores: Was kappa calculated?	Kappa calculated			Only percentage agreement calculated
13 for ordinal scores: Was a weighted kappa calculated?	Weighted Kappa calculated		Unweighted Kappa calculated	Only percentage agreement calculated
14 for ordinal scores: Was the weighting scheme described? e.g. linear, quadratic	Weighting scheme described	Weighting scheme NOT described		

Box D. Content validity (including face validity)				
	excellent	**good**	**fair**	**poor**
General requirements				
1 Was there an assessment of whether all items refer to relevant aspects of the construct to be measured?	Assessed if all items refer to relevant aspects of the construct to be measured		Aspects of the construct to be measured poorly described AND this was not taken into consideration	NOT assessed if all items refer to relevant aspects of the construct to be measured
2 Was there an assessment of whether all items are relevant for the study population? (e.g. age, gender, disease characteristics, country, setting)	Assessed if all items are relevant for the study population in adequate sample size (≥10)	Assessed if all items are relevant for the study population in moderate sample size (5-9)	Assessed if all items are relevant for the study population in small sample size (<5)	NOT assessed if all items are relevant for the study population OR target population not involved
3 Was there an assessment of whether all items are relevant for the purpose of the measurement instrument? (discriminative, evaluative, and/or predictive)	Assessed if all items are relevant for the purpose of the application	Purpose of the instrument was not described but assumed	NOT assessed if all items are relevant for the purpose of the application	
4 Was there an assessment of whether all items together comprehensively reflect the construct to be measured?	Assessed if all items together comprehensively reflect the construct to be measured		No theoretical foundation of the construct and this was not taken into consideration	NOT assessed if all items together comprehen-sively reflect the construct to be measured
5 Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study		Other minor methodological flaws in the design or execution of the study	Other important methodological flaws in the design or execution of the study

Box E. Structural validity				
	excellent	**good**	**fair**	**poor**
1 Does the scale consist of effect indicators, i.e. is it based on a reflective model?				
Design requirements				
2 Was the percentage of missing items given?	Percentage of missing items described	Percentage of missing items NOT described		
3 Was there a description of how missing items were handled?	Described how missing items were handled	Not described but it can be deduced how missing items were handled	Not clear how missing items were handled	
4 Was the sample size included in the analysis adequate?	7* #items and ≥100	5* #items and ≥100 OR 5-7* #items but <100	5* #items but <100	<5* #items
5 Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study		Other minor methodological flaws in the design or execution of the study (e.g. rotation method not described)	Other important methodological flaws in the design or execution of the study (e.g. inappropriate rotation method)
Statistical methods				
6 for CTT: Was exploratory or confirmatory factor analysis performed?	Exploratory or confirmatory factor analysis performed and type of factor analysis appropriate in view of existing information	Exploratory factor analysis performed while confirmatory would have been more appropriate		No exploratory or confirmatory factor analysis performed
7 for IRT: Were IRT tests for determining the (uni-) dimensionality of the items performed?	IRT test for determining (uni)dimension-ality performed			IRT test for determining (uni)dimension-ality NOT performed

Box F. Hypotheses testing				
	excellent	**good**	**fair**	**Poor**
Design requirements				
1 Was the percentage of missing items given?	Percentage of missing items described	Percentage of missing items NOT described		
2 Was there a description of how missing items were handled?	Described how missing items were handled	Not described but it can be deduced how missing items were handled	Not clear how missing items were handled	
3 Was the sample size included in the analysis adequate?	Adequate sample size (≥100 per analysis)	Good sample size (50-99 per analysis)	Moderate sample size (30-49 per analysis)	Small sample size (<30 per analysis)
4 Were hypotheses regarding correlations or mean differences formulated a priori (i.e. before data collection)?	Multiple hypotheses formulated a priori	Minimal number of hypotheses formulate a priori	Hypotheses vague or not formulated but possible to deduce what was expected	Unclear what was expected
5 Was the expected *direction* of correlations or mean differences included in the hypotheses?	Expected direction of the correlations or differences stated	Expected direction of the correlations or differences NOT stated		
6 Was the expected absolute or relative *magnitude* of correlations or mean differences included in the hypotheses?	Expected magnitude of the correlations or differences stated	Expected magnitude of the correlations or differences NOT stated		
7 for convergent validity: Was an adequate description provided of the comparator instrument(s)?	Adequate description of the constructs measured by the comparator instrument(s)	Adequate description of most of the constructs measured by the comparator instrument(s)	Poor description of the constructs measured by the comparator instrument(s)	NO description of the constructs measured by the comparator instrument(s)
8 for convergent validity: Were the measurement properties of the comparator instrument(s) adequately described?	Adequate measurement properties of the comparator instrument(s) in a population similar to the study population	Adequate measurement properties of the comparator instrument(s) but not sure if these apply to the study population	Some information on measurement properties (or a reference to a study on measurement properties) of the comparator instrument(s) in any study population	No information on the measurement properties of the comparator instrument(s)
9 Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study		Other minor methodological flaws in the design or execution of the study (e.g. only data presented on a comparison with an instrument that measures another construct)	Other important methodological flaws in the design or execution of the study
Statistical methods				
10 Were design and statistical methods adequate for the hypotheses to be tested?	Statistical methods applied appropriate	Assumable that statistical methods were appropriate, e.g. Pearson correlations applied, but distribution of scores or mean (SD) not presented	Statistical methods applied NOT optimal	Statistical methods applied NOT appropriate

Box I. Responsiveness				
	excellent	**good**	**fair**	**poor**
Design requirements				
1 Was the percentage of missing items given?	Percentage of missing items described	Percentage of missing items NOT described		
2 Was there a description of how missing items were handled?	Described how missing items were handled	Not described but it can be deduced how missing items were handled	Not clear how missing items were handled	
3 Was the sample size included in the analysis adequate?	Adequate sample size (≥100)	Good sample size (50-99)	Moderate sample size (30-49)	Small sample size (<30)
4 Was a longitudinal design with at least two measurement used?	Longitudinal design used			No longitudinal design used
5 Was the time interval stated?	Time interval adequately described			Time interval NOT described
6 If anything occurred in the interim period (e.g. intervention, other relevant events), was it adequately described?	Anything that occurred during the interim period (e.g. treatment) adequately described	Assumable what occurred during the interim period	Unclear or NOT described what occurred during the interim period	
7 Was a proportion of the patients changed (i.e. improvement or deterioration)?	Part of the patients were changed (evidence provided)	NO evidence provided, but assumable that part of the patients were changed	Unclear if part of the patients were changed	Patients were NOT changed
Design requirements for hypotheses testing				
For constructs for which a gold standard was not available:				
8 Were hypotheses about changes in scores formulated a priori (i.e. before data collection)?	Hypotheses formulated a priori		Hypotheses vague or not formulated but possible to deduce what was expected	Unclear what was expected
9 Was the expected *direction* of correlations or mean differences of the change scores of HR-PRO instruments included in these hypotheses?	Expected direction of the correlations or differences stated	Expected direction of the correlations or differences NOT stated		
10 Were the expected absolute or relative *magnitude* of correlations or mean differences of the change scores of HR-PRO instruments included in these hypotheses?	Expected magnitude of the correlations or differences stated	Expected magnitude of the correlations or differences NOT stated		
11 Was an adequate description provided of the comparator instrument(s)?	Adequate description of the constructs measured by the comparator instrument(s)		Poor description of the constructs measured by the comparator instrument(s)	NO description of the constructs measured by the comparator instrument(s)
12 Were the measurement properties of the comparator instrument(s) adequately described?	Adequate measurement properties of the comparator instrument(s) in a population similar to the study population	Adequate measurement properties of the comparator instrument(s) but not sure if these apply to the study population	Some information on measurement properties (or a reference to a study on measurement properties) of the comparator instrument(s) in any study population	NO information on the measurement properties of the comparator instrument(s)
13 Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study		Other minor methodological flaws in the design or execution of the study (e.g. only data presented on a comparison with an instrument that measures another construct)	Other important methodological flaws in the design or execution of the study
Statistical methods				
14 Were design and statistical methods adequate for the hypotheses to be tested?	Statistical methods applied appropriate		Statistical methods applied NOT optimal	Statistical methods applied NOT appropriate

Section Two: Research Report

A psychometric evaluation of the Psychological Wellbeing Practitioner Competency Rating Scale for Assessment (PWPCS- A) and Treatment (PWPCS-T).

Abstract

Objectives. There are a number of assessment measures of therapist competency in delivering high-intensity CBT. However, there is not currently a psychometrically evaluated assessment for low-intensity CBT. The aim of this research was to evaluate the reliability and validity of the Psychological Wellbeing Practitioner Competency Scale for assessment (PWPCS-A) and treatment (PWPCS-T).

Design. Two studies utilised a quantitative, cross-sectional design, and a cohort, longitudinal, quantitative and qualitative study design.

Methods. Study one collected competency scale ratings from 114 University of Sheffield psychological wellbeing practitioners (PWP) trainees' observed structured clinical examinations. Data was used to determine reliability, responsiveness of scales, and comparative validity. Study two recruited 176 expert, qualified, and novice PWPs who rated a PWP's assessment and treatment session using PWPCS-A and PWPCS-T. Data was analysed to determine the scales reliability and predictive validity.

Results. Excellent reliability, and good comparative and predictive validity was demonstrated for PWPCS-A. The analysis of the PWPCS-T showed moderate reliability and good comparative validity. Neither scales showed responsiveness to change.

Conclusions The PWPCS-A and PWPCS-T are valid and reliable measures of PWP trainee competence. Further research could assess their applicability within clinical practice.

Practitioner Points

- Psychological wellbeing competency scales for Assessment (PWPCS-A) and treatment (PWPCS-T) are reliable and valid measures of practitioner competence in delivering low-intensity CBT interventions to patients with anxiety and depression.

- PWPCS-A and PWPCS-T provide a useful assessment tool for observed structured clinical examinations.
- PWPCS-A and PWPCS-T could be used in further research to investigate therapist effects on patient outcomes.
- Further research is needed to determine the psychometric properties of the PWPCS in clinical settings.
- Further research could explore if the PWPCSs are applicable measures for other mental health conditions.

Introduction

Following growing concerns recognised in the Depression Report (Layard et al, 2006) regarding a lack of availability of evidenced-based psychological treatment, Improving Access to Psychological Therapy (IAPT) services were launched in the UK in 2008 (Care Services and Improvement Partnership Choice & Access Team, 2008). The aim of IAPT services was to address the need for accessible dissemination of evidence-based psychological therapies for people with mental health concerns (Williams, 2015). The model has transformed the NHS delivery of psychological therapy since its inception (Green, Barkham, Kellett & Saxon, 2014).

IAPT service delivery is based on the provision of recognised and researched clinical practice and is consistent with the National Institute for Clinical Excellent (NICE; 2016) guidelines for treating depression and anxiety (Clark, 2011). The IAPT service model offers a stepped care approach, whereby patients are provided with the lowest appropriate service in the first instance, then 'stepped up' when higher intensity treatment is clinically required. (Bower & Gilbody, 2005).

The lowest intensity IAPT service provision (Step 2) involves low-intensity cognitive behavioural therapy (CBT) treatments for patients with mild to moderate anxiety or depression. Within the IAPT framework, patients accessing the service at step 2 receive facilitated self-help delivered by Psychological Wellbeing Practitioners (PWPs) (Robinson, Kellett, King, & Keating, 2012). The PWP's role is to assess common mental health concerns and devise shared treatment plans with the aim of relieving psychological distress (Williams, 2011; British Psychological Society, 2013). Treatment plans are dependent on the presenting mental health concerns and involve cognitive restructuring, problem solving, behavioural activation, and exposure techniques.

In comparison to service delivery for more complex patients, PWPs provide short-term treatments, have briefer sessions, and consequently hold a comparatively high caseload (Clark et al., 2009). Therefore, delivery of Step 2 care requires the PWPs to be highly skilled. Training involves a 1-year Post-Graduate Certificate following a practical, competency-based national curriculum (Richards & Whyte, 2009). The course requires trainee PWPs to work within an IAPT service for its duration, working with service users under close supervision. Assessment of PWP's clinical competence is carried out through Observed Structured Clinical Examinations (OSCEs) throughout the course (Richards & Whyte, 2009).

A meta-analysis by Twomey, O'Reilly and Byrne (2015) showed that low-intensity CBT is an effective treatment model for patients with anxiety and depression. However, there is growing research to suggest that therapist effect can be an influential factor in successful patient outcomes (Crits-Christoph et al., 1991; Firth, Barkham, Kellett & Saxon, 2015). Recent studies, specifically on PWPs have demonstrated that therapist effects can range from 1% (Ali et al., 2014) to 7-9 % (Green et al., 2014; Firth et al., 2015). The results of these studies show that higher rates of reliable and clinically significant change in clinical outcomes were seen for patients who were working with the most effective PWPs. This heterogeneity of effectiveness between PWPs suggests differences in practitioner's competency, highlighting that ensuring consistency of competency in delivery of low intensity approaches is a critical factor in ensuring successful outcomes for patients (Ginzburg et al., 2012).

Competency entails the concurrent application of knowledge, therapeutic skills, clinical reasoning, communication, emotion, values, and understanding (Barber, Sharpless, Klostermann and McCarthy, 2007). In addition to promoting successful client outcomes, ensuring therapist competency in treatment delivery is crucial in providing safe, quality care; enabling the dissemination of evidence-based practice;

improving the validity of comparative research (Fairburn & Cooper, 2011); and refining and evaluating the training and supervision of therapists (Kohrt et al., 2015).

Levels of competency within high-intensity CBT practitioners are assessed through psychometrically evaluated rating scales such as the Cognitive Therapy Scale-Revised (the CTS-R; Blackburn et al., 2001), or through diagnosis specific rating scales such as the cognitive therapy competence scale for social phobia (CTCS-SP; Consbruch, Clark & Stangier, 2011). However, the qualitative differences in the method of delivery between low-intensity and high-intensity treatments mean that different therapist competencies are required (Roth & Pilling, 2007) for PWPs; therefore high-intensity rating scales would not be applicable for their assessment. Currently, there are no validated outcome measures to assess clinical competence in the delivery of low intensity treatment. Burns, Kellett and Donohoe (2015) highlighted the need for the development of a competency measure specifically for low intensity practitioners.

Aim

A method of assessment of PWP competence in delivering low-intensity treatment was developed for patients with mild to moderate anxiety or depression in accordance with the PWP curriculum (Richards and Whyte, 2011). This included two practitioner competence rating scales: the PWP Competency Scale for Assessment (PWPCS-A), measuring practitioner competence in undertaking a patient-centred assessment; and the PWP Competency Scale for Treatment (PWPCS-T) measuring competence in providing CBT-based low-intensity treatment. These are referred collectively as PWPCSs

The aim of this research is to provide extensive analysis of the psychometric qualities of the PWPCSs, through an evaluation of their reliability and validity in order to ensure that the PWPCSs are consistent and accurate measures of PWP competence for the use in training.

Research Question and Hypotheses

The aim of the research is to answer the following research question: Are PWPCSs valid and reliable measures of PWP competency in delivering low intensity treatment for anxiety and depression?

The hypotheses are:

1) Consistent scores of internal consistency will be shown. Good internal consistency demonstrates that items on a scale measure the same construct (Tang, Cui, & Babenko, 2014).
2) There will be consistent agreement between raters using the PWPCS-A and PWPCS-T. Reliability can be demonstrated through an assessment of interrater reliability showing consistency between ratings provided by multiple assessors (Hallgreen, 2012).
3) The PWPCSs will show a good measure of responsiveness to change which will be seen through an increase in ratings when applied over different time points over the year-long PWP training course. Research has shown that competency levels increase as trainees progress through a CBT training course (McManus, Westbrook, Vazquez-Montes, Fennell, & Kennerley, 2010; Muse, McManus, Rakovshik, & Thwaites, 2017).
4) The PWPCSs will show a significant positive relationship with assessed measures of therapeutic alliance. This is based upon past studies which have shown that a high level of therapist competence leads to increased therapeutic alliance (Ackerman & Hilsenroth, 2003; Del Re, Fluckiger, Horvath, Symonds, & Wampold, 2012).
5) The PWPCSs will show good predictive validity by demonstrating that novice PWPs will provide higher ratings of competence (more pass rates) than expert or

qualified practitioners. Brosan, Reynolds and Moore (2008) found that trainee therapists self-assessment of competence was often over-optimistic.

Method

Design

This research is an extensive evaluation of the psychometric qualities of the PWPCSs, testing the research hypotheses by utilising data from across two studies. The first study employed a cohort, longitudinal, quantitative and qualitative design. The second study had a quantitative and cross-sectional design.

PWPCS design. The PWPCS- A and PWPCS- T were designed by PWP trainers (n=3) from the University of Sheffield PWP training course in conjunction with practicing PWPs (n=5). The PWPCSs were developed based on previous competency and adherence rating scales and the PWP national curriculum (Blackburn et al., 2001; Richard & Whyte, 2011). The scale went through five amendment processes prior to completion. An additional 16-page manual for PWPCS-A and a 28-page manual for PWPCS-T were developed to ensure rating accuracy in completing the scales (see Appendix C).

The PWPCSs were developed to assess PWP competencies in delivering assessment and treatment sessions. The scales are appropriate for use with common mental health problems (anxiety disorders and depression). The PWPCSs utilise a 7-point Dreyfus (1989) competency ratings scale. The 7 points are *incompetent* (1), *novice* (2), *advanced beginner* (3), *competent* (4), *proficient* (5), and *expert* (6). Each domain on the PWPCSs provide items for suggested features of the competencies. There are six domains and 34 items for PWPCS-A and six domains and 26 items for PWPCS-T.

The PWPCS-A scale's six competency domains are: *introducing the session*; *establishing and maintaining engagement*; *interpersonal skills*; *gathering problem focused information*; *information giving suitable to the presenting problem*; and *shared planning and decision making.*

The PWPCS-T scale also includes six competency domains and these are: *focusing the session*; *establishing and maintaining engagement*; *interpersonal skills*; *gathering information specific to change*; *delivering within session self-help change methods*; and *planning and shared decision making.*

PWPCS development. Expert PWP trainers (n=3) examined and rated the relevance of each competency domain and items within the domains. The experts had extensive experience in teaching low intensity and high intensity CBT and were qualified IAPT supervisors. They completed the Content Validity Index (CVI) (Lynn, 1986) for PWPCS-A and PWPCS-T. This determined the degree to which the content was relevant and representative to the domain it intended to measure (Haynes, Richard & Kubany, 1995). The CVI was used to determine the content validity of each competency domain and suggested items within the domains. The CVI used a 4-point Likert scale: with 1 being *not relevant*, 2 *somewhat relevant*, 3 *quite relevant*, and 4 as *highly relevant* (Polit & Beck, 2006) (see Appendix D).

Item scores were calculated based on the number of *quite* or *highly relevant* ratings. Convergent scores for each item or domain on the CVI over .67 were considered acceptable (Lynn, 1986), with ratings higher than .9 showing excellent content validity (Polit & Beck, 2006). The results showed agreement for the total competency domain items (T-CVI = 1) except for *Acknowledges the problem by use of complex reflections* (I-CVI= .66). This item on the engagement competency domain was therefore amended to include *simple and complex reflections* for the PWPCS-A and PWPCS-T.

Exploratory and confirmatory factor analysis was carried out (Limon, 2017) to further assess the factor structure of the PWPCSs. The exploratory analysis extracted a unidimensional factor solution, with a latent construct of ‘overall competency’ (47.45%

for PWPCS-A and 54.77% for PWPCS-T). The confirmatory analysis demonstrated adequate model fit for measurement invariance over time for both scales.

Cut-off scores for the PWPCSs were determined using the Singh method (Singh, 2006), which showed an established range between 17-20 for PWPCS-A and 17-18 for PWPCS-T. It was agreed that a score equal to or above 18 would determine the practitioner competence pass rate for PWP trainees (Limon, 2017).

Study One

Procedure. The current PWP competency-based curriculum includes 45-days of training in delivering low intensity psychological treatments for common mental health concerns. The modules include: *engagement and assessment; delivering low-intensity therapeutic interventions; knowledge, respect and understanding for values, policies, culture and diversity;* and *working in social and healthcare settings*. The assessment methods for these modules use standardised scenario role plays (OSCEs; Richards & Whyte, 2011).

Recruitment of participants took place over a two year period, involving three PWP trainee cohorts. Data were collected from trainee, video recorded OSCEs which were rated by PWP course trainers (n=5) using PWPCS. Trainee PWPs had OSCEs to assess competencies in assessment and in delivering treatment. There was no missing data, as PWPCSs were used for course assessment purposes.

OSCEs were carried out at different intervals during the one-year PWP training course. Firstly, PWPs had practice (formative) OSCEs with PWP trainee's peers as clients using a pre-prepared scenario. PWP trainers rated PWP performance in the OSCEs and provided scores on the PWPCSs to inform PWPs on areas of development, for which they received further training and support.

After two weeks, the PWPs completed the assessed (summative 1) OSCE with an actor (as the client, with training and a script). The recordings were assessed by PWP

course trainers (n=7). PWPs total scale scores were passed or failed and those who had received a failed score (<18 total score, or <3 on an individual competency domain) were provided with an hour one-to-one tuition. After a period of one month PWPs completed a further OSCE retake (summative 2) with an actor, which was also recorded and data was collected from the PWPCSs. For each assessment period all actors were asked to perform as clients presenting with the same mental health concern, this changed for each OSCE (formative, summative 1, or summative 2). Table 1 shows the mental health concern presented and treatment method expected for each OSCE assessment period.

PWPs completed formative and summative OSCEs to demonstrate their competence in delivering assessment sessions and treatment sessions. These both followed the same format, except assessment sessions were rated with PWPCS-A and treatment PWCS-T. PWPs completed up to a total (including summative 2) of six OSCEs over the course of the training. Assessment OSCE sessions were 45 minutes long and treatment OSCE sessions were 35 minutes long.

Ten percent of ratings at each stage (formative, summative 1, summative 2) were double marked by another rater (a PWP course trainer). The second raters completed the PWPCSs separately and were unaware of the first marker scores.

Data were also collected from actors involved in the summative OSCEs, who were asked to complete the Working Alliance Inventory (WAI; Horvath & Greenberg, 1989), the Helpful Aspects of Therapy questionnaire (HAT; Llewellyn, 1988) and the Friends and Family test (FFT; NHS England 2014) immediately after each OSCE session. There were no missing data for these questionnaires.

Table 1

Presenting mental health concern for each cohort OSCE (CBT treatment being assessed).

OSCEs	Group		
	2015 (n= 32)	2016 (n= 50)	2017 (n=32)
Formative	Anxiety	-	Anxiety
Summative 1 (Assessment)	Depression	Anxiety	Anxiety and Depression
Summative 2 (Assessment)	Anxiety	Anxiety	Depression
Formative	Depression (cognitive restructuring)	Anxiety (exposure)	-
Summative 1 (Treatment)	Depression (problem solving)	Anxiety (cognitive restructuring)	-
Summative 2 (Treatment)	Anxiety (exposure)	Depression (behavioural activation)	-

Outcome measures. For analysis of the comparative validity the following outcomes were utilised:

Working Alliance Inventory. The 12-item Working Alliance Inventory (WAI; Horvath & Greenberg, 1989) is a post-session, self-report measure used to assess the client's perspective on the therapeutic alliance/relationship and collaborative agreement on goals and tasks. The measure has good internal consistency (0.88) and test-retest reliability (0.78) (Schlosser & Kelso, 2005) (see Appendix H).

Helpful Aspects of Therapy. The Helpful Aspects of Therapy form (HAT; Llewellyn, 1988) is a self-report measure used to determine the client's view on the events that were helpful or hindering in the psychotherapy session. The form contains seven questions, where clients are asked to report on events during the session and provide a rating (9-point Likert scale) on the extent it had been helpful or hindering (see Appendix I). There is currently no evaluation of the measure's psychometric qualities.

Friends and Family Test. The Friends and Family Test (FFT; NHS England 2014) is a self-rating question which asks one question about the likelihood that they would recommend the service to their friends and family. This is rated from *extremely likely* to *extremely unlikely* or *don't know* (see Appendix I).There is currently no psychometric evaluation for this measure.

Participants. The participants in study 1 were the PWP trainees, the raters, and the actors involved in the OSCEs. Participants were provided with information regarding the study (see Appendix D) and were informed that their data would be used in a study to investigate the validity and reliability of the PWP competency scales. Participants included in the study signed consent for the use of their data (see Appendix E).

PWP trainees. Data was collected from three cohorts on the University of Sheffield PWP training course (n= 37 for 2015, n= 50 for 2016, n= 32 for 2017). As the

training is at entry level, none of the trainees had prior experience specifically in delivering CBT interventions before the course.

Raters. The OSCE raters (n=7) were PWP trainers on the University of Sheffield PWP training course. Three were qualified high intensity CBT trainers, three were PWP trainers, and one was a clinical psychologist. They all had extensive experience working, educating, and supervising trainees within IAPT. They all received training on how to use the PWPCSs, and received the PWPCS manuals when rating (see Appendix B).

Actors. The actors (n=5) were employed by the University of Sheffield to play clients for the PWP trainee OSCEs. The same professional actors were consistent throughout the three cohorts and all had previous experience in playing roles within OSCEs.

Data analysis.

Data analyses were completed using SPSS version 21 (IBM Corp, 2012).

Internal consistency. Internal consistency (hypothesis one) was determined through an analysis of Cronbach's alpha scores, item-total correlations, and Guttmann split-half reliability. Cronbach's alpha was calculated using the domain scores for the OSCE PWPCS-A (n=267) and PWPCS-T (n= 164). Scores above .8 were considered acceptable. Item-total calculations of the six domain scores utilised all the data from PWPCS-A (n= 380) and PWPCS-T (n=326) from study one and study two. Inter-item correlation coefficient scores above .30 were deemed acceptable (Cristol et al., 2007; Streiner & Norman, 2003). Guttmann split-half reliability coefficients were also calculated to assess the split-half reliability of the PWPCS-A (n=380) and PWPCS-T (n=326) data collected from both study one and two. Coefficients above .8 demonstrated good correlations when the PWPCS data is randomly split into two halves.

Interrater reliability. Previous studies of the psychometric qualities of competency rating scales have tested reliability using various methods, but there is currently no 'gold standard' for reliability assessment of rating scales (Gordon, 2006; von Consbruch, Clark, & Stangier, 2011). Therefore, to ensure accuracy, the interrater reliabilities of the PWPCSs were analysed across both studies.

For study one, to test hypothesis two, two-way mixed effects intra-class Correlation Coefficients (ICC; Shrout & Fleiss, 1979) with absolute agreement were calculated for the first and second markers for the OSCE data for PWPCS-A and PWPCS-T (n=70). Data were interpreted using Koo and Li (2016) ranges: values were defined as less than .5, .5 to .75, .75 to .9, and greater than .90. These were *poor, moderate, good* and *excellent* respectively.

Scale responsiveness. To determine the responsiveness of the PWPCSs to detect change (hypothesis three) the ratings between each OSCE stage (formative, summative 1, summative 2) were compared. PWPCS responsiveness was assessed with T-tests to determine whether the study groups significantly differed from each other. Total scale scores means were compared between formative and summative 1 for PWPCS-A (n=63) and PWPCS-T (n=70), and between summative 1 and summative 2 OSCEs (n=28 for PWPCS-A and n=16 for PWPCS-T).

Comparative validity.

Pearson's correlation coefficients were calculated to assess whether there was a relationship between the PWPCSs and other outcome measures of similar construct (WAI, FFT and HAT form) (hypothesis four).

A chi-squared test was used to assess the goodness of fit between PWPCS-A and PWPCS-T ratings with the FFT question ('*would you recommend this PWP to friends or family?'*). The percentage of PWPs who failed the OSCE and who would not recommended by the actor (FFT) was graphically presented.

To determine the relationship between the HAT results and the PWPCSs, both quantitative and qualitative methods were utilised. Pearson's correlation coefficient was calculated to assess the relationship between the total HAT form scores and PWPCS total scale scores. The hindering aspect scores were inverted. The percentage of negative comments for passed and failed PWPCS- A and PWPCS- T were calculated. For the qualitative data, a thematic analysis of the actors' written responses was carried out using the Braun and Clark's (2006) recommendations. For each theme, the PWP's domain failure was calculated and presented. This was discussed, along with the qualitative data.

Study Two

Procedure.

Recruitment was undertaken over a two-year period between September 2015 and September 2017. Participants were recruited from three groups of PWP's (novice, qualified, and expert). Participants were asked to sign consent forms (see Appendix E) after reading the study information sheet (see Appendix D) which informed them that their data would be used to investigate the validity and reliability of the PWP competency scales. They were also asked to complete a demographic information page.

***PWP recorded session*.** Each group was asked to view the same video recording of a PWP trainee completing a 45-minute assessment session and a 35-treatment session (video A). They were asked to complete the PWPCSs to rate the PWPs competency with the 'client'. The PWP trainee (from a previous cohort) in the film consented to the use of the recording, as did the PWP trainer who played the role of the client. The 'client' in the assessment session presented with depression and anxiety symptoms in the treatment session.

Participants. In Study two the participants consisted of three subgroups: experts, qualified, and novice PWPs.

Expert group. PWP trainers from various institutions across England attended PWP continuing professional development training events either in London or in Sheffield. The participants (n=24) viewed Video A and rated the competency items and domains using the PWPCSs. Participants were asked not to discuss or alter the results of the PWP competency scales after viewing the film to ensure data were not biased.

Qualified group. Qualified PWPs (n=59) attended the PWP conference in Sheffield and were asked to view Video A. The video of the session was projected onto the screen in the auditorium. The qualified PWPs were asked to complete both PWPCS-A and PWPCS-T during the viewing. The completed scales were collected at the end of the day prior to the qualified PWPs leaving the conference. Participants were asked not to discuss or alter the results of the PWPCSs after viewing the film until the scales were collected to ensure data was not biased.

Novice group. Two cohorts of PWP trainees (novice) (n=30 for PWPCS-A and n=79 for PWPCS-T) were asked to view video A as part of their initial induction onto the PWP training course. They were asked to rate the trainees performance using the PWPCSs, as a learning experience to determine the criteria for competence assessment using OSCEs. Ratings were not discussed prior to collection to avoid bias.

Table 2 presents the demographic information for each of the subgroups. Participants were required to complete each domain section of the PWP competency scales to be included in the final sample. The final research sample was N= 109. All expert PWPs had supervisory experience, 66% of qualified PWPs had been supervising, for an average of 2 years.

Table 2

Demographics of expert, qualified, and novice PWPs

	Group		
	Expert (n= 24)	Qualified (n=55)	Novice (n=30/79)
Females (%)	71	81	90
Males (%)	29	19	10
Mean age in years (SD)	35 (7.27)	38 (11.06)	27 (7.00)
Mean no. of years qualified as PWP (SD)	3 (2.51)	4 (2.91)	0

Note: 7 cases with missing data that could not be allocated for analysis, total N=109 (6% missing data).

Data Analysis. Data analyses were completed using SPSS version 21 (IBM Corp, 2012).

Internal consistency. Cronbach's alpha was calculated to test hypothesis one, using the domain ratings for all group data (n=113 for PWPCS-A and n= 162 for PWPCS-T). Cronbach's alpha (Cronbach, 1951) ranges from 0 (domains independent) and 1 (identical). Scores above .8 were considered reliable (Nunnally & Bernstein, 1994).

Interrater reliability. To determine the interrater reliability (hypothesis two), Intraclass Correlation Coefficients (ICC; Shrout & Fleiss, 1979) were calculated for each participant group for PWPCS-A and PWPCS-T: Novice (n= 30/79); Qualified (n=59/59); Expert (n=24/24). A two-way ICC mixed effects approach with absolute agreement was used as several raters assessed the same session. Data was interpreted using Koo and Li (2016) interpretation ranges of the ICC.

Predictive validity. Hypothesis six was determined by graphically representing the mean total scale scores to show the difference between the expert, qualified, and novice group PWPCS-A and PWPCS-T ratings. The percentage pass rates were calculated. A one-way analysis of variance (ANOVA) was undertaken to determine whether there was significant difference between group means and the Tukey post-hoc test was used to determine specificity between the group differences.

Ethical Considerations

Ethical approval was granted by The University of Sheffield Department of Psychology Research Ethics Committee (see Appendix G).

Results

Descriptive Statistics

Study one. The mean and standard deviations for each cohort for formative, summative 1, summative 2 were calculated (Table 3). For PWPCS-A, the 2016 cohort had the highest mean scores and the 2017 cohort had the lowest. Summative 2 had the highest overall means of all three cohorts.

Table 3

Total rating score Means (SD) for PWP cohorts for formative, summative 1, and summative 2 for PWPCSs

OSCE	Cohorts		
	2015	2016	2017
PWPCS-A			
Formative	20.54 (6.36)	-	20.68 (2.36)
Summative 1	20.27 (3.72)	23.08 (4.12)	22.27 (2.98)
Summative 2	22.20 (2.91)	24.14 (3.22)	22.86 (3.06)
PWPCS-T			
Formative	24.11 (3.16)	24.83 (2.82)	-
Summative 1	23.50 (4.23)	24.71 (3.77)	-
Summative 2	24.27 (5.83)	24.25 (3.49)	-

Note. Missing data presented were data was not available.

The results of an ANOVA comparing means based on presenting mental health condition at each OSCE stage is presented in Table 4 and showed that there were significant differences between means anxiety ($F_{2,3} = 14.91 p < .001$) at formative

OSCEs (depression could not be determined as only one group). At summative 1 there were also significant differences for anxiety ($F_{1,2} = 4.26$ $p=.04$), and depression ($F_{1,2} = 12.27$ $p<.001$). However, there was no significant difference between means at summative 2 ($F_{1,2} = 2.79$ $p=.06$ for anxiety, $F_{1,2} = 3.25$ $p=.08$ for depression).

Study two. The mean and standard deviation for expert, qualified and novice groups are presented in Table 4. The results show discrepancies in the mean scores for the novice group for PWPCS-A compared to similar scores for the expert and qualified groups. For PWPCS-T, the qualified group has the highest mean and the novice group has the lowest total rating score mean.

Table 4

Total rating score Means (SD) for expert, qualified, and novice PWPs for PWPCSs

	Groups		
	Expert	Qualified	Novice
PWPCS-A	16.67 (2.16)	16.11 (2.74)	21.48 (2.77)
PWPCS-T	21.13 (2.47)	23.43 (3.64)	20.98 (2.26)

Hypothesis 1: Internal Consistency

Study One. The calculation of Cronbach's alpha for the total scale scores showed excellent internal consistency for both PWPCSs (α= .91 for PWPCS-A and α=.92 for PWPCS-T).

Study two. Internal consistency of total scale scores for PWPCS-A (α= .87) and PWPCS-T (α= .85) were good for the domain scores for all groups. The average inter-item correlation coefficients were calculated for each domain, and total scale scores for PWPCS-A and PWPCS-T (Table 5). All domains correlated (>.3 using Cristol et al., 2007 cut off) and therefore, it can be assumed that the domains were evaluating the

same constructs. Internal consistency remained valid when tested for domain exclusions. The item total analysis indicated good correlation between domains (>.3).The Guttmann split-half coefficients were calculated from the total scale rating scores and showed excellent internal consistency results, with *rSHG*= .85 for PWPCS-A and *rSHG*= .85 for PWPCS-T.

Table 5

Item-total and inter-item correlations for PWPCS-A and PWPCS-T

Competency domains	Item-total (if deleted)	Cronbach alpha (if deleted)	Competency domains					
Competency domains			Introduction	Engagement	Interpersonal	Info gathering	Change method	Shared planning
Introduction	.64	.86	1.00	-	-	-	-	-
Engagement	.70	.85	.56	1.00	-	-	-	-
Interpersonal	.70	.84	.47	.66	1.00	-	-	-
Info gathering	.69	.85	.57	.51	.58	1.00	-	-
Information giving	.70	.84	.46	.59	.57	.56	1.00	-
Shared planning	.63	.86	.49	.44	.50	.50	.58	1.00

Competency domains	Item-total (if deleted)	Cronbach alpha (if deleted)	Competency domains					
			Introduction	Engagement	Interpersonal	Info gathering	Change method	Shared planning
Focusing session	.52	.85	1.00	-	-	-	-	-
Engagement	.74	.81	.46	1.00	-	-	-	-
Interpersonal	.61	.84	.37	.61	1.00	-	-	-
Info gathering	.61	.84	.39	.47	.46	1.00	-	-
Change method	.64	.83	.43	.60	.44	.46	1.00	-
Shared planning	.74	.81	.46	.66	.53	.58	.56	1.00

Hypothesis 2: Interrater Reliability

Study one. The intra-class correlation coefficients were calculated between the ratings of the first and second (double) marker. The results showed excellent inter-rater agreement (ICC(2, 70)= .91, 95% .82- .96).

Study two. The results of the ICC (Shrout & Fleiss, 1979) showed good reliable correlation scores for PWPCS-A and variable interrater reliability for PWPCS-T for expert, qualified and novice groups (Table 6).

The expert group (n=24) showed excellent interrater reliability for total scale scores for PWPCS-A. (ICC(2,24)= .93, 95% .80-.99). The domain ICCs varied from .81 (95% .37-.99) to .91 (95% .81-.97) showing domain rating scores were within the good to excellent range (using Cicchetti, 1994). For PWPCS-T, the total scale ICC score was within the moderate range (ICC (2,24)= .68, 95% -2.11-.93), with the 95% confidence interval suggesting a large discrepancy between raters' agreement about therapist competence during the treatment session. The lowest domain ICC was for the *interpersonal* competency domain for PWPCS-A (ICC (2,24)= .81, 95% .37-.99) and *change method* competency domain for PWPCS-T (ICC(2,24)= .35, 95% -.94-.92).

The qualified participant group (n=59) also showed excellent interrater reliability for total scale scores (ICC(2, 59)= .96, 95% .91-.99) for PWPCS-A and good interrater reliability for the total scale scores (ICC(2, 59)= .76, 95% .36-.96) for PWPCS-T. Competency domain ICCs are within moderate to excellent range (.79, 95% .52-.95 to .92, 95% .76 -1) for PWPCS-A. The lowest domain ICC was *Interpersonal*. For PWPCS-T the domain ICCs were within moderate range, except *shared planning* which was within the poor range (ICC(2,59)= .36, 95% -1.07-.95).

Table 6

Interclass correlation coefficients (95% confidence intervals) for expert, qualified, and novice groups for PWPCS-A and PWPCS-T.

Competency domains	Expert (n=24)	Qualified (n=59)	Novice (n=30/79)
Introduction	.89 (.73 - .98)	.91 (.77 - .98)	.92 (.80 - .99)
Engagement	.83 (.54 - .98)	.92 (.76 - 1)	.78 (.42 - .96)
Interpersonal	.81 (.37 - .99)	.78 (.41 - .96)	.85 (.56 - .98)
Information gathering	.89 (.79 - .95)	.79 (.52 - .95)	.97 (.93 - .99)
Information giving	.86 (.11 - 1)	.82 (.29 -.99)	.74 (-.04 - .99)
Shared planning	.91 (.81 - .97)	.87 (.59 - 1)	.62 (-.11 - .95)
Total scale score	.93 (.80 - .99)	.96 (.91- .99)	.80 (.46 - .97)

Competency domains	Expert (n=24)	Qualified (n=59)	Novice (n=30/79)
Focusing session	.68 (-2.11 - .93)	.78 (-.29 - 1)	.95 (.82-1)
Engagement	.62 (-.03 - .94)	.73 (.28 - .96)	.90 (.74-.98)
Interpersonal	.81 (.36 - .99)	.81 (.44 - .98)	.85 (.60-.98)
Information gathering	.66 (.20 - .92)	.82 (.56 - .96)	.92 (.79-.98)
Change method	.35 (-.94 - .92)	.77 (.25 - .98)	.80 (.43-.98)
Shared planning	.75 (-.19 - .95)	.36 (-1.07-.95)	.84 (.50-.99)
Total scale score	.68 (-2.11 - .93)	.76 (.36 – 96)	.64 (.06-.94)

The novice participant group (n=30/79) showed good interrater reliability for total scale scores for PWPCS-A (ICC (2,30)= .80, 95% .46- .97) and moderate reliability between raters for PWPCS-T (ICC (2,79)= .64, 95% .06- .94). The domain ICCs for PWPCS-A were within moderate to excellent range, with the lowest domain coefficient being *shared planning* (ICC (2,30)= .62, 95% -.11- .95). The domain ICCs for PWPCS-T were mostly within the excellent range with the lowest being *change method* (ICC (2,79)= .80, 95% .43-.98).

The results showed little difference between the interrater reliability of the three groups. For PWPSC-A, all panel groups were within the good to excellent range, and for PWPSC-T, all groups were within the moderate to good range.

Hypothesis 3: Responsiveness

Responsiveness was determined by analysing whether the PWPCSs could detect change over time. The mean domain and total scale scores for all OSCEs are presented in Table 7 to show whether PWPs increased in competence levels whilst progressing through the training course. The means show an increase from formative to summative OSCE stages for the assessment sessions. The PWPCS-T results showed a decrease in means from formative to summative 1, then an increase to summative 2. The standard deviations scores were highest for PWPCS-T summative 1 and summative 2 (which showed a larger range of scores than other assessment stages).

Table 7

Domain and total scale scores mean (SD) for formative, summative 1 and summative 2 for the PWPCS-A and PWPCS-T.

Competency domains	Formative (n=63/70)	Summative (n=176/78)	Summative 2 (n=28/16)
Introduction	4.02 (.66)	4.27 (.84)	4.56 (.71)
Engagement	3.61 (.66)	3.46 (.79)	3.75 (.62)
Interpersonal	3.61 (.70)	3.84 (.89)	3.91 (.73)
Information gathering	3.44 (.68)	3.72 (.79)	3.70 (.55)
Information giving	3.53 (.65)	3.54 (.82)	3.77 (.73)
Shared planning	3.38 (.75)	3.18 (1.02)	3.52 (.67)
Total scale score	21.26 (3.18)	22.31 (3.89)	23.13 (3.08)

Competency domains	Formative (n=63/70)	Summative (n=176/78)	Summative 2 (n=28/16)
Focusing session	4.61 (.68)	4.31 (.88)	5.03 (1.16)
Engagement	3.88 (.68)	3.83 (.83)	3.63 (.83)
Interpersonal	4.18 (.70)	4.04 (.78)	4.00 (.82)
Information gathering	4.08 (.59)	3.92 (.83)	3.69 (.86)
Change method	4.07 (.68)	3.72 (1.07)	3.56 (.98)
Shared planning	3.78 (.75)	3.47 (.95)	3.81 (1.12)
Total scale score	24.51 (2.98)	23.27 (4.19)	23.72 (4.62)

Figure 1 is a graphical representation of the total scale rating score means for PWPCS-A and PWPCS-T at formative, summative 1, and summative 2 for all OSCEs. The red line shows the pass/fail cut off score. The graph shows that means were above 18 (passed range) for all OSCE stages and scores were clustered in a range of 21 to 24.

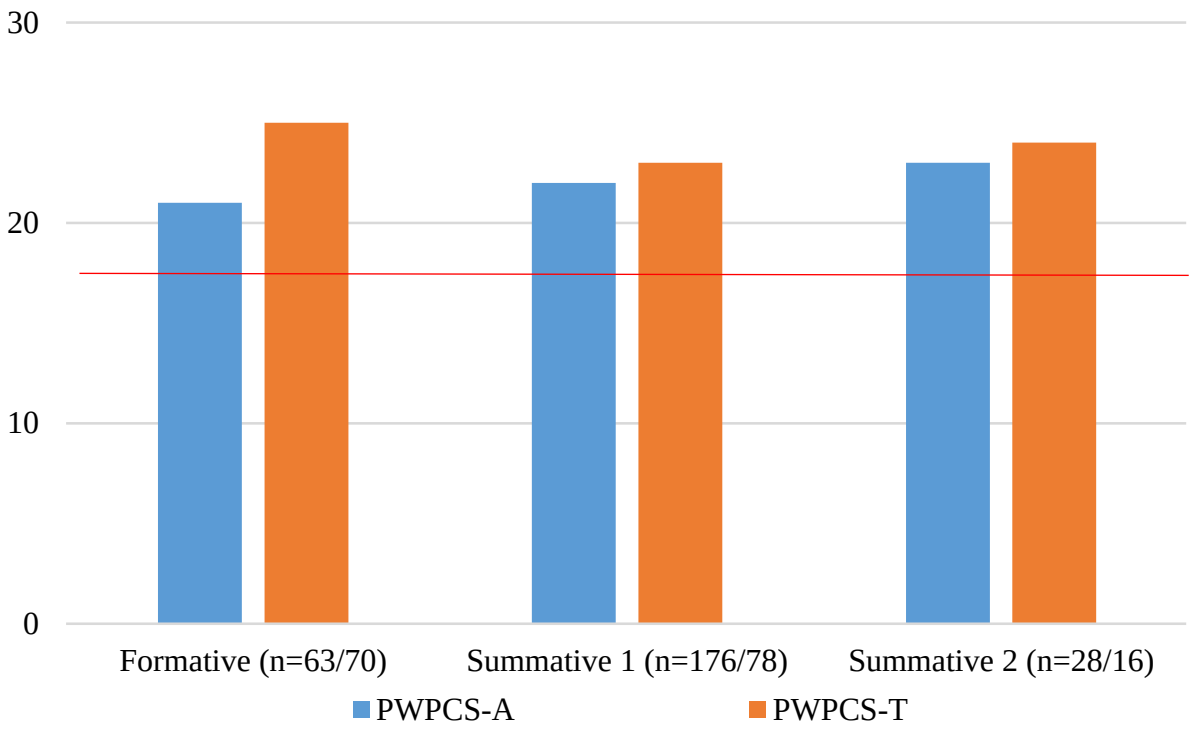


Figure 1. Graphical representation of the mean ratings scores at formative, summative 1, and summative 2 for PWPCS-A and PWPCS-T.

The analysis of the comparison of means (T-tests) showed no significant difference between the means of the assessment formative and summative 1 ratings (t= 1.33 p=.23 for PWPCS-A, t= -2.40 p=.05 for PWPCS-T) or for PWPCS-T summative 1 and 2 (t= .89 p=.41). However, there was a significant difference in the means between the summative 1 and summative 2 ratings for PWPCS-A (t= 2.85 p=.03).

The percentage pass rates at formative and summative 1 were 81% for PWP assessment OSCE and 100 % at summative 2. For the treatment session the pass rate was 90% for the formative, 79% at summative 1, and 90% at summative 2 (see figure 2).

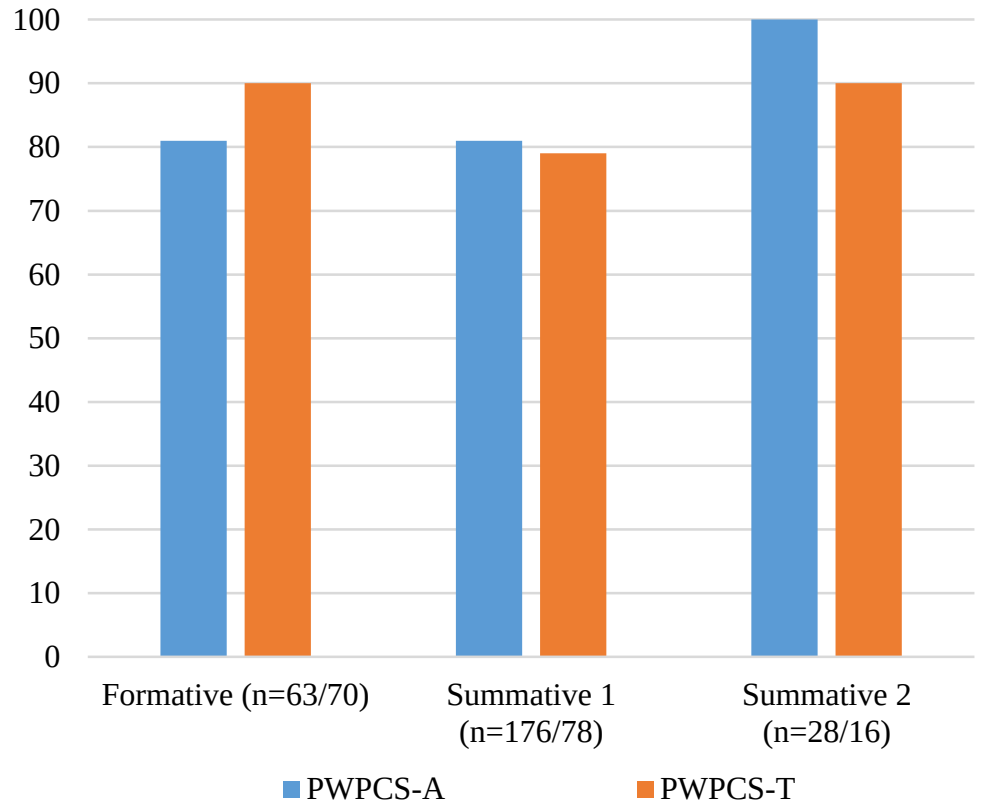


Figure 2. Graphical representation of percentage pass rate on PWPCS-A and PWPCS-F at formative, summative 1 and summative 2.

Hypothesis 4: Comparative validity

The results of the Pearson's correlation coefficient calculations between the PWPCSs and the other measures of similar construct (WAI, HAT and FFT) are presented in Table 8.

Table 8

Correlation (significance) between the PWPCS-A and PWPCS-T and other measures (WAI, HAT and FFT)

Competency domains	WAI				HAT		FFT
	Task	Bond	Goal	Total	Helpful	Hindrance	Total
Introduction	.33 (.06)	.34 (.05)*	.34 (.05)*	.36 (.04)*	-	-	-
Engagement	.47 (.01)**	.43 (.01)**	.62 (.00)**	.54 (.00)**	-	-	-
Interpersonal	.52 (.00)**	.51 (.00)**	.51 (.00)**	.54 (.00)**	-	-	-
Information gathering	.52 (.00)**	.48 (.00)**	.64 (.00)**	.58 (.00)**	-	-	-
Information giving	.67 (.00)**	.60 (.00)**	.56 (.00)**	.64 (.00)**	-	-	-
Shared planning	.49 (.00)**	.33 (.06)	.47 (.00)**	.46 (.00)**	-	-	-
Total scale score	.66 (.00)**	.57 (.00)**	.69 (.00)**	.67 (.00)**	.29 (.11)	.49 (.01)**	.54 (.00)**

Competency domains	WAI				HAT		FFT
	Task	Bond	Goal	Total	Helpful	Hindrance	Total
Focusing session	.17 (.34)	.06 (.74)	.15 (.40)	.08 (.64)	-	-	-
Engagement	.47 (.01)**	.46 (.00)**	.41 (.02)*	.42 (.02)*	-	-	-
Interpersonal	.22 (.23)	.26 (.15)	.24 (.18)	.17 (.37)	-	-	-
Info gathering	.34 (.06)	.35 (.05)*	.31 (.09)	.36 (.04)*	-	-	-
Change method	.66 (.00)**	.61 (.00)**	.64 (.00)**	.65 (.00)**	-	-	-
Shared planning	.28 (.11)	.22 (.22)	.27 (.14)	.24 (.20)	-	-	-
Total scale score	.51 (.00)**	.47 (.01)**	.49 (.00)**	.46 (.00)**	.69 (.00)**	.48 (.01)**	.64 (.00)**

Note. *= $p<.05$ **= $p<.01$

Good significant correlation was demonstrated for all PWPCS-A total scale scores and each of the WAI subsections, as well as the WAI total score. All the domain scores correlated with the WAI, with the exception of the *introduction* competency and the *shared planning* with the *bond* subsection of the WAI. These results demonstrate that higher ratings of competence on PWPCS-A correlated well with higher scores on the WAI.

Correlations were variable for PWPCS-T and WAI. The PWPCS-T total scale scores significantly correlated with the subsection totals of the WAI. However, WAI total scores only correlated with three of the competency domain totals. Only the *engagement* and *change method* showed significant correlation with WAI subsections.

The results of the PWPCSs and the FFT showed good significant correlation, demonstrating that higher competency ratings on the PWPCSs correlated with higher FFT scores. PWPs with a higher level of competency correlated positively with higher recommendation ratings scored by clients (actors).

The Pearson's Chi-square correlation coefficient showed a significant relationship (goodness of fit) between PWP competency ratings and actors *recommendation* scores on the FFT. For PWPCS-A χ^2 (1, 204) = 14.59, p<.001 and for PWPCS-T χ^2 (1, 94)= 5.06, p< .05. Therefore, suggesting a significant relationship between competence and recommendation.

Figure 3 shows the percentage of PWP that had passed or failed on PWPCS-A and PWPCS-T and were not recommended by clients (actors) on the FFT.

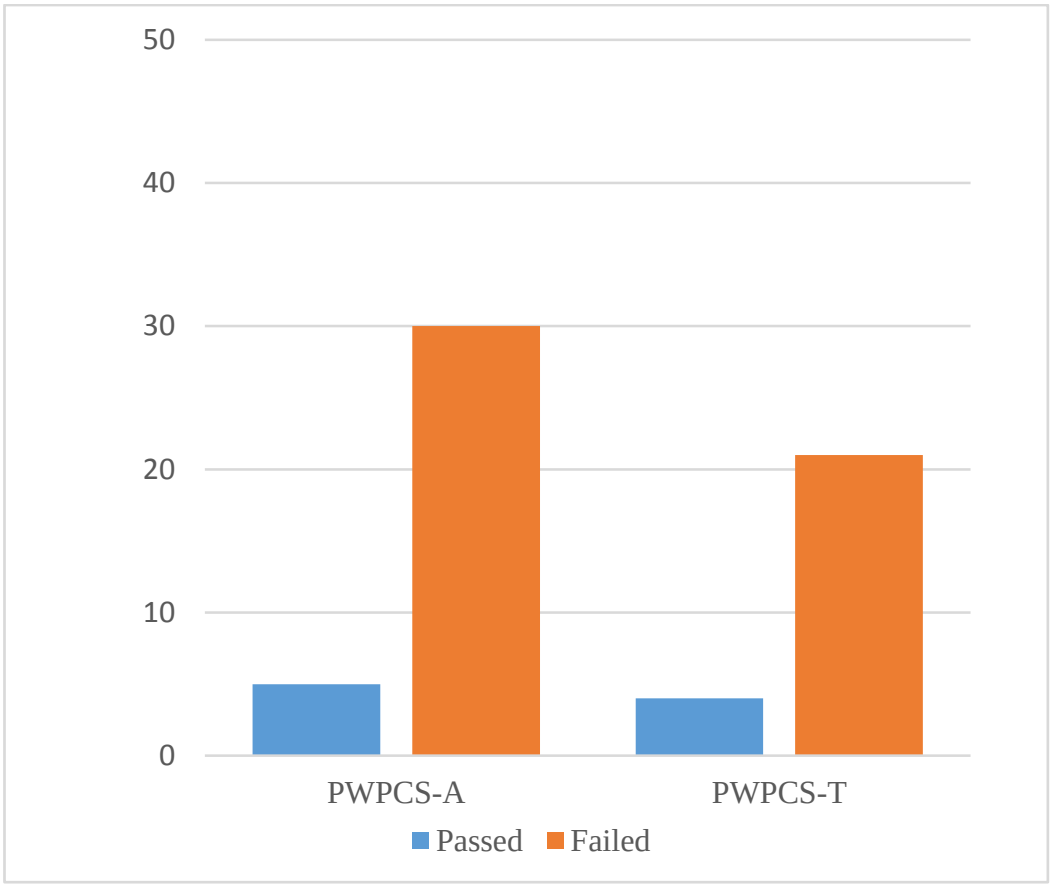


Figure 3. Percentage of passed or failed PWPs who did not receive a recommendation on FFT.

The percentages (in Figure 3) demonstrate that 30% of failed PWPs on PWPCS-A and 21% on PWPCS-T would not be recommended by the client (actor) compared to just 5% (PWPCS-A) and 4% (PWPCS-T) of PWPs that passed.

The Pearson's correlation coefficients were calculated between PWPCS ratings and client scores on the helpful and hindering aspects of therapy (HAT) form. The results showed that PWPCS-A did not correlate with the helpful scores from the HAT. A significant correlation was seen between PWPCS and hindrance aspect scores, thus

showing lower PWPCs ratings correlated with higher scores of hindering aspects of therapy.

The thematic analysis of the qualitative feedback from the HAT produced three themes for the helpful aspects and four themes for the hindering aspects of sessions. The helpful aspect themes were: *an experience of being listened to, empathised with, and reassured; collaborative and structured sessions; confident and knowledgeable PWPs*. The hindering aspect themes were: *experience of not being listened to and being 'rail roaded'; a nervous, uncomfortable, and unprepared PWP; poor timing and pacing of the session; lack of clarity and related missed opportunities during session.*

The actors provided answers for the helpful aspects question for all PWPs (100%). Twenty eight percent of passed PWPs received hindering aspect comments compared to 73% of failed PWPs(scored <18 or <3 on a domain).

The frequency of comments was assessed to determine how many were received for PWPs who had failed, and to which theme comments were relating to. Most of the 62 PWPs had failed in multiple domains and received comments relating to one or more theme. All comments were included for each failed domain. Table 9 demonstrates the total number of helpful aspect comments received for each theme for each domain failure and table 10 shows the hindering comments.

Table 9

Total number of comments (themes) reported by actors as helpful aspects of therapy received for PWPs who had received a failed competency score.

	Competency domain Failure*					
	Introduction	Engagement	Interpersonal	Info gathering	Info giving/ Change method	Shared planning
An experience of being listened to, empathised with, and reassured	4	8	2	7	6	3
Collaborative and structured sessions	1	9	5	8	7	9
Confident and knowledgeable PWPs	2	3	12	1	10	10

Note. *Domain failure- rating scores below 3.

Table 10
Total number of comments (themes) reported by actors as hindering aspects of therapy received for PWPs who had received a failed competency score.

	Competency domain Failure*					
	Introduction	Engagement	Interpersonal	Info gathering	Info giving/ Change method	Shared planning
Not listened to and 'railroaded'	**2**	**9**	**10**	**10**	**12**	**14**
Nervous, unconfident, and unprepared PWP	**0**	**4**	**4**	**4**	**3**	**5**
Poor timing and pacing	**1**	**2**	**2**	**1**	**4**	**9**
Lack of clarity	**1**	**3**	**3**	**1**	**3**	**7**

Note. *Domain failure- rating scores below 3.

Experience of not being listened to and being 'rail-roaded'. The most frequently stated hindrance aspect *was not being listened to* and *'rail-roaded'*. Several actors expressed that within sessions they felt they had not been listening to by the PWP and felt the session had been directed by an agenda set by the PWP rather than collaboratively.

'His guidance in 'reasons against' was driven by him; he didn't use examples to illustrate clearly where he was getting his ideas.' (PWPCS score 12)

'*I didn't feel listen to and I don't think he thought about my concerns. He seemed to want to get through his agenda as quickly as possible.'* (PWPCS score 16)

PWPs who had received this comment were more likely to have failed in multiple areas on the PWPCS (as seen in Table 9). The most failures were seen for the Information giving and shared planning domains. The results show that PWPs that failed on the competencies which focus on collaboration and problem solving were also reported by actors to lack skills in joint working.

Nervous, unconfident and unprepared PWP. The least frequent comment for PWPs that had failed (yet more frequently reported for PWPs who had passed) was regarding the PWPs nervousness and consequently feeling the session was unprepared. Actors highlighted that a hindering aspect of therapy was the PWP behaving overly nervous, unconfident about their practice, and unstructured and unprepared for leading the session.

'... seemed quite nervous.' (PWPCS score 20.5)

'He seemed a little all over the place.' (PWPCS score 17)

The results showed competency failures in interpersonal, engagement, and collaborative working on the PWPCS.

Poor timing and pacing of the session. Actors highlighted that poor timing and paced of the session was hindering, and this was associated with feeling rushed or parts were too slow that other areas were missed.

'I felt rushed and "capped off" at times.' (PWPCS score 19.5)

'The start was so quick I felt a little bewildered, jumped into it, could have spent more time in the intro'. (PWPCS score 19.5)

The highest failure domain rate for PWPs who had received this comment was for the shared decision making competency. This domain was failed most frequently due to the competencies not being met due to timing.

Lack of clarity and missed opportunity during the session. The actors expressed that an aspect of sessions that was unhelpful was a lack of clarity or guidance about CBT. The actors also stated feeling frustrated that the PWP had missed opportunities to gain more information from them (to help guide the CBT intervention).

'Going into the 5 areas model I didn't feel like I understood what the exercise was about and therefore I wasn't quite sure how to answer the questions to fill in each area.' (PWPCS score 20)

'It would have been helpful to have spent a little more time going through the 5 areas once it had been filled in, to help me start to understand how my problem is maintained.' (PWPCS score 15)

'I felt like some of the areas we discussed were not fully explored.' (PWPCS score 19.5)

The results show that PWPs who received this comment on the HAT had a high failure rate on the *shared planning* competency domain.

The helpful aspects of therapy themes are presented below.

Experience of being listened to, empathised with, and reassured. One of the most valued aspects of the therapy session highlighted by the actors was an empathetic PWP. They expressed how they felt comfortable within the session as they felt listened to and their feelings validated.

'*I felt very comfortable and her questioning and empathy instilled trust.'* (PWPCS score 29)

'It was very easy to talk to her because she seemed interested and acknowledged several times about the difficulties I was having. I felt listened to.' (PWPCS score 26)

Collaborative, and structured sessions. A further theme identified was from comments regarding clear and confident PWPs, who were structured in their approach, and remained collaborative.

'The goal setting discussion was very collaborative and the PWP used things I had said previously to prompt me to set my own goals.' (PWPCS score 22)

'…was very clear in his explanations of why we were talking about each section. I felt this helped me to answer more specifically and understand what we were doing.' (PWPCS score 30.5)

Confident and knowledgeable PWPs. The actors highlighted their appreciation of the PWPs positive manner, reassured by their confidence, and that they benefited from their knowledge about the model. The highest frequency of comments relating to this theme were for given to PWPs who had failed on the PWPCSs.

'Her explanations of the 5 areas sounded very encouraging that it would be beneficial for me.' (PWPCS score 24)

'I felt positive about the treatments suggested and therefore optimistic about future sessions.' (PWPCS score 22)

'A really nice efficiently warm and professional manner. I felt I was in safe hands.' (PWPCS score 26)

Hypothesis 5: Predictive Validity

A further analysis was used to examine the differences between expert, qualified and novice ratings of the PWPCSs to test the hypothesis that the scales will show that novice raters will give overly-generous ratings when compared to the other groups.

Figure 4 is a representation of the mean scores for each panel group for the assessment and treatment scales. The red line shows the pass cut off score.

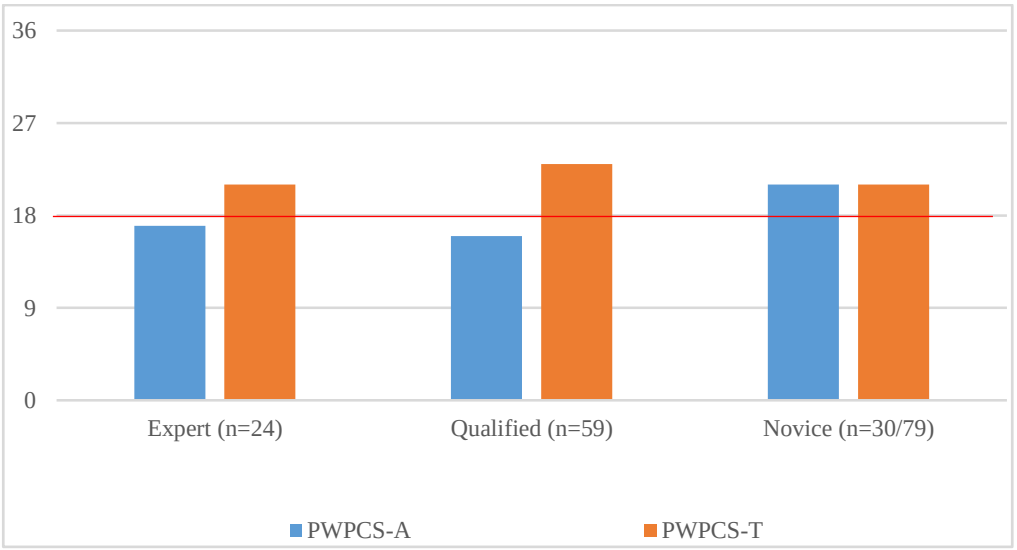


Figure 4. A graphical representation of the mean ratings scores for each group for PWPCS-A and PWPCS-T.

The results show expert and qualified groups ratings increase from the assessment to the treatment whereas the novice group ratings were the same for both sessions (Table 11). The expert and qualified both had mean rating scores below the pass cut off for the assessment and above for the treatment. Experts had the lowest percentage pass rate compared to the other groups (17% for assessment and 83% for treatment). Nearly half qualified PWP group ratings passed (49%) for assessment and 93% for treatment. The novice group had the highest percentage pass rate (89% for assessment and 91% for treatment).

Table 11

Mean (SD) and ANOVA for expert, qualified, and novice group for the PWPCS-A and PWPCS-T.

Competency domains	Groups					
	Expert (n=24)	Qualified (n=59)	Novice (n=30/79)	F (df=2)	*P*	Tukey post-hoc
Introduction	3.46 (.48)	3.93 (.71)	4.17 (.59)	8.38	.00*	N > Q, E
Engagement	2.65 (.65)	3.19 (.70)	3.33 (.69)	7.16	.00*	E < Q, N
Interpersonal	2.38 (.65)	2.80 (.73)	3.15 (.66)	8.30	.00*	N > Q, E
Information gathering	2.75 (.54)	3.29 (.58)	3.58 (.59)	14.24	.00*	E < Q, N
Information giving	2.92 (.49)	3.16 (.75)	3.72 (.65)	10.26	.00*	E < Q, N
Shared planning	2.63 (.65)	2.92 (.92)	3.53 (.76)	8.65	.00*	N > Q, E
Total scale score	16.67 (2.16)	16.11 (2.74)	21.48 (2.77)	41.79	.00*	N > Q, E

Competency domains	Groups					
	Expert (n=24)	Qualified (n=59)	Novice (n=30/79)	F (df=2)	*P*	Tukey post-hoc
Focusing session	3.64 (.63)	3.94 (.69)	3.72 (.60)	2.68	.07	-
Engagement	3.50 (.40)	3.86 (.69)	3.45 (.51)	9.84	.00*	N < Q > E
Interpersonal	3.67 (.57)	3.81 (.84)	3.18 (.54)	2.45	.09	-
Information gathering	3.36 (.60)	3.97 (.60)	3.67 (.51)	3.84	.02	-
Change method	3.40 (.90)	3.97 (.73)	3.48 (.58)	1.98	.14	-
Shared planning	3.39 (.74)	4.28 (.64)	3.52 (.64)	13.11	.00*	Q > N, E
Total scale score	21.13 (2.47)	23.43 (3.64)	20.98 (2.26)	5.17	.00*	Q> N

*Note. * p<.01*

A one-way Analysis of Variance (ANOVA) was calculated and significant differences between PWPCS-A total scale score means were found between the three groups ($F(2, 3)= 41.79$, $p<.001$). Post-hoc comparisons, using the Tukey HSD, indicated that the mean score for the novice group (M= 21.48, SD=2.16) was significantly different from the qualified and expert groups. There were significant differences shown for each competency domain.

The ANOVA for the PWPCS-T also showed significant differences between the mean total scale scores ((F2, 3)= 5.17, $p<.001$). The post hoc comparisons suggested that the mean score for the qualified group (M=23.43, SD= 3.64) was significantly different from the novice group (M=20.98, SD= 2.26). The expert group was not significantly different from either group. For the competency domains only engagement and shared planning showed significance.

Discussion

The aim of this research was to answer a research question by testing a number of hypotheses. The research question was to determine whether the PWPCSs are valid and reliable measures of PWP competency in delivering low-intensity treatment for mild to moderate anxiety and depression. The results tested five hypotheses and showed that the PWPCS-A had excellent internal consistency, excellent interrater reliability, and good comparative and predictive validity. Excellent internal consistency was also shown for the PWPCS-T, moderate interrater reliability, good comparative validity, but was not able to show predictive validity. Neither scale was responsive to changes over time.

Reliability

Results showed that PWPCSs had excellent degrees of internal consistency among competency domain. These results are consistent with findings from other studies of therapist competency rating scales for high-intensity CBT, which also showed excellent internal consistency reliability (Blackburn et al., 2001; Muse et al., 2017).

Interrater reliability was assessed in both studies. An analysis of expert, qualified and novice PWP raters scores showed excellent rater agreement for the PWPCS-A, yet only moderate agreement for PWPCS-T. When exploring differences between scales, the PWPCS-A focuses more on therapist global competencies, in comparison to PWPCS-T, which has more treatment specific competencies. The lowest ICC domain scores for the PWPCS-T were for *change methods* (ICC= .35) for expert PWPs and *shared planning* competencies (ICC= .36) for the qualified group. The results suggest that the differential interrater reliability scores may be due to rater's difficulties agreeing on how specific low-intensity CBT techniques should be applied.

Previous studies have shown that a high level of assessor training is needed to achieve good interrater reliability for CBT rating scales (Barber et al., 2007; Blackburn

et al., 2001; Gordon, 2007; Muse et al., 2017). The lower reliability scores for the PWPCS-T may highlight a need for more intensive training in assessing PWP competency in delivery of low-intensity CBT treatments.

Von Cronsbruch et al. (2012) found that higher levels of interrater agreement are seen when assessing less competent therapists. The mean scores and pass rates suggest that the practitioner seen in video A (study two) was less competent in the assessment session, than treatment. Therefore, the greater agreement between ratings on PWPCS-A than on PWPCS-T, may be reflective of lower levels of practitioner competence seen in video A. This further highlights the need for training to assess PWPs at all levels of competence.

The reliability results for qualified PWPs were excellent for PWPCS-A (ICC= .96) and good for PWPCS-T (ICC=.76). Over 60% of qualified PWP participants were supervising within clinical settings. The high levels of agreement show that the PWPCSs may be appropriate competency rating scales for clinical supervision. However, further research would be needed to determine the validity of PWPCSs in clinical settings.

Validity.

The validity of the scales was assessed by determining whether the PWPCSs could show expected changes over time, whether scales significantly correlated with scores from measures of similar construct, and whether they were able to show predicted outcomes (that novice PWPs would show overly-generous ratings of competency).

Discriminant validity. The results showed that PWPCSs were not responsive in detecting expected changes in levels of competency. Ratings over three assessment time periods during the PWP training course did not show significant increases in

competence levels. The mean scores for PWPC-T even showed a decrease in practitioner competence from formative to summative 1 OSCEs.

The lack of responsiveness of PWPCSs could be due to methodological limitations. Ratings were undertaken immediately after each OSCE period. Therefore, scores may have been subject to bias due to cohort effects. Assessment of PWPs competence could have been influenced by the general level of ability of the cohort group at each assessment. The group may have all improved during the progression of the course and yet competency ratings remain consistent as they were made based on comparisons with others in the cohort. Furthermore, PWP trainers' expectations of trainees is likely to change over the duration of the course which may also influence scoring (and prevent significant increases in scores over time). Previous studies of therapist competency scales have shown significant increases in ratings over the progression of a CBT training course (Blackburn et al., 2001; Muse et al, 2017). However, their methodologies differed from this study, as all video tapes of sessions were collected throughout the course and assessed collectively, using scales, at the end of training, thus reducing the impact and influence of possible cohort effects.

Discrepancies in mean scores may also have been influenced by examination process factors. Formative OSCE sessions were conducted with peers, whereas the summative sessions were assessed examinations with actors. This could also account for the decrease in mean scores from formative to summative 1 seen in PWPCS-T results. PWP's were likely to have felt more nervous and under pressure in summative sessions which could have impacted in their ability to perform clinically.

Comparative validity. Research has shown that a high level of therapist competence leads to increased therapeutic alliance (Ackerman & Hilsenroth, 2003; Del Re et al., 2012). The analysis of the relationship between the PWPCSs and WAI showed significant positive correlation between competency ratings and therapeutic alliance

scores. The highest correlation scores were shown between PWPCS domains and the *goal* WAI subscale. This is expected as the low-intensity CBT treatment model focuses on collaborative goal setting with patients (Twoney et al., 2015). Higher scores of therapeutic alliance were consistent with in higher ratings of therapist competency on the PWPCS, demonstrating that the scales were measuring a competency construct.

The results showed that the weakest relationship was between WAI and *introduction/ focusing session* domains on PWPCS. This domain rates the practitioner's ability to provide information about themselves, their role, and the session. Though this competency is an important aspect of a session, if not completed it is unlikely to impact significantly on the relationship with the client, therefore explaining why low ratings for this domain would not necessarily be reflected in low therapeutic alliance scores.

The PWPCS ratings were also compared with client (actor) qualitative and quantitative responses on the HAT form. The PWPCSs showed that lower levels of competency significantly correlated with higher scores for hindering aspects of therapy. However, the results showed no significant relationship between higher competency ratings and helpful aspects (for PWPCS-A). An explanation could be that actors completing HAT forms are more likely to provide positive scores irrespective of their experience in session, knowing that trainees were part of an examination process, and were likely to receive feedback. This is reflected in the total responses received on the HAT forms (100% completion of qualitative comments for helpful aspects of therapy, compared to less than 50% for hindering aspects).

The analyses of the qualitative data support the findings of the relationship between PWPCSs and the WAI. For example, the *information giving/change method* and *shared planning* domain competencies focus on collaborative working and planning shared treatment goals with patients, when the frequency of HAT comments were assessed in relation to PWP competency failures, comments related to *PWPs not*

listening and *'rail roading'* in the session were most frequently given to PWPs who had failed in those domains on the PWPCSs. The results further provide evidence of PWPCS validity.

The results of the analyses of the relationship between PWPCSs ratings and FFT scores showed a significant positive relationship and association.. This showed, as predicted, that the PWPs with higher ratings of competency received more recommendations from patients (actors).

Predictive validity. Research by Brosan et al.(2008) showed that trainee CBT therapists were more likely provide over-optimistic self-assessments of their competence in delivering therapy. This study aimed to demonstrate PWPCSs predictive validity in showing that novice PWPs rated the practitioner shown in video A at a higher competency level than qualified or expert PWPs. The results showed support for this hypothesis for the PWPCS-A. The mean, ANOVA, and post-hoc test results showed that the novice group ratings were significantly higher than the other groups. The novice group had an 89% pass rate for assessment compared to 17% expert and 49% qualified.

There were no significant differences between the total scores for expert and novice groups for PWPCS-T ratings. However, if the trainee's level of competence had improved from assessment to treatment sessions then discrepancies between groups for PWPCS-T may be more difficult to determine.

The results showed that the qualified group ratings were significantly higher than the novice group. One explanation for this could be that the novice group may only have a limited knowledge of low-intensity treatment techniques, and therefore be unable to recognise practitioner competence in delivery. It may also be considered that the expert group, who are PWP trainers, may be viewing video A from a training perspective and be more likely to be rating whilst identifying trainee development

needs. The qualified group could be less likely to have a training agenda when rating, yet they should have a thorough understanding of competency and low-intensity CBT intervention delivery.

Limitations

This study provided an in-depth evaluation of the reliability and validity of the newly developed PWPCSs. The methodology ensured psychometric quality by meeting criterion set by the Consensus-based Standards for the Selection of health measurement Instruments (COSMIN; Mokkink et al., 2010). The study utilised a number of methods to determine an overall evaluation of the psychometric properties of the PWPCS-A and PWPCS-T.

However, the study did present a number of methodological limitations.

Limitations with the sample population. This research was limited as, within study one, all participants were recruited from the same training institution. Furthermore, data was collected from a homogenous sample group (PWP trainees) and therefore, conclusion about the analysis can only be applied to the application of PWPCSs within a training context.

The studies were limited, in evaluating practitioner competencies in delivering appropriate low intensity interventions, to only two mental health concerns: anxiety and depression. Conclusions therefore, cannot be made about the PWPCSs reliability and validity with different mental health conditions or co-morbidity.

Trepka, Rees, Shapiro, Hardy, & Barkham (2004) state that there are therapist and client factors involved the therapeutic process. The PWPCSs do not assess client related factors which may impact on therapist competence, such as severity of clients' mental health symptoms.

Limitations in the analyses. Previous studies (Karterud et al., 2012; Vallis et al.,1986) showed that interrater reliability decreased when the number of raters reduced. Study two utilised ratings from a large number of participants (n=117). The evaluation in this study did not assess whether interrater reliability remained consistent when fewer raters scores were analysed. However, the results of ICC for double markings of the OSCEs did show excellent interrater reliability (with just two raters).

There may have been some bias associated with the double markings of the OSCEs. Though 10% of OSCEs were meant to be randomly selected for additional assessment to ensure agreement between raters, it was evident through the process of data collection that the majority of double marked OSCEs were for PWPs who had the lowest competency scores. This is likely to be due to trainers wishing to seek further clarity and agreement on scores given. This is likely to bias the level of agreement as second markers may have assumed a failed score had already been given by the first marker. Furthermore, ICCs are more likely to be higher for practitioners with lower competency (von Consbruch et al., 2011) and therefore, the results in study one may not be providing an accurate assessment of agreement at all levels of practitioner competence.

The use of OSCEs as a means of assessment when evaluating psychometric quality may present limitations. Research has been shown that OSCEs are successful and valid method of assessment, however may not be a true representation of clinical practice and consequently, may be subject to bias (Sheen, McGillivray, Gurtman & Boyd, 2015; Yap, Bearman, Thomas & Hay, 2012).

A further limitation of the analysis was that the PWPCSs were assessed for their validity by comparing ratings with scores from the HAT and FFT. Neither of these outcome measures have been psychometrically evaluated and therefore, the usefulness of comparative results may be questionable. Furthermore, the measures of therapeutic

alliance were completed by actors and not by real clients, and therefore the analysis only offers a speculative look on the client/ PWP experience and alliance.

Clinical Implications

Assessment of therapist competency is needed to ensure that quality and skilful therapy is delivered to patients with mental health concerns (Bennett & Parry, 2004; Fairburn & Cooper, 2011; Kohrt et al., 2015). The PWPCSs provide a reliable and validated measure of practitioner competency in delivering low intensity CBT to patients with mild to moderate anxiety and depression. Despite some identified methodological limitations, the PWPCS-A and PWPCS-T can be utilised during training to determine PWPs level of competence, and can help to identify individual developmental needs. The scale can provide a useful tool in the assessment of individual competence, as well as an overview of cohort levels. The PWPCSs, as assessment tools, can provide training institutions with the means of evaluating competence to ensure that trainee PWPs are adequately able to deliver low-intensity CBT treatments skilfully.

The PWPCSs could be useful tools in further investigation into the potential effect of therapist competence on patient outcomes, as well as comparative measures of validity for other assessments of competency in low-intensity CBT.

Further research could be carried out obtain a larger sample of data from across training institutions to further assess psychometric quality. Furthermore, studies could be conducted to determine the PWPCSs utility as supervision tools for clinical practice.

Conclusions

The research showed that the PWPCS-A and PWPCS-T are valid and reliable measures for assessing trainee PWP competencies in delivering low-intensity CBT treatment with clients with mild to moderate anxiety or depression . The scales tested five hypotheses, of which four were accepted. The results showed excellent internal

consistency and interrater reliability, and good comparative and predictive validity for PWPCS-A. The PWPCS-T was moderately reliable with good comparative validity. The results showed that PWPCSs were not responsive to expected changes over time. Discrepancies between scales and the lack of scale responsiveness may be due methodological limitations, and highlight the need for more intensive training on competency rating. Despite limitations, it can be concluded that the PWPCSs have good psychometric properties. Further research could assess the application of the PWPCSs within a clinical context, and for different theoretical models and mental health conditions.

References

Ackerman, S. J., & Hilsenroth, M. J. (2003). A review of therapist characteristics and techniques positively impacting the therapeutic alliance. *Clinical Psychology Review, 23*, 1-33, Doi: 10.1016/S0272-7358.

Ali, S., Littlewood, E., McMillan, D., Delgadillo, J., Miranda, A., Croudace, T., & Gilbody, S. (2014). Heterogeneity in patient-reported outcomes following low-intensity mental health interventions: A multilevel analysis. *PloS ONE, 9,* e99658.

Barber, J. P., Sharpless, B, A., Klostermann, S., & McCarthy, K, S. (2007). Assessing intervention competence and its relation to therapy outcome: A selected review derived from the outcome literature. *Professional Psychology: Research and Practice*, 38, 493-500. doi:10.1037/0735-7028.38.5.493

Bennett, D., & Parry, G. (2004) A measure of psychotherapeutic competence derived from cognitive analytic therapy, *Psychotherapy Research, 14*, 176-192. Doi:10.1093/ptr/kph016.

British Psychological Society (2013) Psychological Wellbeing Practitioner Training Accreditation Handbook (3rd edition). Improving access to Psychological services. Retrieved from http:// www.bps.org.uk/system/files/Public%20files/2013_pwp_handbook_3rd_ed_final.pdf

Bjaastad, J. F., Haugland, B. S. M., Fjermestad, K. W., Torsheim, T., Havik, O. E., Heiervang, E. R., & Öst, L.-G. (2016). Competence and Adherence Scale for Cognitive Behavioral Therapy (CAS-CBT) for anxiety disorders in youth: Psychometric properties. *Psychological Assessment, 28*, 908-916. Doi: 10.1037/pas0000230.

Blackburn, I., James, I., Milne, D., Baker, C., Standart, S., Garland, A., & Reichelt, F. (2001). The Revised Cognitive Therapy Scale (CTS-R): Psychometric properties. *Behavioural And Cognitive Psychotherapy*, 29, 431-446. doi:10.1017/s1352465801004040

Bower, P., & Gilbody, S. (2005) Stepped care in psychological therapies: access, effectiveness and efficiency. The British Journal of Psychiatry, 186 (1), 11-17. Doi: 10.1192/bjp.186.1.11

Brosan, L., Reynolds, S., & Moore, R. (2008). Self-Evaluation of Cognitive Therapy Performance: Do Therapists Know How Competent They Are? *Behavioural and Cognitive Psychotherapy*, *36*(5), 581-587. Doi:10.1017/S1352465808004438

Burns, P., Kellett, S., & Donohoe, G. (2016). "Stress Control" as a Large Group Psychoeducational Intervention at Step 2 of IAPT Services: Acceptability of the Approach and Moderators of Effectiveness. *Behavioural and Cognitive Psychotherapy*, *44*, 431-443. Doi:10.1017/S1352465815000491

Care Services and Improvement Partnership Choice and Access Team (2008) *Improving Access to Psychological Therapies (IAPT) Commissioning Toolkit*. London, UK: Department of Health.

Cicchetti D. V. (1994) Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, 6*, 284–290. Doi: 10.1037/1040-3590.6.4.284

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. Doi:10.3109/09540261.2011.606803

Clark, D.M., Layard, R., Smithies, R., Richards, D.A., Suckling, R. & Wright, B.

(2009). Improving access to psychological therapy: Initial evaluation of two UK demonstration sites. Behaviour Research and Therapy, 47, 910-920. Doi: 10.1016/j.brat.2009.07.010

Crits-Christoph, P., Baranackie, K., Kurcais, J., Beck, A., Carroll, K., Perry, K., Luborsky, L…. & Zitrin, C. (1991). Meta-analysis of therapist effects in psychotherapy outcome studies. *Psychotherapy Research, 1,* 81-91. Doi: 10.1080/10503309112331335511

Cronbach L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16,* 297-334. doi:10.1007/BF02310555

Del Re, A. C., Flückiger, C., Horvath, A. O., Symonds, D., & Wampold, B. E. (2012) Therapist effects in the therapeutic alliance–outcome relationship: A restricted maximum likelihood meta-analysis. *Clinical Psychology Review, 32,* 642-649. Doi:10.1016/j.cpr.2012.07.002.

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

Firth, N., Barkham, M., Kellett, S., & Saxon, D. (2015). Therapist effects and moderators of effectiveness and efficiency in psychological wellbeing practitioners: A multilevel modelling analysis. *Behaviour Research And Therapy*, 69, 54-62. Doi:10.1016/j.brat.2015.04.001

Ginzburg, D., Bohn, C., Höfling, V., Weck, F., Clark, D., & Stangier, U. (2012). Treatment specific competence predicts outcome in cognitive therapy for social anxiety disorder. *Behaviour Research And Therapy*, 50, 747-752. Doi:10.1016/j.brat.2012.09.001

Gordon, P. K. (2006). A comparison of two versions of the Cognitive Therapy Scale. *Behavioural and Cognitive Psychotherapy 35,* 343.Doi: 10.1037/pas0000372

Green, H., Barkham, M., Kellett, S., & Saxon, D.(2014) Therapist effects and IAPT Psychological Wellbeing Practitioners (PWPs): A multilevel modelling and mixed methods analysis. *Behaviour Research and Therapy, 63*. 43-54. Doi: 10.1016/j.brat.2014.08.009.

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. Doi: 10.1017/S1352465801002089

Hallgren, K. A. (2012). Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial. *Tutorials in Quantitative Methods for Psychology*, *8*, 23–34. Doi:10.20982/tqmp.08.1.p023

Haynes, S., Richard, D., & Kubany, E. (1995). Content validity in psychological assessment: A functional approach to concepts and methods. *Psychological Assessment*, *7*, 238-247. Doi:10.1037//1040-3590.7.3.238

Horvath, A. O., & Greenberg, L. S. (1986). Development of the Working Alliance Inventory. In Greenberg, L. S. & Pinsoff, W. M. (Eds.), *The psychotherapeutic process: A research handbook*, 529-556. New York, NY: Guilford.

IBM Corp. (2012). IBM SPSS Statistics for Windows, Version 21.0. Armonk, NY: IBM Corp.

Improving Access to Psychological Therapies (2008).*Improving Access to Psychological Therapies Implementation plan: Curriculum for low-intensity therapies workers.* London, UK: Department of Health.

Koo, T. K., & Li, M. Y. (2016). A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of Chiropractic Medicine*, *15*, 155–163. Doi:10.1016/j.jcm.2016.02.012

Kohrt, B., Jordans, M., Rai, S., Shrestha, P., Luitel, N., & Ramaiya, M. et al. (2015). Therapist competence in global mental health: Development of the Enhancing Assessment of Common Therapeutic factors (ENACT) rating scale. Behaviour Research and Therapy, 69, 11-21. doi:10.1016/j.brat.2015.03.009

Layard, R., Bell, S., Clark, D., Knapp, M., Meacher, M., Priebe, S., Thornicroft, G., Turnbull, A. & Wright, B. (2006). The depression report: A new deal for depression and Anxiety disorders. *Centre for Economic Performance's Mental Health Policy Group.* Retrieved from: EconPapers.repec.org/RePEc:cep:cepsps:15.

Llewelyn, S. (1988). Psychological therapy as viewed by clients and therapists. *British Journal Of Clinical Psychology*, 27, 223-237. doi:10.1111/j.2044-8260.1988.tb00779.x

Limon, E. (2017). *Competencies in delivering guided self-help: exploratory and confirmatory factor analysis* (Unpublished dissertation). University of Sheffield, UK.

Lynn, M. (1986). Determination and quantification of content validity. *Nursing Research*, *35*, 382-386. Doi:10.1097/00006199-198611000-00017.

Martin, D. J., Garske, J. P., & Davis, M. K. (2000). Relation of the therapeutic alliance with outcome and other variables: A meta-analytic review. *Journal of Consulting and Clinical Psychology, 68*, 438-450. Doi: 10.1037/0022-006X.68.3.438

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, Doi: 10.1016/j.brat.2010.08.002

Mokkink, L. B., Terwee, C. B., & de Vet, H. C. W. (2012) COSMIN: Consensus-based

standards for the selection of health status measurement instruments. *Encyclopedia of Quality of Life and Well-Being Research*, 1309-1312.

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. Doi: 10.1037/pas0000372

National Institute for Clinical Excellent (2016) Depression in adults: recognition and management (CG90). From https://www.nice.org.uk/guidance/cg90

NHS England (2014) Friends and Family Test. Retrieved from: https://www.england.nhs.uk/wp-content/uploads/2014/07/fft-imp-guid-14.pdf

Polit, D., & Beck, C. (2006). The content validity index: Are you sure you know what's being reported? critique and recommendations. *Research in Nursing Health, 29,* 489-497. Doi:10.1002/nur.20147

Richards, D. & Whyte, M. (2009). *Reach Out: National programme student materials to support the delivery of training for Psychological Wellbeing Practitioners delivering low intensity interventions.* 2nd Edition. Rethink, UK.

Robinson, S., Kellett, S., King, I., & Keating, V. (2012). Role Transition from Mental Health Nurse to IAPT High Intensity Psychological Therapist. *Behavioural and Cognitive Psychotherapy, 40,* 351-366. Doi:10.1017/S1352465811000683.

Roth, A., & Pilling, S. (2007). 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.* Doi:10.1017/s1352465808004141

Schlosser, L., & Gelso, C. (2005). The advisory working alliance inventory-advisor version: scale development and validation. J*ournal Of Counseling Psychology*, 52, 650-654. Doi:10.1037/0022-0167.52.4.650

Singh, K. (2007) *Quantitative Social Research Methods* London, UK: Sage Publications Inc.

Sheen, J., McGillivray, J., Gurtman, C. and Boyd, L. (2015), Assessing the Clinical Competence of Psychology Students Through Objective Structured Clinical Examinations (OSCEs): Student and Staff Views. Australian Psychologist, 50: 51–59. Doi:10.1111/ap.12086

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86,* 420–428. Doi: 10.1037/0033-2909.86.2.420

Tang, W., Cui, Y., & Babenko, O. (2014). Internal Consistency: Do we really know what it is and how to assess it? *Journal of Psychology and Behavioral Science, 2,* 205-220.

Trepka, C., Rees, A., Shapiro, D.A., Hardy, G. E. & Barkham, M.(2004) Cognitive Therapy and Research, *28,* 143. Doi:10.1023/B:COTR.0000021536.39173.66

Twomey, C., O'Reilly, G. & Byrne, M. (2015) Effectiveness of cognitive behavioural therapy for anxiety and depression in primary care: a meta-analysis. *Family Practice, 32*(1), 3-15. Doi: 10.1093/fampra/cmu060

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. Doi: 10.1017/S1352465811000622

Vu, N. V., & Barrows, H. S. (1994) Use of standardized patients in clinical assessments: recent developments and measurement findings. *Educational Researcher, 23,* 23-30. Doi: 10.3102/0013189X023003023

Webb, C.A., DeRubeis, R.J., & Barber, J.P. (2010). Therapist adherence/competence

and treatment outcome: A meta-analytic review. *Journal of Consulting and Clinical Psychology, 78,* 200-211. Doi: 10.1037/a0018912.

Williams, H. (2011). Is there a role for Psychological Wellbeing Practitioners and Primary Care Mental Health Workers in the delivery of low intensity cognitive behavioural therapy for individuals who self-harm?. *The Journal Of Mental Health Training, Education And Practice*, 6, 165-174. Doi:10.1108/17556221111194509

Williams, C. H. J. (2015), Improving Access to Psychological Therapies (IAPT) and treatment outcomes: Epistemological Assumptions and Controversies. *Journal of Psychiatric and Mental Health Nursing, 22,* 344–351. Doi:10.1111/jpm.12181

Wu, S. M., Whiteside, U., & Neighbors, C. (2007) Differences in inter-rater reliability and accuracy for a treatment adherence scale. *Cognitive Behavioural Therapy, 36,* 230-239. Doi: 10.1080/16506070701584367

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

Appendices

Appendix A- PWPCS- A

LOW INTENSITY ASSESSMENT SESSION **COMPETNENCY RATING:_____/36**

Worker: ……..... Date: ……..

Introduction to the Assessment Session Competency						
	INCOMPETENT	NOVICE	ADVANCED BEGINNER	COMPETENT	PROFICIENT	EXPERT
Introduces self by name ↓						
Elicits patient's full name ↓						
Role of the PWP made clear ↓						
Confidentiality (e.g. supervision, GP, MDS) ↓						
Describes purpose/agenda of assessment and methods used ↓						
Defines time scale for the assessment						
Overall Section Competency Rating	0	1	2	3	4	5 6

Initial Engagement Competencies						
	INCOMPETENT	NOVICE	ADVANCED BEGINNER	COMPETENT	PROFICIENT	EXPERT
Collaborative approach						
Acknowledges the problem by use of simple reflections						
Acknowledges the problem by use of complex reflections						
Use of capsule summaries						
Use of major summaries						
Ratio of questions to feedback						
Overall Section Competency Rating	0	1	2	3	4	5 6

Information Gathering: Problem(s) Focus							
COM-B		INCOMPETENT	NOVICE	ADVANCED BEGINNER	COMPETENT	PROFICIENT	EXPERT
CAPABILITY MOTIVATION OPPORTUNITY	Questioning skills						
CAPABILITY	Elicitation of problem description						
CAPABILITY	Elicits bodily symptoms of presenting problem						
CAPABILITY MOTIVATION OPPORTUNITY	Elicits behavioural symptoms of the presenting problem						
CAPABILITY MOTIVATION	Elicits key cognitions of presenting problem						
CAPABILITY MOTIVATION	Elicits emotions of the patient associated with the presenting problem						

OPPORTUNITY	**Determines the impact of the problem on valued life activities**						
CAPABILITY MOTIVATION OPPORTUNITY	**Risk Assessment**						
	Sensitively integrates MDS into assessment and provides result feedback						
CAPABILITY MOTIVATION OPPORTUNITY	**Recognition of co-morbidity**						
CAPABILITY MOTIVATION	**Assesses other important issues, such as;** • Why do they want help now • Current & past treatments • Current medication and attitude to this treatment						
Overall Section Competency Rating		**0**	**1**	**2**	**3**	**4**	**5**	**6**

Information Giving: Suitable to the Problem						
	INCOMPETENT	**NOVICE**	**ADVANCED BEGINNER**	**COMPETENT**	**PROFICIENT**	**EXPERT**
Co-creates accurate ABC or 5-areas conceptualisation ↓						
Co-creates patient centred problem statement ↓						
SMART treatment goal planning						
Overall Section Competency Rating	**0**	**1**	**2**	**3**	**4**	**5**	**6**

Shared Planning and Decision Making Competencies						
	INCOMPETENT	NOVICE	ADVANCED BEGINNER	COMPETENT	PROFICIENT	EXPERT
Treatment options – offered to patient suitable to presentation ↓						
Agrees plans & actions ↓						
Ends the session						
Overall Section Competency Rating	0	1	2	3	4	5 6

Appendix B - PWPCS- A manual

LOW INTENSITY COGNITIVE BEHAVIOURAL

COMPETENCY SCALE MANUAL

Assessment Sessions

INTRODUCTION

Low intensity cognitive behavioural interventions are often delivered by Psychological Wellbeing Practitioners (PWP) who provide guided self-help (GSH) in a 'coaching' style to patients with mild- moderate common mental health problems. A crucial aspect of the PWP role is the assessment of patients, aiming to identify the patient's main presenting problem and evaluate the suitability of the specific style of the low intensity clinical method and model of intervention for the patient, their problems and their goals. Assessment competencies are also essential in ensuring the safety of the patient and in the right choice of treatment.

ASSESSING FOR BEHAVIOUR CHANGE

Consideration of behaviour change theory is fundamental to the low intensity cognitive behavioural approach. It is essential the practitioners are able to consider the way in which behaviour change underpins the low intensity method and apply this knowledge within the assessment. The integrative model of behaviour and behaviour change that informs PWP work is the COM-B model (Michie et al., 2014). The model conceptualises behaviour change as resulting from the interaction of three factors (a) capability to perform behaviour change (b) the opportunity to carry out necessary behaviour change and (c) the motivation for behaviour change. During assessment, practitioners should utilise the COM-B model to inform and influence the gathering and synthesis of information to aid clinical decision-making and treatment planning. There are no scales measuring the use of COM-B, but the model should be used to inform the assessment process.

The three areas are outlined:

CAPABILITY

Does the patient have sufficient knowledge or skills to change their behaviour/reasoning/executive functioning through understanding of their common mental health problems?

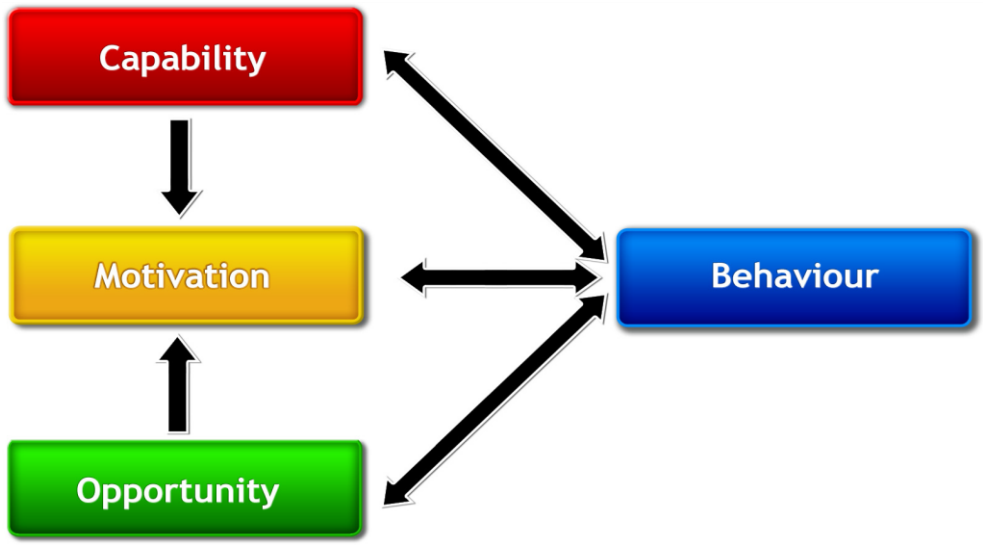
OPPORTUNITY

What factors in the patient's environment maintain the problem behaviour and make behaviour change difficult? Does the patient have sufficient access to resources? What barriers to change need to be considered?

MOTIVATION

What is the patient's current readiness for change? What factors are currently impacting the patient's motivation? Is avoidance currently making change difficult or maintaining the problem? What other factors may play a role in decreasing motivation e.g. drugs/alcohol?

The COM-B model has been mapped to the PWP assessment tool to highlight areas where it will facilitate the PWP with their assessment of the patient and their presenting problem. The model should be applied such that the 3 factors are considered in relation to their impact on the patient's ability to engage in behaviour change, and ultimately to engage in the PWP approach. The model is applied such that it informs PWP treatment planning, informs treatment goals and enables the PWP to anticipate challenges in behaviour change.



LOW INTENSITY COGNITIVE BEHAVIOURAL COMPETENCY SCALE MANUAL

This scale is used to measure the level of competency in practitioners delivering low intensity cognitive behavioural *assessment sessions.* The scale does not measure adherence to the PWP assessment approach (i.e. whether something was done), but rather the competency with which the PWP completed the assessment (e.g. the skilfulness of the assessment and the methods used). The scale contains 6 items to enable raters to examine a range of key competencies:

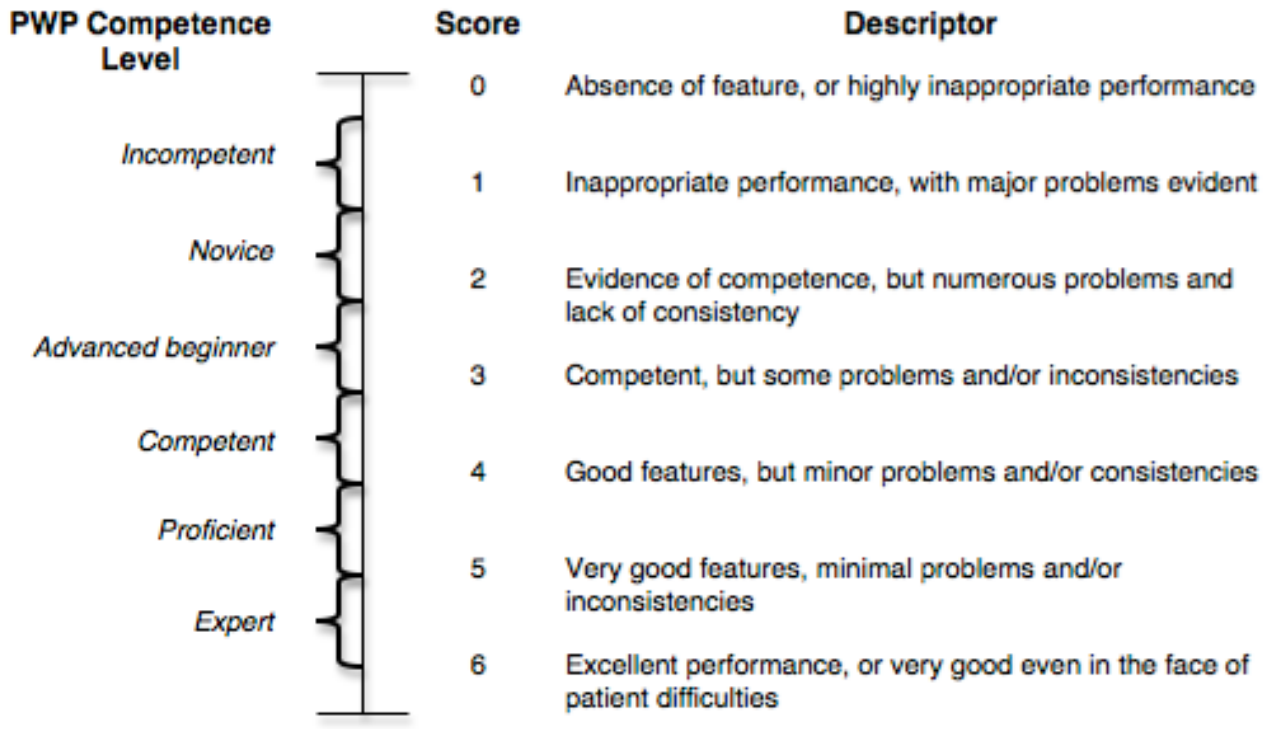
- Introduction to the assessment session
- Engagement competencies
- Interpersonal competencies
- Information gathering competencies: problem focused
- Information giving competencies: suitable to the problem
- Shared planning and decision making competencies

The low intensity cognitive behavioural competency measure is a rating scale to be used by supervisors, trainers and managers to assess practitioner's performance in *assessment sessions*. The examples included within the manual are considered as guidelines. The examples provide both descriptive and explanatory examples for reference. As practice is complex, then raters need to be able to use the manual as guidance to ratings, as exhaustive descriptors cannot be provided.

The scale and manual is suitable for use in benchmarking the competencies of both trainee and qualified PWPs.

SCORING

The low intensity cognitive behavioural assessment competency scale scoring system uses the Dreyfus system (1990), whereby competencies are rated on a Likert scale (0-6). Each level has been defined in detail to conform to the levels of competence. This has been set out in the table below.



For a low intensity practitioner to be graded as competent in an assessment session, the session has to score ≥18 overall (range 0-36). The PWP must score 3 or more on the summary rating in each of the six sections - half-point scoring is accepted.

The summary rating of each section is NOT the average of the ratings given on specific aspects and is not cumulative.

The competency-rating tool is designed to be appropriate for assessment sessions lasting 30-45 minutes.

Raters are encouraged to use the whole scale during competency assessment. A 6 is often characterised by the application of competencies "in the face of patient difficulties." It is possible to score a 6 in the absence of patient difficulties should the rater feel this provides the most accurate rating of the practitioners competence.

Competency Rating Criteria

Introduction to Assessment Session

The low intensity cognitive behavioural practitioner or PWP should demonstrate competence in introducing themselves and clarifying their role, as well as providing information on the process and features of the assessment – this should be fluently and confidently presented. The practitioner should ensure that the patient understands what to expect will occur in the initial assessment appointment. The key features of the 'introduction to assessment' item as outlined in the low intensity cognitive behavioural competency scale are as follows:

Key features:

- PWP's introduce themselves and gain the patient's full name and preferred name
- Role clarification
- Outline confidentiality and its boundaries
- Describing the purpose of the assessment session and what methods will be used
- Defining a time scale for the assessment session

At the start of the assessment session the practitioner should *introduce their name and their role*. This should be welcoming and clear.

Confidentiality should be described fully. The patient should be informed that information discussed in session will not be shared with anyone beyond the Primary Care team, in terms of record keeping and supervision. In terms of risk concerns then the practitioner should inform the patient about who they would share information with in such circumstances that there is concern about the level of risk posed to the patient or others. Confidentiality should be agreed with the patient.

The practitioner should explain the purpose of the assessment is to develop a *shared understanding of the problems to inform appropriate treatment or signposting*. The assessment methods should be explained to the patient for example; defining exactly what the problem is, completing outcome measures and discussing appropriate treatment options.

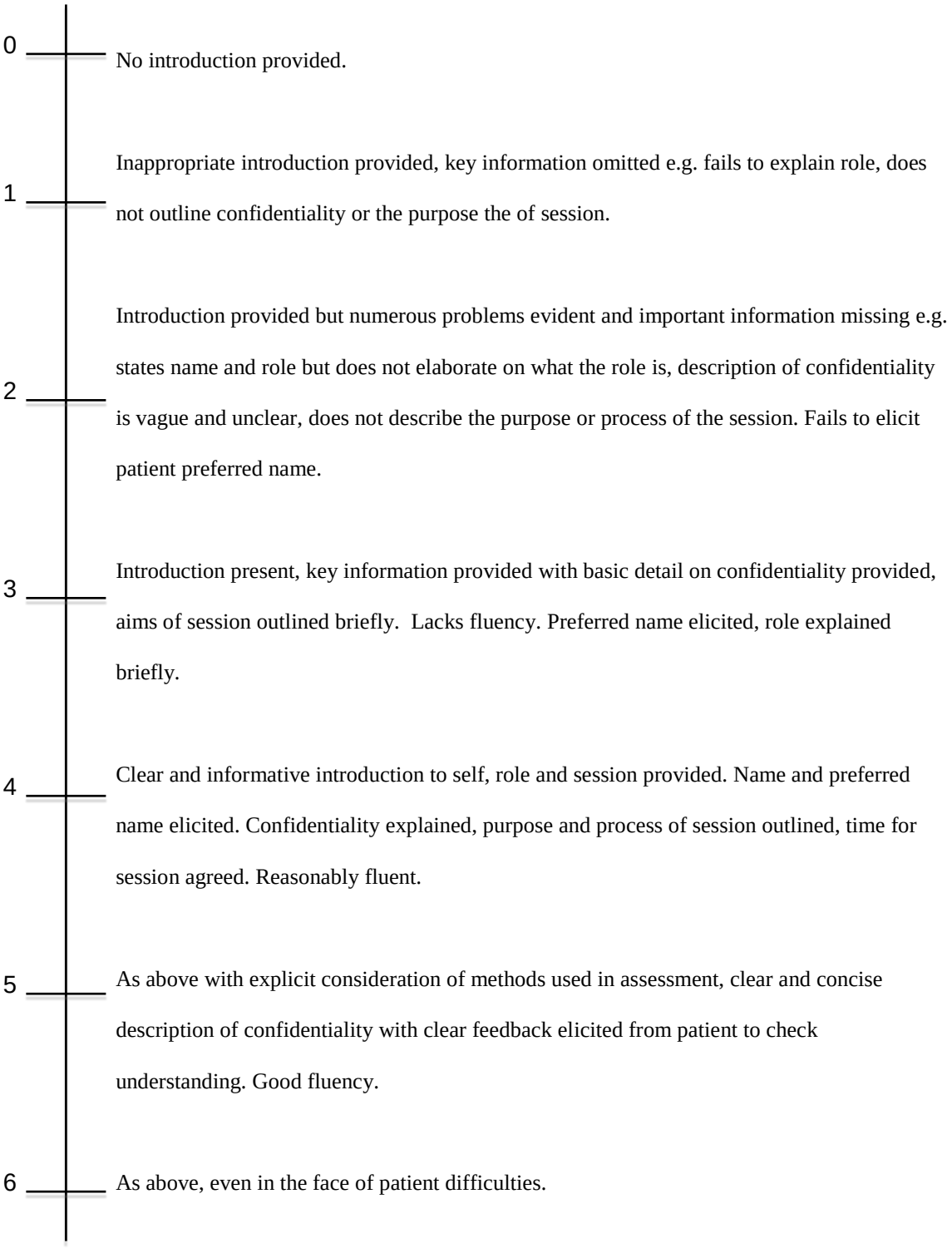
A time scale should be defined and then the session adhere to this time scale.

Checklist-

- Has the practitioner stated their name and asked for the client's full name?
- Have they clarified their job title and given a description of their role?
- Did the practitioner appear confident in their introductions, so putting the patient at ease?
- Has the practitioner outlined how the sessions will be set out (i.e. the methods used)?
- Did the practitioner explain and agree confidentiality and boundaries (e.g. information discussed with supervisor, GP, risk assessment)?
- Was there a time scale for the assessment session clarified?
- Did the practitioner check understanding of all the above when and if necessary?

Introduction to Assessment Session

Competency ratings:



Establishing and Maintaining Engagement

The low intensity cognitive behavioural practitioner or PWP should demonstrate their ability to engage the patient throughout the assessment session. The aim is that the patient feels heard and that their problems are appropriately acknowledged and validated – this is done by a combination and blend of a collaborative stance/approach, reflections, summaries and the key absence of any 'interrogatory' style. The key features of the 'establishing and maintaining engagement' item as outlined in the low intensity cognitive behavioural competency scale are as follows:

Key features:

- Ensuring a collaborative approach
- Acknowledge the problem by reflection
- Using capsule summaries
- Using major summaries
- Appropriate ratio of questions to feedback

The practitioner should ensure a *collaborative stance and approach* is taken during the session to develop a shared understanding of the patient's problems and difficulties. Language should be collaborative in nature (e.g. shall we have a look at how your low mood is impacting on your home life at the moment?). The practitioner should not *falsely collaborate* (e.g. let's look at how we are coping with that' or 'shall we move on?'). When conceptualising, the PWP should ensure that the patient can see and contribute to the conceptualisation.

The practitioner should ensure that problems are *acknowledged by simple and complex reflections* so that the patient feels listened to and that they feel that their problems are validated. The simple reflections should provide a narrative of the current difficulties and enable the practitioner and patient to work towards developing a problem statement (e.g. "so you felt

like you were having a heart attack" or "so you've been feeling really low and crying often."). Complex reflections should be used as appropriate.

The practitioner should ensure that the patient feels listened to be providing appropriate, accurate and regular capsule summaries and also section summaries. The capsule summaries are used to show the patient that the practitioner recognises certain themes or collections of statements about, for example, how the patient has been feeling, acting or thinking. Section summaries are used to create transfer from one section of the assessment process to another. The practitioner should not over chunk or over summarise. The assessment section should end with a brief summary from the practitioner of the process, content and outcomes from the assessment.

There should be an *appropriate ratio of questions to feedback*. This is to ensure that there is not an interrogatory approach to the assessment, and is feedback *to* the patient. Feedback should be elicited *from* the patient to clarify information and ensure an accurate description of the problem is being gained.

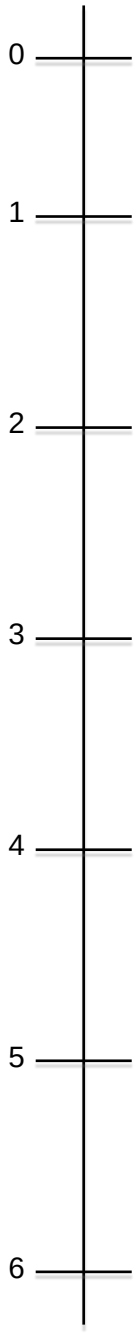
Checklist-

- Was there a collaborative approach to discussing the patient's difficulties?
- Was collaborative language used?
- Was there any false collaboration?
- Was the effort to engage the patient evident across the session?
- Did the practitioner offer a variety of simple and complex reflections?
- Did the practitioner provide capsule and major summaries of the patient's difficulties, without over summarising?
- Were the reflections and summaries appropriate and accurate to the patient's descriptions?
- Was there an appropriate ratio of questions to feedback?
- Was feedback elicited from the patient?

- Did the PWP work with the patient when conceptualising the problem?

Establishing and Maintaining Engagement

Competency ratings:



0 No evidence of attempts to engage patient.

1 Inappropriate or ineffective engagement of the patient, absence of collaboration, absence of summaries. Absence of feedback. An interrogatory style.

2 Attempts to engage patient somewhat patchy across the session. Limited use of summaries and reflections or alternatively over summarising. Limited collaboration and opportunities to build engagement regularly missed. Written material not shared. Tending towards an interrogatory style.

3 Engagement evident but with some problems. Some capsule summaries and major summaries evident, but sporadic in frequency and accuracy. Reflections are utilised.

4 Collaborative approach present, but problems evident. Some sharing of the written material. Clear demonstration of engagement. Both capsule and major summaries are used well. Complex and simple reflections are also present. There is a good level of feedback. Patient involved in the written material. Occasional inconsistent collaboration.

5 As above with regular and very effective use of capsule summaries and major summaries. Correct amount of simple and complex reflections evident. Question:feedback ratio is very well balanced. Patient fully involved in the written material (e.g. adding own written material). Clear collaborative stance.

6 As above, even in the face of patient difficulties.

Interpersonal skills

The low intensity cognitive behavioural practitioner should demonstrate their interpersonal skills in developing and maintaining an effective therapeutic relationship with the patient in the assessment session. The key features of the 'interpersonal skills' item as outlined in the low intensity cognitive behavioural competency scale are as follows:

Key features:

- Empathises through verbal communication
- Non-verbal communication
- Normalising and non-judgmental stance
- Warmth, compassion and rapport
- Pacing

The practitioner should be able to establish a trusting and containing therapeutic relationship with the patient. This should be emphasised through the practitioner's *use of verbal communication*, such as paraphrasing, empathy and clarification.

A competent practitioner should also demonstrate their interpersonal skills in *non-verbal communication skills*, such as maintaining eye contact, smiling when appropriate, using appropriate facial expressions, having an open posture, and considering the seating arrangements. The practitioner should not take notes in a manner that disrupts or inhibits their interpersonal effectiveness.

The practitioner should be able to convey *warmth and compassion* with the patient. This should enable the patient to feel contained and able discuss their problems within the session. The patient's concerns and difficulties should be appropriately *normalised* and not dismissed. The

practitioner should be able to *establish rapport*, building a trusting and warm relationship with the patient to encourage the development of optimism about treatment, as well as motivate the client to want to continue with the treatment process (if indicated).

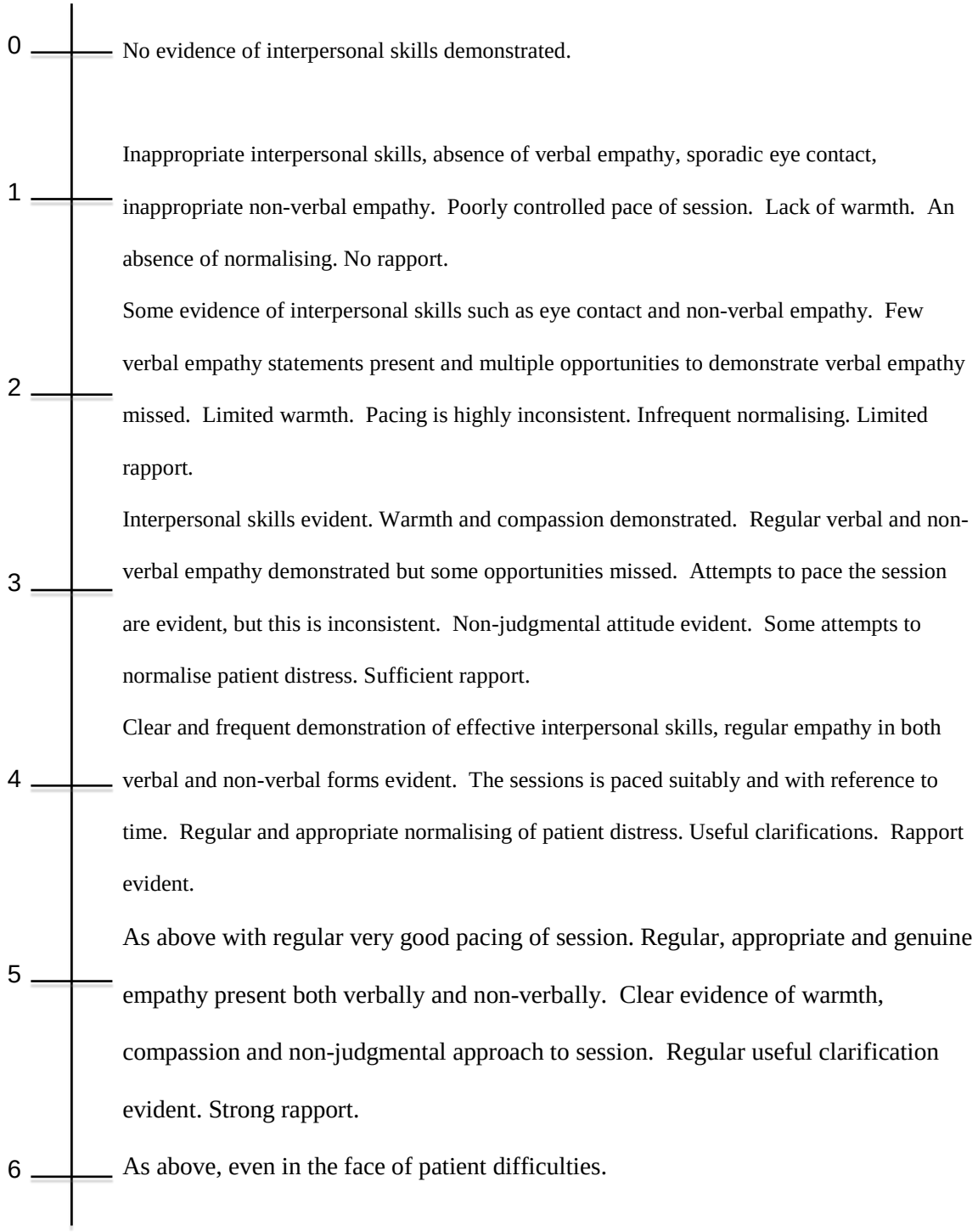
Pacing should be patient-centred to ensure that the patient feels listened to and that they feel their problems are validated. The practitioner should be able to follow the assessment process without the patient feeling unheard or rushed. The session should not be so slow, that the key aspects are not covered.

Checklist-

- Did the practitioner make attempts to develop a therapeutic relationship with the patient?
- Did the practitioner use good body language?
- Did the practitioner demonstrate verbal empathy?
- Did the practitioner demonstrate non-verbal empathy?
- Did the practitioner have an empathetic and warm approach?
- Was there evidence to suggest that the client felt listened to and their problems validated?
- Did the practitioner engender hope via realistic and accurate assurances and explanations?
- Was the patient was given enough time to talk and think?
- Was the practitioner patient-centred and adapted the session to the patient's needs?
- Was the pacing appropriate and flexible?

Interpersonal skills

Competency ratings:



Information Gathering: Problem Focused

The low intensity cognitive behavioural practitioner should demonstrate their competency in gathering information from the patient regarding their problem(s), difficulties and impact of these problems and difficulties are having upon their life. The key features of the 'information gathering' item as outlined in the low intensity cognitive behavioural competency scale are as follows:

Key features:

- Elicits a problem description
- Uses an appropriate questioning style
- Elicits cognitive/behavioural/emotional and physical symptoms of presenting problem
- Elicits onset, triggers for and moderators of the problem
- Determines the impact of the problem on valued activities
- Completes appropriate risk assessment
- Sensitively integrates outcome measures and provides feedback on result
- Recognises of co-morbidity (both psychological and physical)
- Gather information about other relevant issues (e.g. why access help now, past treatments, current medication)

The practitioner should elicit *a problem description* from the patient. The 4 W's; What is the problem? Where does the problem occur? With whom is the problem better or worse? When does the problem happen? Has it happened before? When did it start? Triggers should be elicited to include examples of current situations or stimuli that trigger the problem in the here and now.

The practitioner uses an *appropriate questioning style* to elicit relevant information. A process of funnelling is used to elicit patient centred problem identification by the appropriate use of open questions, specific open questions, closed questions, summarising and clarification.

Following the low intensity model the practitioner should ensure that information is gained in regards to the *behavioural aspects* of the problem, any *physiological symptoms*, the *emotional response*, and *key cognitions*. This will aid in the conceptualisation of the problem as well as enabling patients to recognise and reflect of the different aspects of their difficulties.

The practitioner should gather information about the modifying factors relating to the problem, which includes identifying the maintaining factors.

The practitioner should determine *the impact of problem* on the patient's life and their valued interests and activities.

A full *risk assessment* MUST be undertaken and responded to appropriately. Risk assessment should include identification of intent, presence and nature of suicidal thoughts, hopelessness, thoughts of self-harm, plans, actions past and present, access to means and protective factors. Other risk factors such as alcohol, substance misuse, and risk to/from others should also be gleaned. Self-neglect and neglect of others. Absence of risk assessment leads to an automatic 0 score on this item.

Outcome measures should be sensitively integrated into the assessment. The results should be feedback (use of measure cut-offs) and discussed in an appropriate and compassionate manner.

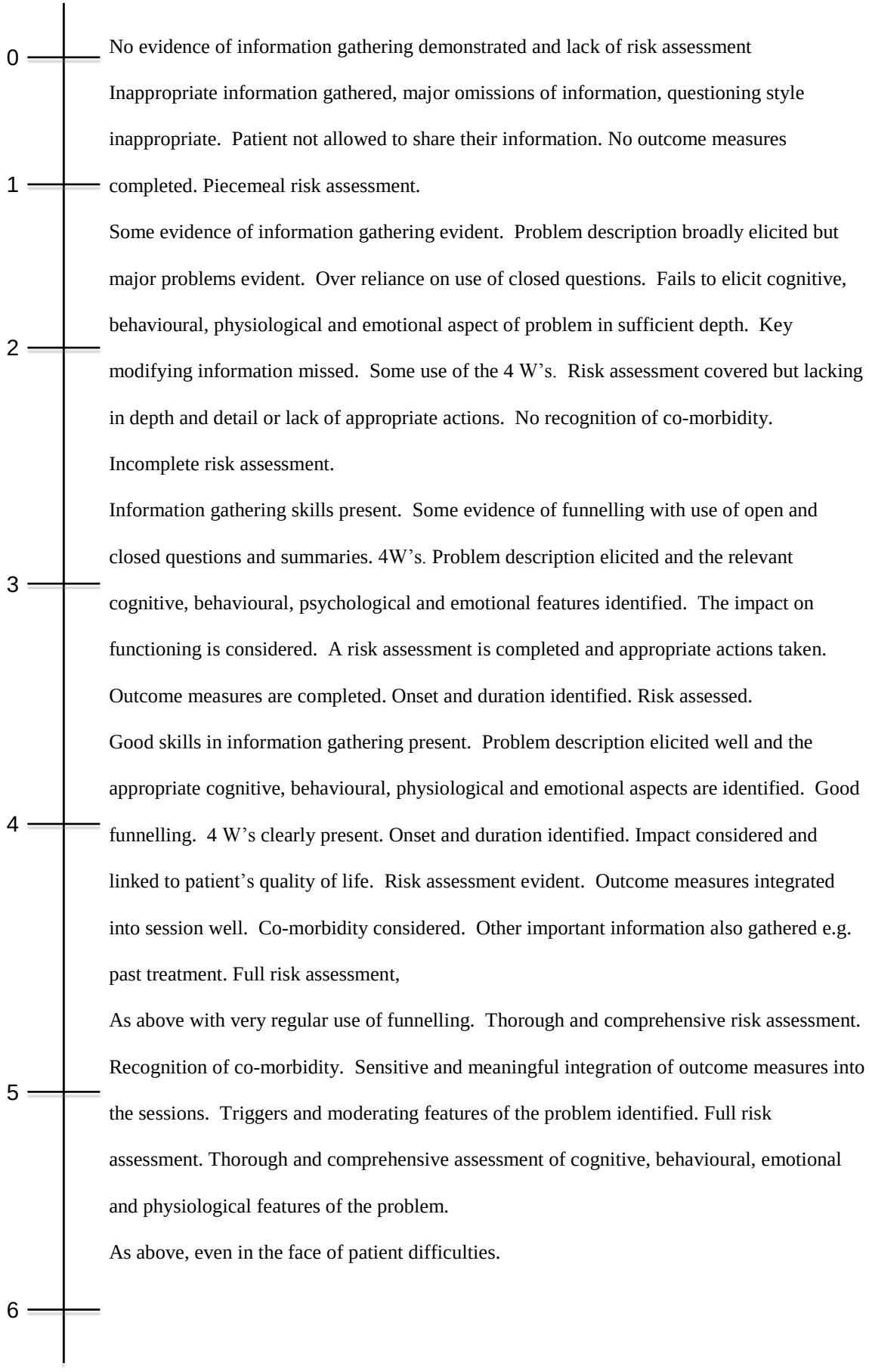
Practitioners should also address *any other issues* that may affect the patient's motivation to engage in guided self-help (e.g. such as past treatment, physical health problems and current medication). The practitioner therefore asks about previous treatments for previous episodes.

Checklist-

- Did the practitioner elicit a problem description from the patient
- Did the practitioner assess the 4 W's of the problem?
- Did the practitioner identify physical symptoms of the problem?
- Did the practitioner identify behavioural aspects of the problem?
- Did the practitioner identify the emotional impact of the problem?
- Did the practitioner identify key cognitions?
- Did the practitioner assess the impact on the patient's valued life activities?
- Did the practitioner elicit the triggers?
- Did the practitioner complete a full risk assessment? And was this dealt with appropriately?
- Was the onset and duration of the problem identified?
- Were modifying factors considered?
- Was information about alcohol and substance misuse elicited?
- Was information gained regarding possible co-morbidity?
- Were outcome measures completed by the patient? And the results discussed?
- Were other relevant issues discussed?

Information Gathering: Problem Focused

Competency ratings:



Information Giving: Focal to the Problem

The low intensity cognitive behavioural practitioner should demonstrate their competency in providing information that is appropriate, focal and suitable to the patient's problem.

The key features of the 'information giving: suitable to the problem' item as outlined in the low intensity cognitive behavioural competency scale are as follows:

Key features:

- Co-creates an accurate ABC or 5-areas conceptualisation
- Co-creates patient centred problem statement

The practitioner should work with the patient to provide a low intensity cognitive behavioural *conceptualisation* of the patient's difficulties using either the ABC or 5-areas technique. The practitioner should attempt to ensure that the patient has a clearer understanding of their difficulties via the conceptualisation.

The patient and practitioner should work together to *create a problem statement*. This will provide a summary of the main features of the problem and a rationale for the treatment method. Much of the problem statement is brought forward from the information gathering and repetition is to be avoided. The problem statement may also provide possible goals for treatment. The problem statement should summarise the triggers, behavioural/cognitive/physiological/emotional aspects of the problem, and should outline the impact of the problem on functioning. The problem statement should be written in the *first person.*

During the assessment session the practitioner should not drift into treatment and should be careful not to provide too much information too early. The practitioner can decide whether it is more useful to complete the problem statement or the conceptualisation first. The practitioner

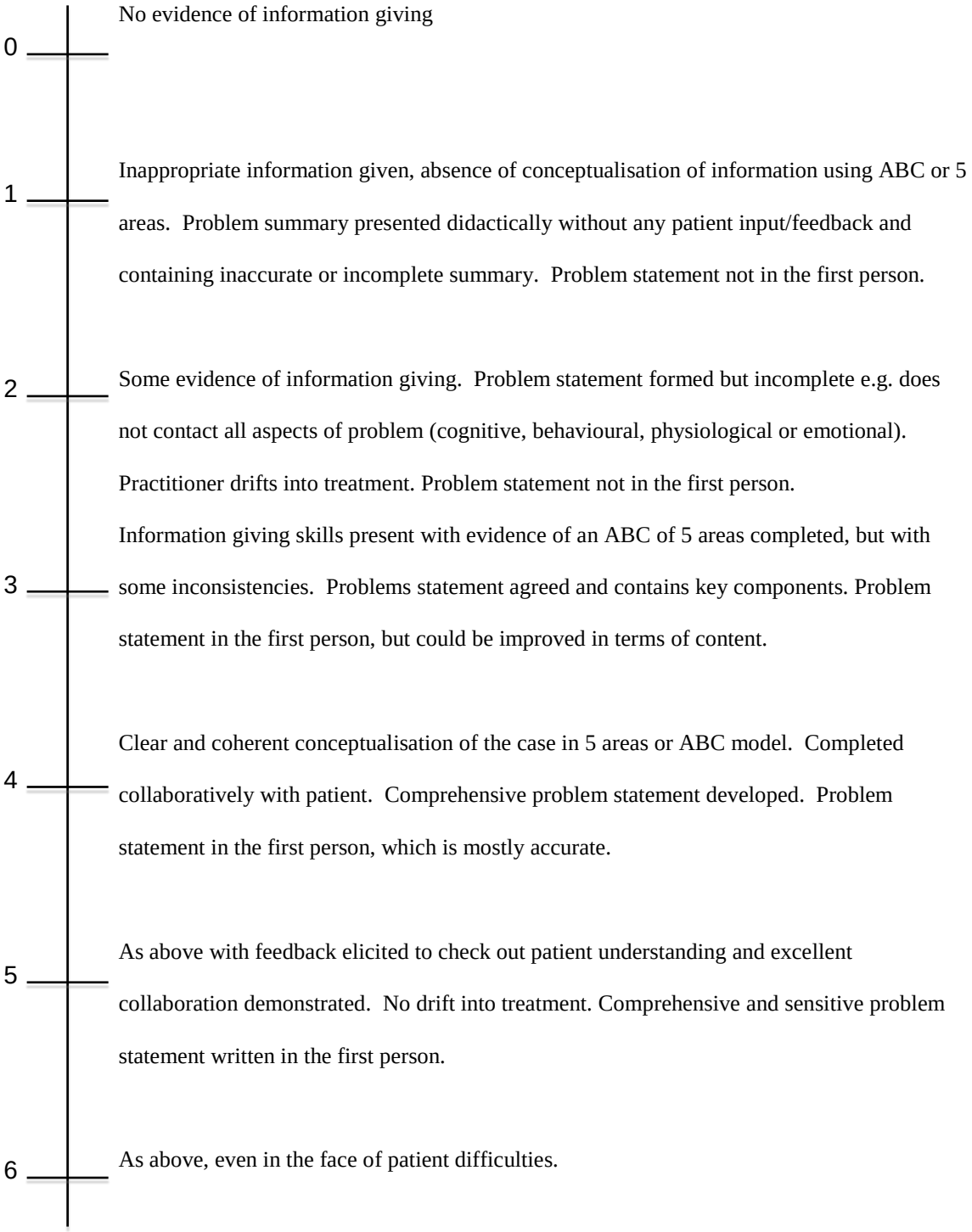
may want to suggest areas that could be worked on within treatment, however the practitioner should focus primarily on giving information linked to the information gathered during assessment and its conceptualisation.

Checklist-

- Did the practitioner conceptualise the problem using an appropriate ABC or 5 areas approach?
- Did the practitioner elicit feedback as to the patient's understanding of the conceptualisation?
- Was the practitioner able to explain the conceptualisation in an accessible way?
- Did the problem statement include triggers, behavioural, cognitive, physiological, and emotional aspects of the problems, alongside the impact on functioning?
- Did the practitioner collaboratively generate a patient-centred problem statement that was succinct and also written in the first person?

Information Giving: Focal to the Problem

Competency ratings:



Shared Planning and Decision Making

The low intensity cognitive behavioural practitioner should demonstrate their competency in identifying suitable treatment options (including signposting), as well as working with the patient to agree plans and actions subsequent to the session (e.g. provide appropriate psycho-education) and also define the goals of the guided self-help.

The key features of the 'shared planning and decision making' item as outlined in the low intensity cognitive behavioural competency scale are as follows:

Key features:

- Suitable treatment options offered
- A rational for treatment provided
- Overall goals for treatment agreed
- Agreed plans and actions subsequent to the session (i.e. between session work)
- Effective ending to the session

The practitioner and the patient should work collaboratively to identify *suitable treatment options* based on the information gathered, the patient's goals and the relevant evidence base. Factors impacting behaviour change as per the COM-B model should be considered. The practitioner should provide information about treatment options and discuss with the patient which would be appropriate and achievable. For example guided self-help interventions such as Behavioural Activation for Depression and medication support, alternative step 2 interventions such as C-CBT, group based interventions such as workshops, step 3 interventions or signposting to other services.

The practitioner should provide a *rationale for treatment* which should involve the consideration of the presenting problem, the patient's goals and the evidence base. The practitioner should not drift into treatment delivery at this point, but should provide an overview of what the patient could expect from their chosen treatment and how this links to information gathered at assessment.

The practitioner should work with the patient to create overall goals for the low intensity intervention. In the assessment session, efforts should be made to make these as SMART as possible. These are not the goals for the next session.

The practitioner should work with the patient to *agree appropriate plans and actions* subsequent to the assessment session. This is the work that is focal to the next session and might involve provision of psycho-educational material, starting to keep a thought diary, or doing some behavioural self-monitoring and so on. The practitioner should consider what adaptations the patient may require to access and engage in this work.

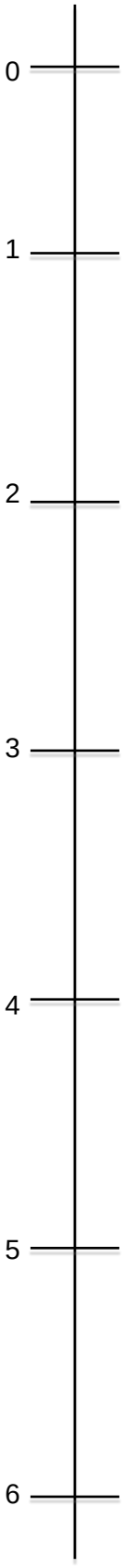
The practitioner should complete the assessment with an appropriate *ending to the session.* The practitioner should ensure the patient has a clear plan and information about appropriate treatment methods. Arrangements should be made regarding an agreement about next step in terms of contact arrangements, appointment etc. The patient should leave the assessment feeling optimistic and confident about the process and confident in attending subsequent sessions. There should be a brief session summary that captures the key aspects of the assessment and outlines the information gathered and decisions made. The practitioner should elicit feedback from the patient about their experience of the session.

Checklist-

- Were treatment options discussed and decided or a plan for when this would take place decided (e.g. after the patient has read about the various treatment options)?
- Did the practitioner create SMART goals for treatment?
  - Was there evidence of shared decision making?
  - Did the practitioner identify suitable treatment options based on the information gathered during the assessment?
  - Was the agreed outcome and planned actions in line with the assessment, patient goals and the low intensity model?
  - Did the practitioner describe the next steps of treatment and outline what the patient should expect?
  - Did the practitioner provide a brief outline of the rationale for the agreed treatment?
  - Did the practitioner and patient agree any the actions subsequent to the session (i.e. the between session work)?
  - Did the practitioner consider the COM-B when making decisions with the patient?
  - Did the practitioner review the session and the patient's experience?
  - Did the practitioner appropriately end the session?
  - Was there a useful session summary?
  - Did the patient leave the session with a clear plan?

Shared Planning and Decision Making

Competency ratings:



0 No evidence of shared decision making or planning. Fails to achieve an agreed outcome to the session. No goals. No actions subsequent to the session. Inappropriate sign-posting.

1 Inappropriate decisions made about treatment. Decisions made unilaterally by the practitioner without any collaboration with patient. Rationale not discussed or outlined. Session ended abruptly. No goals. The actions subsequent to the session are unclear. No use of COM-B.

2 Appropriate outcome and treatment choice identified. Unilateral decision made. Brief and vague rational for treatment choice provided. Vague plans and agreements for treatment established. Session ends without summary. Vague goals discussed. Little specificity to subsequent actions. Some sporadic use of COM-B.

3 Appropriate outcome and treatment chosen. Some evidence of inclusion of patient within decision making process. Rational is either too brief with detail omitted or overly detailed or bordering on treatment. Ending of session evident with vague agreement for next steps. Sufficient evidence of COM-B features e.g. opportunity considered but does not consider motivation or capability. Specific goals agreed.

4 Treatment and outcome to session agreed collaboratively with patient. A concise rationale provided. Agreed actions and plans are clear and feedback elicited from patient to check understanding. Sessions ends well with summary and clear outcome. At least 2 elements of the COM-B model are considered. SMART goals.

5 As above with excellent end of session summary, concise and well informed rationale, collaboration and shared decision making evidence. 3 elements of COM-B are considered, (motivation, capability and opportunity) and this is discussed with regards to consideration of treatment and outcome of session. Actions subsequent to the session are appropriate and helpful. SMART goals.

6 As above, even in the face of patient difficulties.

Appendix C - PWPC- T

Introduction to the Treatment Session Competency						
	INCOMPETENT	NOVICE	ADVANCED BEGINNER	COMPETENT	PROFICIENT	EXPERT
Agrees collaborative agenda with patient ↓						
Subsequent adherence to agenda						
Overall Section Competency Rating	0	1	2	3	4	5 6

Continued Engagement Competencies						
	INCOMPETENT	NOVICE	ADVANCED BEGINNER	COMPETENT	PROFICIENT	EXPERT
Collaborative approach concerning change						
Acknowledges progress or difficulties by use of simple reflections						
Acknowledges progress or difficulties by use of complex reflections						
Use of capsule summaries regarding progress or difficulties						
Use of major section summaries						
Ratio of questions to feedback to facilitate change						
Overall Section Competency Rating	0	1	2	3	4	5 6

Interpersonal Competencies						
	INCOMPETENT	NOVICE	ADVANCED BEGINNER	COMPETENT	PROFICIENT	EXPERT
Empathises through verbal communication						
Non-verbal behaviour						
Non-judgemental stance						
Warmth and compassion						
Pacing						
Overall Section Competency Rating	0	1	2	3	4	5 6

Information Gathering Competencies Specific to Change							
COM-B		INCOMPETENT	NOVICE	ADVANCED BEGINNER	COMPETENT	PROFICIENT	EXPERT
CAPABILITY OPPORTUNITY MOTIVATION	Questioning skills						
CAPABILITY OPPORTUNITY MOTIVATION	Problem statement review						
CAPABILITY OPPORTUNITY MOTIVATION	Review of goal progress						
CAPABILITY OPPORTUNITY MOTIVATION	Homework review						

	Review of medication (if appropriate)						
CAPABILITY MOTIVATION OPPORTUNITY	**Risk Assessment**						
	Outcome monitoring						
Overall Section Competency Rating		**0**	**1**	**2**	**3**	**4**	**5**	**6**

Within Session Self-Help Change Method Competencies							
COM -B		INCOMPETENT	NOVICE	ADVANCED BEGINNER	COMPETENT	PROFICIENT	EXPERT
CAPABILITY OPPORTUNITY MOTIVATION	**Adherence to principles of PWP intervention** Displays fidelity to low intensity treatment (eg. providing educational materials for use in session) ↓						
CAPABILITY MOTIVATION OPPORTUNITY	**Appropriateness of PWP intervention** Relevant PWP intervention appropriate for patient, stage of intervention and presenting problem ↓						
	Use of written material (eg. diaries, ABC or 5-areas conceptualisation to drive the low intensity change methods)						
Overall Section Competency Rating		**0**	**1**	**2**	**3**	**4**	**5**	**6**

Planning and Shared Decision Making Competencies							
COM -B		INCOMPETENT	NOVICE	ADVANCED BEGINNER	COMPETENT	PROFICIENT	EXPERT
CAPABILITY OPPORTUNITY MOTIVATION	Agrees next steps of treatment and the between session work ↓						
CAPABILITY MOTIVATION OPPORTUNITY	Agrees the implementation plan for the between session work ↓						
	Session review and ending						
Overall Section Competency Rating		0	1	2	3	4	5 6

Appendix D- Information sheet

Information sheet

Research Project Title:

Competency of assessment and treatment during low intensity cognitive-behaviour therapy: A validation study

You are being invited to participate in a research project. Before you decide it is important for you to understand why the research is being done and what it will involve. Please take the time to read the following information carefully and discuss with others if you wish. Ask us if there is anything that is not clear or if you would like more information.

What is the research study?

Psychological wellbeing practitioners (PWPs) use low intensity cognitive behavioural interventions to treat people with mental health concerns. We would like to test a scale which measures the level of competency shown by PWPs in assessment and treatment sessions. This research will study whether the *low intensity cognitive behavioural competency scales* are valid, reliable and have good internal consistency.

Measuring practitioner competencies in delivering assessment and treatment with clients is very important. Firstly it will provide information to trainers, supervisors, PWPs and trainees that will allow them to develop their skills. Also by ensuring that PWP have a high level of competence we will be able to assure that patients are receiving a quality and safe provision of care.

How will the scale be tested?

The research will involve a number of phases. Firstly we will ask an expert panel to review the items to ensure that we are measuring the appropriate competencies. Then we will ask PWP trainers and qualified PWPs to rate a pre-recorded assessment and treatment session. Using their data we will test the inter-rater reliability to see whether they show similar ratings scores.

In addition we will also be asking PWP trainees to be involved in the research by collecting the ratings from their OSCEs and practice sessions (using the competency scales) and comparing these results to measure the test-retest reliability. Actors involved in the OSCEs will be asked to complete questionnaires about how they felt during the session. This will allow us to see if the practitioners who were viewed by the actors as being the most helpful were also rated highly on the competency scales.

Who will be asked to be involved in this research?

We will be requesting the involvement of:

- PWP trainers (attending the North/South PWP trainers conferences)
- Qualified PWPs (attending the Yorkshire and Humberside PWP conference)
- PWP trainees (at University of Sheffield)
- Actors (involved with trainees OSCEs at University of Sheffield)

Do I have to take part?

Participation in this research is voluntary. If you decide to take part you will be given this information sheet to keep (and will be requested to fill in a consent form). You can withdraw your rating and/or responses at any time without it being viewed negatively. For PWP trainees,

withdrawal will not affect your grades or be detrimental to your place on the course. For actors, withdrawal will not affect your payment or relationship with the University of Sheffield.

What will I have to do?

The expert panel will be asked to rate the relevance of each item on the *low intensity cognitive behavioural competency scale.*

The PWP trainers and qualified PWPs will be asked to view a pre-recorded (OSCE) assessment and treatment session. They will be asked to complete ratings on practitioner's level of competence using the scales.

The PWP trainees will complete their practice sessions and OSCEs during their course. The ratings from the course staff will be collected (recorded sessions will only be used in the ratings by the university and will not be passed on to the research team).

The actors involved in the OSCE will be asked to complete 2 short questionnaires after each session with a trainee.

Will the data collected by confidential?

All the data collected will remain confidential. You will not be identified or identifiable within any reports or publications. You name will be replaced by a participant Identification number during the research.

Ethical consent was obtained for this study from Sheffield University Ethics Committee.

Thank you for participating in this research**.**

Appendix E- consent form

Competency of assessment and treatment during low intensity cognitive-behaviour therapy: A validation study

Lucy Hughes

Participant Id number for this project:

Please initial box

1. I confirm I have read and understand the information sheet dated August 2015 explaining the above research project and I had the opportunity to ask questions about the project.

2. I understand that my participation is voluntary and that I am free to withdraw my data at any time without giving any reason and without there being any negative consequences. Please contact Lucy Hughes (pcp12la@shef.ac.uk).

3. I give permission for members of the research team to have access to my anonymised responses. I understand that my name will not be linked with the research materials, and I will not be identified or identifiable in the report that result in the research.

4. I agree that the data collected from me to used in future research.

5. I agree to take part in the above research.

Appendix F - Ethical Approval

From: s.kellett@sheffield.ac.uk

Ethics approval has been accepted. See below

-------- Original Message --------

Subject: Ethics Application 006168

Date: Sun, 16 Aug 2015 15:53:25 +0100

From: R&IS <no-reply@sheffield.ac.uk>

Reply-To: t.webb@sheffield.ac.uk

To: s.kellett@sheffield.ac.uk

This is a notification from the online ethics application system.

Your application (006168) has been returned to you and can now be viewed.

You can log in to the system to view and take action on this application here

http://ethics.ris.shef.ac.uk/

Best wishes

R&IS

Appendix G - WAI

NAME _____ PWP Trainee _____

On the following pages there are sentences that describe some of the different ways a person might think or feel about his or her PWP.

Work fast, your first impressions are the ones we would like to see. (PLEASE DON'T FORGET TO RESPOND TO EVERY ITEM.)

Thank you for your cooperation.

	Never	Rarely	Occasionally	Sometimes	Often	Very Often	Always
I felt uncomfortable with the PWP							
The PWP and I agreed about the things I will need to do in therapy to help improve my situation.							
I am worried about the outcome of future sessions.							
What I did in session gave me a new way of looking at my problem.							
The PWP and I understood each other.							

	Never	Rarely	Occasionally	Sometimes	Often	Very Often	Always
The PWP perceived accurately what my goals were.							
I found what I did in the session confusing.							
I believed the PWP liked me.							
I wish the PWP and I could have clarified the purpose of our session.							
I disagreed with the PWP about what I ought to get out of therapy.							

	Never	Rarely	Occasionally	Sometimes	Often	Very Often	Always
I believe the time the PWP and I spent together was not spent efficiently.							
The PWP did not understand what I was trying to accomplish from therapy.							
I am clear on what my responsibilities will be in therapy.							
The goals of this session are important for me.							

	Never	Rarely	Occasionally	Sometimes	Often	Very Often	Always
I found what the PWP and I were doing in therapy is unrelated to my concerns.							
I felt that the things I did in therapy will help me to accomplish the changes that I want.							
I believed the PWP is genuinely concerned for my welfare.							
I am clear as to what the PWP wanted me to do in this session.							

	Never	Rarely	Occasionally	Sometimes	Often	Very Often	Always
The PWP and I respected each other.							
I felt that the PWP was not totally honest about his/her feelings toward me.							
I am confident in the PWP's ability to help me.							
The PWP and I were working towards mutually agreed upon goals.							
I felt that the PWP appreciates me.							

	Never	Rarely	Occasionally	Sometimes	Often	Very Often	Always
We agreed on what was important for me to work on.							
As a result of this session I am clearer as to how I might be able to change.							
The PWP and I trusted one another.							
The PWP and I had different ideas on what my problems were.							
The PWP and I collaborated on setting							

	Never	Rarely	Occasionally	Sometimes	Often	Very Often	Always
goals for my therapy.							
I was frustrated by the things I was doing in therapy.							
We established a good understanding of the kind of changes that would be good for me.							
The things that the PWP asked me to do didn't make sense.							
I don't know what to							

	Never	Rarely	Occasionally	Sometimes	Often	Very Often	Always
expect as the result of my therapy.							
I believe the way we worked with my problem was correct.							
I felt that the PWP cares about me even when I did things that he/she did not approve of.							

Appendix H - HATs

Of the events which occurred in this session, which one do you feel was the most helpful or important for you personally? (By "event" we mean something that happened in the session. It might be something you said or did, or something your PWP said or did.)
Please describe what made this event helpful/important and what you got out of it.

How helpful was this particular event? Rate it on the following scale. (Put an "X" at the appropriate point)
HINDERANCE ——————————————————- Neutral ——————————————————- HELPFUL 1 2 3 4 5 6 7 8 9
Did anything happen during the session which might have been hindering?
YES / NO
If yes, please rate how much of a hindrance was this event was: HINDERANCE ——————————————————- Neutral ——————————————————- HELPFUL 1 2 3 4 5 6 7 8 9

Please describe the event briefly:
How likely are you to recommend this PWP to friends and family if they needed similar care or treatment?
1 Extremely unlikely 2 Unlikely 3 Neither likely or unlikely 4 Likely 5 Extremely Likely 6 Don't know
Would you come and see this PWP again?
YES / NO