# Development and psychometric evaluation of the Assessment of Core CBT Skills (ACCS): An observation based tool for assessing Cognitive Behavioural Therapy competence

Kate Muse, Freda McManus, Sarah Rakovshik, Richard Thwaites

## Abstract

This paper outlines the development and psychometric evaluation of the Assessment of Core CBT Skills (ACCS) rating scale. The ACCS aims to provide a novel assessment framework to deliver formative and summative feedback regarding therapists' performance within observed cognitive-behavioural treatment sessions, and for therapists to rate and reflect on their own performance. Findings from three studies are outlined: 1) a feedback study ($N$ = 66) examining content validity, face validity and usability, 2) a focus group ($N$ = 9) evaluating usability and utility, and 3) an evaluation of the psychometric properties of the ACCS in 'real world' CBT training and routine clinical practice contexts. Results suggest that the ACCS has good face validity, content validity, and usability and provides a user-friendly tool that is useful for promoting self-reflection and providing formative feedback. Scores on both the self and assessor-rated versions of the ACCS demonstrate good internal consistency, inter-rater reliability, and discriminant validity. In addition, ACCS scores were found to be correlated with, but distinct from the Revised Cognitive Therapy Scale (CTS-R) and were comparable to CTS-R scores in terms of internal consistency and discriminant validity. Additionally, the ACCS may have advantages over the CTS-R in terms of inter-rater reliability of scores. The studies also provided insight into areas for refinement and a number of modifications were undertaken to improve the scale. In summary, the ACCS is an appropriate and useful measure of CBT competence that can be used to promote self-reflection and provide therapists with formative and summative feedback.

Key words: competence, skill, assessment, training, cognitive-behavioural, CBT.

Competence in delivering psychological treatments can be defined as the degree to which a therapist demonstrates the general therapeutic and treatment-specific knowledge and skills required to appropriately deliver evidence-based interventions (Barber, Sharpless, Klostermann, & McCarthy, 2007; Kaslow, 2004). Within the context of Cognitive Behavioural Therapies (CBT), Roth and Pilling (2007) identify five key domains of competence required to deliver effective treatment. One domain reflects generic therapeutic competences, such as knowledge of mental health and patient engagement. The other four domains relate to CBT-specific competences, including basic CBT competences such as knowledge of cognitive-behavioural principles, the use of specific CBT techniques, problem-specific competences (aka protocol or disorder-specific interventions), and metacompetences such as the ability to select and apply appropriate CBT methods. Tools for measuring competence in delivering CBT provide a means of assessing the training of new CBT therapists and ensuring the quality of treatment provision within routine practice, provide a framework for delivering formative feedback, promote ongoing self-reflection, and are essential to establishing treatment integrity in research trials (Dobson & Singer, 2005; Laireiter & Willutzki, 2003; McHugh & Barlow, 2010; Weck, Bohn, Ginzburg, & Ulrich, 2011). As such, it is imperative that therapists, assessors, and researchers alike have access to valid, reliable, and usable measures for assessing CBT competence.

A recent review identified ten key methods for assessing CBT competence (Muse & McManus, 2013). It is argued that each method focusses on different aspects of Miller's (1990) hierarchical framework for assessing clinical skill, ranging from therapists' knowledge of CBT ('knows'), their practical understanding ('knows how'), their skill within artificial clinical simulations ('shows how'), and their skill within real clinical practice settings ('does'). Therapists' skill within real clinical practice settings is potentially the most complex aspect of CBT competence to operationalise and assess. However, in order to confidently conclude that a therapist is competent, it is important to establish that they can appropriately and effectively apply their generic and treatment-specific knowledge and skills within the cultural and organisational context of clinical practice settings (Miller, 1990; Roth & Pilling, 2007). Indeed, this aspect of clinical skill is viewed by experts in the field as being at the heart of delivering competent CBT (Muse & McManus, 2015). To date, the 'gold standard' for assessing therapists' skill within practice has been ratings of therapists' in session performance using standardised rating scales

which outline and behaviourally operationalise the skills involved in the competent delivery of CBT. However, there is a need for further refinement of the observation-based scales that are currently available (Fairburn & Cooper, 2011; Muse & McManus, 2013; Muse & McManus, 2015). In particular, there is a need for more comprehensive and up to date rating scales with improved validity, reliability, and usability that can be used for both formative and summative purposes. Thus, the current study focuses on developing an observation-based tool for assessing whether therapists can demonstrate the skills necessary to effectively deliver CBT within a treatment session. A copy of the ACCS rating scale, manual, and submission cover sheet is available from www.*removed for anonymity*.

The most prominent existing tools for assessing therapists' in session performance are the Cognitive Therapy Scale (CTS, or Cognitive Therapy Rating Scale: CTRS, www.beckinstitute.org) and the revised version of the CTS (CTS-R: Blackburn et al., 2001). Although widely used, the CTS and CTS-R have been criticised for lacking capacity for formative feedback, poor definitional clarity, unclear rating guidelines that lack depth, unnecessary item overlap, multiple concepts addressed by single items, lack of applicability across Axis 1 disorders, lack of applicability across a range of both cognitive and behaviourally focused therapies, and failure to account for recent advances in CBT (see Muse & McManus, 2014 for a recent review). The Assessment of Core CBT Skills (ACCS) aims to address these limitations by: breaking down broad aspects of CBT competence into discrete components, providing clearer behavioural anchors for scale points, reducing the degree of ambiguity and assessor inference required, updating the content of the scale in light of recent advancements in CBT practice, including additional aspects of CBT competence, increasing capacity for formative feedback, and incorporating the use of supporting materials. Hence the ACCS builds upon these existing scales to provide an assessment framework for delivering formative and summative feedback on therapists' performance within observed CBT treatment sessions, and for therapists to rate and reflect on their own performance.

The ACCS aims to assess core general therapeutic and CBT-specific skills required to competently deliver CBT interventions that reflect the current evidence-base for treatment of the patient's presenting problem (i.e. 'limited-domain intervention competence': Barber et al., 2007; Kaslow, 2004). As illustrated in Figure 1, the ACCS features 22 items, organised thematically into eight competence domains. Following a

deductive approach (Burisch, 1984), a review of relevant literature (Muse & McManus, 2013) was used to guide the development of scale items. In particular, the authors drew upon the CTS (www.beckinstitute.org), the CTS-R (Blackburn et al., 2001), Roth and Pilling's (2007) competence framework, and relevant CBT treatment manuals and protocols. Items were included because relevant theory or research indicated that the skill is an important aspect of CBT competence.

*Insert Figure 1 about here*

The skills assessed within the ACCS are transdiagnostic (i.e. focus on competences which are not specific to any one diagnosis or protocol) and relate to therapists' performance within active treatment sessions. It could be argued that the ideal method of assessing competence is to use rating scales that are specific to a particular treatment protocol and address all of the disorder-specific skills evident across each stage of treatment (e.g. video feedback in social phobia, reliving in PTSD, goal setting, relapse prevention etc.). This approach would require a different competence measure for each treatment protocol as well as the inclusion of a vast range of items, many of which would not be applicable to the majority of sessions being rated. Given the proliferation of different diagnosis specific treatment manuals, this approach would undermine the feasibility of this method of assessment, increase the complexity of rating competence, and make it difficult to draw comparisons across therapists (Farchione et al., 2012). This would be especially problematic in training and practice settings where clinicians deliver a variety of CBT protocols and work with patients experiencing a wide range of mental health problems and high rates of co-morbidity (Barber et al., 2007). It was, therefore, decided to focus on skills which are evident in active treatment sessions and are relevant across different treatment groups and protocols.

All items are rated on a four-point scale measuring clinical skill (1- *limited* 2 – *basic*, 3- *good*, and 4- *advanced*). As respondents rarely endorse negative scale points (Schwarz, Knauper, Hippler, Noelle-Neumann, & Clark, 1991), only values above zero were used. The optimal length of a rating scale is between four and seven points as this allows for sufficient reliability, variability, sensitivity, and usability (Krosnick & Fabrigar, 1997). Thus four response options were used to allow adequate discrimination between levels of competence without making the scale unwieldy. Given that some respondents will choose a neutral response in order to avoid making a choice (Van Vaerenbergh & Thomas, 2012) and that the purpose of this scale is to determine whether a therapist can

demonstrate competence or not, an even number of response options was used to force respondents to make a commitment in the direction of competence or incompetence. Both a total score (range 22 to 88) and an average item score is provided. As little is known about whether some CBT skills are of more importance than others, equal weight is given to each item.

The accompanying ACCS manual provides guidance for assessors in making judgements about the skilfulness of therapists' performance. Generic anchors are provided for each scale point, which is used to provide an overarching framework for scale ratings (see Figure 2). Item-specific 'exemplar therapist behaviours' also provide examples of the type of performance consistent with each scale point (see Figure 2 for an example). This approach was used because respondents are more satisfied when all scale points are labelled (Wallsten, Budescu, Zwick, & Kemp, 1993) and using behaviourally anchored scale points improves inter-rater agreement, reduces the halo effect, and improves measurement validity (Krosnick & Fabrigar, 1997). The ACCS manual also specifies implementation guidelines, recommending that ACCS ratings are completed on the basis of viewing a recording of a full CBT treatment session in combination with key contextual information (e.g. stage of therapy, patient's presenting problem, formulation etc.) provided by therapists in the ACCS submission cover sheet.

*Insert Figure 2 about here*

The ACCS is designed to be a developmental tool and therefore provides space for in-depth narrative feedback in addition to numerical ratings. Assessors can draw on the exemplar behaviours provided as part of the scale and the specific session material to give examples of strengths and areas for improvement, as well as highlighting strategies for further development. Such in-depth formative feedback plays an integral role in the ongoing development of competent, reflective practitioners and is well received by those being assessed (Govaerts, van der Vleuten, Schuwirth, & Muijtjens, 2005; Milne, 2007; Van der Vleuten et al., 2010).

This paper presents findings from three studies examining the ACCS scale. All three studies received ethical approval and were funded by a grant from the British Association of Behavioural and Cognitive Psychotherapies. Study 1 presents a large-scale feedback study, which involved collecting formal feedback about the ACCS from both expert and novice CBT therapists. This feedback was used to examine content validity,

face validity, and perceived usability. Study 2 provided a more in-depth insight into how useful and user-friendly the ACCS is in practice by conducting a focus group to examine assessors' experiences of using the ACCS. Finally, Study 3 involved investigating the psychometric properties of the ACCS in 'real world' CBT training and routine practice contexts in order to evaluate the reliability and validity of scores on the assessor-rated and self-rated ACCS scale. Overall, it is hoped that findings from these three studies will help to determine whether the ACCS is suitable for use in clinical practice and training settings.

## Study 1: Feedback Study Examining Content Validity, Face Validity and Perceived Usability of the ACCS

Review from subject matter experts is an essential ingredient in improving the quality of rating scales during the developmental phase (Brewer & Hunter, 2005), and it is also useful to gain feedback from the target population to better understand how they comprehend and respond to items (Campanelli, Martin, & Rothgeb, 1991). Hence, Study 1 collected feedback about the ACCS from experts within the field of CBT, with experience of assessing competence, and from relative novices with limited CBT experience, who are likely to receive feedback on the ACCS and use the tool to rate their own competence. The primary aim was to examine face validity (i.e. appropriateness, credibility and plausibility of items as measures of CBT competence), content validity (i.e. adequate representation of CBT competence), and perceived usability. Participants' feedback was also used to identify areas where the ACCS required refinement.

### Method

#### Participants

The study recruited two groups of participants: expert and novice CBT therapists. Experts were broadly defined as individuals with significant experience in the provision of CBT interventions and involvement in evaluating the competence of CBT therapists. Experts were identified through professional involvement in the training, selection, or evaluation of CBT therapists' and/or publication of research in the domain. Novice participants were broadly defined as individuals who were new to, and inexperienced in delivering CBT (e.g. trainees, recently qualified CBT practitioners). Novices were identified through current or recent involvement in training courses that included a

significant CBT training component (e.g. clinical psychology doctorate courses, post graduate diplomas in CBT). Snowball sampling, whereby participants were asked to forward the information about the study, was also used to reduce researcher bias. Due to the recruitment strategy, it is not known how many therapists were given study information. Forty-one experts and 25 novices completed the questionnaire (see Table 1 for demographic characteristics).

*Insert Table 1 about here*

## Materials

**Face and content validity questions.** Participants rated the items' relevance (1- *not relevant* to 4- *very relevant*) and clarity (1- *not clear* to 4- *very clear*). A content validity index (CVI: Yagmale, 2003) was calculated for each domain by identifying the percentage of experts who rated the item as being both relevant and clear (i.e. a rating of ≥ three). Participants were asked whether any important aspects of CBT competence were omitted (i.e. any key competences the scale neglected) and, if so, what these were. Participants were asked to identify domains that inappropriately overlapped (i.e. measured the same construct). Finally, a yes/no response was used to indicate any items inappropriately assessing multiple aspects of CBT competence (rather than specific and discrete constructs).

**Usability questions.** Participants rated how easy they thought the scale would be to use (1- *not easy* to 4- *very easy*), the overall style, appearance and layout of the scale (1- *poor* to 4- *very good*), and how appropriate they found the scoring system (1- *not appropriate* to 4- *very appropriate*). Participants also indicated whether they felt the scale provided adequate opportunity for in-depth feedback using a yes/no response. If participants circled no, they were asked to indicate what they felt was missing.

**Qualitative feedback.** Where participants provided a rating of three or below, for the relevance or clarity of the domain, the appropriateness of the scoring system, style appearance and layout, or ease of use they were asked to indicate potential improvements. Participants were also asked whether they had any other comments or suggestions for improvements. Recurrent patterns were identified using thematic analysis (Braun & Clarke, 2006). Initial codes were generated by summarising the key issues highlighted in each comment. Codes with similar meanings were then combined to create overarching

themes. Analysis was carried out by the first and second author (XX and XX), with coherence and replicability being checked by an independent researcher.

## Results

### Face and Content Validity

Content validity scores for each ACCS domain are presented in Table 2. Both novices and experts found all domains at least 'quite' relevant and clear, with no significant differences between the scores assigned by novices and experts. The content validity index (i.e. the percentage of participants who rated the domain as ≥ three for both relevance and clarity) was above the suggested threshold of 70% (Lynn, 1986) for all domains. No items were identified as assessing multiple concepts or as overlapping with other items by the majority (>50%) of participants. For the agenda setting domain, over 30% of total participants indicated that items inappropriately assessed multiple aspects of CBT competence. Nineteen participants (28.79 % of the total sample) indicated that they felt guided discovery / Socratic method was missing.

*Insert Table 2 about here*

### Usability

All participants rated the scale as at least 'quite' easy to use, with at least 'good' style, appearance and layout, and at least a 'quite' appropriate scoring system. Mann-Whitney tests revealed no significant differences between novices' and experts' scores for style, appearance and layout (novices $M$ = 3.88, $SD$ = 0.33 vs. experts $M$ = 3.68, $SD$ = .52; $U$ = 422.50, $p$ = .10) or appropriateness of the scoring system (novices $M$ = 3.80, $SD$ = .41 vs. experts $M$ = 3.51, $SD$ = .71; $U$ = 407.50, $p$ = .08). However, novices assigned a significantly higher rating for ease of use compared to experts (novices $M$ = 3.56, $SD$ = .65 vs. experts $M$ = 3.20, $SD$ = .71; $U$ = 366.50, $p$ = .03). All novice participants and 87.80 % of experts ($n$ = 36) felt the scale provided ample opportunity for feedback.

### Qualitative Feedback

Four key areas of strength were identified. First, participants felt the ACCS was a clear and comprehensive rating scale, commenting that that the ACCS was "very clear and useful", "extremely comprehensive" and "very thorough". Second, participants liked the intuitive and user-friendly style of the ACCS, which made it seem "very easy to use". In

particular, participants highlighted the layout, the organisation of items into different domains, the use of colour-coded icons, and the inclusion of general and item-specific guidelines and exemplar behaviours. As one participant noted, these features made the ACCS "much easier to make sense of quickly in comparison to other scales". The third strength reflected the useful developmental functions of the ACCS, both in terms of facilitating self-reflection and as a tool for providing in-depth formative feedback. This theme can be summarised by the following quotation "full of opportunity to on the one hand provide constructive feedback, while on the other to provide a standard to work towards and better oneself by". Finally, the fourth strength identified was the ACCS's increased specificity and coherence, the separation of skills into discrete sections, and the inclusion of core CBT skills that have not previously been captured. These strengths resulted in the view that the ACCS is "a useful addition to our box of tools in supervision".

Participants also identified some limitations and offered suggestions for overcoming these. Participants suggested adding "missing elements" such as patient difficulty, skilfulness of delivery of interventions, guided discovery, collaboration, and more behavioural aspects of CBT in the descriptors. Participants also suggested improving clarity and usability by providing additional information within the rating guidelines, re-phrasing terminology, re-structuring the scale, allowing more opportunity for formative feedback, and making the scale anchors more concise. Finally, some participants questioned whether the ACCS would be applicable to all disorders and protocols and others noted that there was some "inevitable" overlap between items and domains due to the complex nature of CBT competence.

The scale was refined in the light of participants' feedback. First, changes were made to the scoring system i.e., adding space for formative feedback within each domain, reducing the five-point scale to a four-point scale, using more positive banding titles, and using an average item scoring system in addition to a total sum scoring system. Second, changes were made to improve usability. This included re-phrasing and clarification of anchor descriptions, reducing the length of anchor descriptions, reducing ambiguity and increasing behavioural specificity of anchor descriptions, including additional rating guidance, adding a submission coversheet to be submitted with session recordings, and updating the order in which domains appeared in the scale. Finally, changes were made to the specific content of items, including focusing more explicitly on behavioural elements, including collaboration as a separate item, providing further clarification and guidance for

the measuring change domain, further emphasising guided discovery within scale items, re-naming and re-structuring the conceptualisation domain, expanding the CBT interventions domain, and re-structuring and extending the homework domain.

## Discussion

Feedback from expert and novice CBT therapists was elicited to examine the usability, face validity, and content validity of the ACCS and identify areas for improvement. The majority of novice and expert participants found the domains in the scale both relevant and clear and only a very small percentage of participants indicated that items in the scale assessed multiple concepts or overlapped with other items. Qualitative feedback about the ACCS was generally very positive, with participants finding the ACCS to be a comprehensive, clear, and user-friendly tool that would be helpful for promoting self-reflection and providing formative feedback. Both experts and novices felt the scale would be easy to use, was visually appealing (i.e. had good style, appearance and layout), and had an appropriate scoring system. Taken together, these results suggest that the ACCS has good face validity, content validity and perceived usability. Results from this study were also used to improve the clarity and usability of the scale, enhance capacity for formative feedback, and to address missing elements of skill.

## Study 2: An In-depth Focus Group Evaluating Usability and Utility of the ACCS

Study 2 utilised a focus group to obtain in-depth assessor feedback on the usability of the ACCS scale, with the intention of identifying what did and did not work well in practice, as well as areas where the ACCS required further refinement.

## Method

### Participants

Nine individuals who assessed therapists using the ACCS within the 2013/14 intake of the Postgraduate Diploma (PGDip) in Cognitive Behavioural Therapy course run by the Oxford Cognitive Therapy Centre (OCTC) participated (for a description of the course see McManus, Westbrook, Vazquez-Montes, Fennell, & Kennerley, 2010). Participants were all BABCP accredited CBT therapists who had been practicing CBT for

between 13 and 30 years ($M$ = 20.22, $SD$ = 6.24). Four assessors were clinical psychologists, three were nurses, one was a psychiatrist, and one was a counsellor.

### Data Collection

A focus group was used to obtain assessors' feedback on using the ACCS. A semi-structured interview schedule consisting of open-ended questions and minimal prompts was used to guide the discussion (Kvale, 1996). Within the schedule, emphasis was placed on reflection of personal experience in relation to the scale in general (e.g. "What has been your experience of using the ACCS? How have you found it?"), and more specifically in relation to clarity and relevance of the items, appropriateness of the scoring system, and usability (e.g. "How easy or difficult was it to use the ACCS?"). Where problems or difficulties arose, participants were asked whether the issue could be resolved and, if so, how.

### Data Analysis

Qualitative analysis comprised of the 'framework technique' (Ritchie & Spencer, 2002), chosen because it provides a simple framework for describing the key advantages, disadvantages, and areas for improvement commonly highlighted by participants. Emergent themes were used to identify an initial thematic framework, which was then systematically applied to the data. Following this the content of the recording notes was distilled into a summary and entered into a chart of key themes. Finally, a 'map' of key themes was created by aggregating patterns of data, weighing up the importance and dynamics of issues, searching for an overall structure in the data, and synthesising the findings. The primary author (XX) took the lead in analysis and validation was conducted by an independent third party with no involvement in the development of the ACCS.

## Results

Results of the focus group are structured within two overarching themes: 1) key strengths and 2) areas for improvement[1].

[1] Direct participant quotations are not provided to support the key themes identified in the text. This is because participants did not provide consent for direct quotations to be used for research purposes.

**Key Strengths**

Eight key strengths were identified: 1) relevance, 2) clarity, 3) simplicity, 4) detail and specificity, 5) well-operationalised, 6) layout and style, 7) formative function, and 8) usability. Participants felt the ACCS was highly relevant to CBT competence, had a clear instruction manual, was attractive in appearance, and clearly and simply defined and operationalised CBT competence. They also liked the detail and specificity of the ACCS, i.e. the way the ACCS broke down the core competences required to deliver CBT into smaller components. Participants felt that the ACCS was easy to use and commented that the increased detail and specificity offered a useful template for providing more in-depth feedback to therapists and thus could serve as a useful 'good practice guide' for helping therapists to understand and remember the essential CBT skills. Participants did not feel that this increased specificity meant that the ACCS was too lengthy or took too long to complete, especially once they had become familiar with the scale and had practiced using it.

**Areas for improvement**

Four topics were highlighted as areas for improvement. First, participants felt there was a need for a 'clearer and more detailed feedback form'. A number of assessors completed ACCS ratings without referring back to the manual, leading them to complete ratings on the basis of banding headings alone (i.e. 1- limited, 2- basic, 3- good, 4- advanced). This issue arose due to time constraints and because assessors only provided numeric feedback rather than providing qualitative feedback to support and justify their ratings. There was also some uncertainty amongst participants about whether and when to use supporting documentation in addition to session recordings (particularly for the formulation and measuring change domains). Participants felt that these issues could be resolved by including additional information in the feedback form and provision of training. Second, participants wanted 'clarification within the rating guidelines', particularly about whether the domain 'measuring change' included informal measures of change (e.g. simple visual analogue scales) as well as standardised questionnaires. Participants also felt the title 'coherent formulation' did not adequately reflect the idea that the item refers to whether the formulation was actively used and updated.

Third, participants felt there was a need for some 'modification to the scoring system'. The number of points on the ACCS scale was debated. Some participants wanted

more than four scale points to improve sensitivity, whilst others felt the use of ½ marks for those who fell between two of the descriptors would be helpful. A number of participants found the use of mean scores for each domain confusing and unnecessary. The reason for including this was discussed (i.e. to prevent the different number of items in the domains resulting in uneven domain weighting in the total score). Participants felt that usability and uncertainty about which aspects of CBT skill are more predictive of patient outcome than others negated this argument. Some participants found it difficult to interpret the total ACCS score and suggested adding information about the total score if a therapist's performance was consistently rated as 1-limited, 2-basic, 3-good, or 4-advanced. The possibility of including an 'appropriately omitted' option was discussed. Some participants felt this would be helpful (e.g. if no idiosyncratic or standardised questionnaires were used or the formulation wasn't explicitly referred to). However, most felt it should always be possible to rate these items, providing the supporting documentation was used.

The final theme refers to 'debates about the items included in the scale'. Participants initially questioned whether the formulation domain could be further broken down into discrete constituent parts (e.g. shared with the patient, revised in light of new information etc.). However, upon further discussion it became clear that these additional components would often not be evident in a given single session. Some participants felt that Socratic enquiry was not evident in the ACCS and suggested adding an extra item. Other participants recognised that Socratic enquiry was evident to some degree in the collaboration, reviewing homework, and reviewing interventions items and felt that it could be drawn out further within these items. There was some debate amongst participants about the inclusion of the formulation domain within the ACCS. Some questioned whether the formulation is, or should be evident in every session whilst others felt was evident in each session. It was felt that viewing the written formulation alongside the session recording would be helpful. A similar discussion was held about whether it was always appropriate to measure change in symptoms, associated features (e.g. beliefs, behaviours, feelings), and movement towards goals. It became evident that two areas of confusion seemed to underlie the discussion: whether supporting documentation could be used, and whether this domain referred only to standardised, formal measures. Most participants felt that the domain could be rated for every session if supporting documentation was used and if informal, idiosyncratic measures were considered.

In response to participant feedback, revisions to the ACCS were made within four areas: 1) the provision of further rating guidance on the feedback form (addition of item and generic banding descriptions and clarification about the use of supporting documentation), 2) improvements to the manual rating guidelines (e.g. clarification regarding the use of informal measures of change, amending the formulation item title, and providing further information about how to interpret ACCS scores), 3) modifications to the scoring system (e.g. providing guidance about the use of ½ marks, removing total domain scores, and adding a mean item score), and 4) clarifications and revisions to the wording of the item descriptors.

## Discussion

This study sought to obtain feedback from assessors with experience of using the ACCS scale in order to examine assessors' views about usability and utility. Discussion about areas for improvement focussed on the need for a clearer and more detailed feedback form, further clarification within the rating guidelines, and modification to the scoring system and revisions were made to the ACCS scale in response to this feedback. There was also debate about which items should or should not be included in the scale, reflecting the broader question of what constitutes CBT competence. A number of strengths of the ACCS were also identified. In particular, participants felt the scale items were relevant, well operationalised, detailed and specific. Participants also commented that the ACCS had an attractive layout and style, performed a useful formative function, and was clear, simple and easy to use. Overall findings indicate that the ACCS worked well in practice, with only minor refinements being necessary to further improve usability.

## Study 3: Psychometric evaluation of the ACCS

Study 3 investigated the psychometric properties of the assessor-rated and self-rated ACCS scale in 'real world' CBT training and routine practice contexts. The following psychometric properties were examined: the ability of items to discriminate levels of competence (i.e. their ability to adequately capture and differentiate between different levels of competence); internal consistency (i.e. the degree to which items assess the same underlying construct); inter-rater reliability (i.e. the level of agreement between different assessors' ratings on the ACCS); discriminant validity (i.e. whether the ACCS is sensitive to improvements in competence); and convergent validity (i.e. whether ACCS

scores correlate well with a previous measure of competence: the Revised Cognitive Therapy Scale [CTS-R]). A secondary aim of the study was to provide an exploratory comparison of the psychometric properties of scores on the ACCS and the CTS-R (Blackburn et al., 2001) in terms of internal consistency, inter-rater reliability and discriminant validity.

## Method

### Participating CBT Centres

Two centres participated in this study: the Oxford Cognitive Therapy Centre (OCTC) and First Step. Within OCTC, participants were recruited from the 2013/14 intake of the Postgraduate Diploma (PGDip) in CBT course run in collaboration with the University of Oxford (for a detailed description of the course see McManus, Westbrook, Vazquez-Montes, Fennell, & Kennerley, 2010). First Step is a National Health Service Improving Access to Psychological Therapies (IAPT) treatment service for people with mild to moderate depression and anxiety disorders (Department of Health, 2008).

### Participants

Participants were: 1) therapists who completed self-ratings and submitted recordings for assessors to rate (herein referred to as ‘therapists’) and 2) senior therapists who rated the submitted recordings (herein referred to as ‘assessors’). Within First Step, 35 CBT practitioners participated as therapists and 12 participated as assessors. As these groups were not mutually exclusive, seven participants participated as both therapists and assessors. Within OCTC, the 23 therapists enrolled on the PGDip in CBT were invited to participate as therapists. Twenty therapists (86.96%) agreed that their supervisors could rate their recordings using the ACCS and nineteen (82.60 %) agreed to complete self-ratings using the ACCS. Eleven senior CBT practitioners employed as supervisors on the PGDip by OCTC participated as assessors. Table 3 shows participants’ demographic characteristics.

*Insert Table 3 about here*

## Patients

Due to confidentiality, little information about the patients in the submitted recordings was available. However, patients' primary presenting problem(s) were identified by the therapists and assessors rated the complexity of patients in the recordings (from 1- *very straightforward* to 4- *very complex*). Patients in both sites presented primarily with depression (37.14% in First Step and 30.26% in OCTC) or an anxiety disorder (60.00% in First Step and 57.91% in OCTC). There was no significant difference between the perceived complexity of the patients in First Step and OCTC ($U$ = 1209.00, $p$ = .55), with patients in both sites being rated as 'somewhat straightforward' (First Step $M$ = 2.23, $SD$ = 0.70; OCTC $M$ = 2.14, $SD$ = 0.81).

## Rating Procedure

**Oxford Cognitive Therapy Centre.** As part of their training, therapists submitted six recordings of CBT sessions with patients. Data was collected from the first two terms (providing up to four recordings per therapist: see Figure 3). Recordings were selected by therapists who completed an ACCS self-rating of their performance within the recorded session. The recordings were also rated using the CTS-R (Blackburn et al., 2001) and the ACCS by their course supervisors. In addition, 20 session recordings (26.32%) were selected at random and blind double rated by one of the authors (XX). Assessors and therapists were provided with a copy of the ACCS manual.

**First Step.** Therapists routinely submitted video recordings of CBT treatment sessions for feedback within supervision. One recording per therapist was independently viewed and rated using the ACCS by three people: the therapist's supervisor ($n$ = 11), a Senior Psychotherapist within First Step ($n$ = 4), and the therapist themselves (i.e. self-ratings, $n$ = 35). Assessors and therapists were provided with a copy of the ACCS manual and attended a one-day training course.

*Insert Figure 3 about here*

## Materials

**Cognitive Therapy Scale-Revised.** Recordings submitted by therapists within OCTC were rated using the Cognitive Therapy Scale-Revised (CTS-R: Blackburn et al., 2001). This is a 12-item scale that assesses general therapeutic skills and CBT specific

skills on a seven-point scale (0 – *incompetent/non-compliance* to 6- *expert: compliance + high skill*). Total CTS-R scores range from 0 to 72, with higher scores representing a higher level of skill.

**Assessment of Core CBT Skills.** Recordings submitted by therapists within OCTC and First Step were rated using the Assessment of Core CBT Skills (ACCS). All 22-items are rated on a four-point scale measuring clinical skill (1- *limited* to 4- *advanced*). Total ACCS scores range from 22 to 88, with higher scores representing a higher level of skill.

**Session recordings.** The total number of session recordings available was as follows. Within OCTC there were 41 self-ACCS ratings, 76 assessor-ACCS and CTS-R ratings, and 20 ACCS double-assessor ratings. Within First Step there were 35 ACCS self-ratings and 35 ACCS double-assessor ratings. Therapists also completed a supporting information cover sheet providing assessors with information about the therapeutic context of the recording (e.g. session number, presenting problem, treatment goals, etc.).

## Data analysis

Within First Step, all recordings were rated twice on the ACCS: once by a supervisor and once by a senior psychotherapist. To enable generalisability of the results to settings which do not have the resources to use multiple assessors, only the supervisor's ratings was used within the psychometric analysis (with the exception of inter-rater reliability).

**Descriptive statistics.** The mean, standard deviations and range of scores were calculated for the 22 ACCS items and total score. This data was examined to establish whether the items discriminated well, whether any items demonstrated strong positive or negative skew, and to check for floor or ceiling effects. These were completed independently for 76 self-ratings on the ACCS (41 from OCTC and 35 from First Step) and for 111 assessor-ratings on the ACCS (76 from OCTC and 35 from First Step).

**Internal consistency.** To examine how highly each item correlated with the overall scale, corrected item-total correlations were calculated. Cronbach's alpha (α) was used to examine correlations amongst individual items and the α if item deleted was also examined for each item. Internal consistency was calculated independently for 76 self-ratings on the ACCS (41 from OCTC and 35 from First Step) and for the 111 assessor-ratings on the ACCS (76 from OCTC and 35 from First Step). To enable comparison, corrected item-

total correlations and Cronbach's alpha were also calculated for 76 assessor-ratings and 41 self-ratings on the CTS-R (all from OCTC).

**Inter-rater reliability.** Because a pool of assessors was used within OCTC and First Step, the same two raters did not assess every recording. Within OCTC supervisors were allocated as rater 1 and the second marker was allocated as rater 2. Within First Step, supervisors were allocated as rater 1 and senior psychotherapists were allocated as rater 2. Agreement between raters was examined for the total score and individual items by calculating intraclass correlation coefficients (ICC), treating the raters as random effects (Strout & Fleiss, 1979, Model 2, 1). ICC values were calculated for the 20 pairs of assessor-ratings on the ACCS completed in OCTC and for the 35 pairs of assessor-ratings on the ACCS completed in First Step.

**Discriminant validity.** The competence of therapists undertaking Diploma-level CBT training has been shown to increase during training (McManus et al., 2010; Williams, Moorey, & Cobb, 1991). Thus, it would be expected that trainees' ACCS scores would increase as they develop skills during training. A repeated measure ANOVA was used to examine whether assessor-rated ACCS total scores increased significantly over the course of training and Bonferroni post-hoc pairwise comparisons were used to test for differences between first and subsequent recordings. To enable comparison, this analysis was also conducted for CTS-R ratings. Analysis was conducted for the 17 OCTC therapists for whom a full data set were available.

**Convergent validity.** The correlation (Pearson's coefficient) between the 76 ACCS and CTS-R ratings completed by assessors within OCTC was examined to explore the relationship between the scales. As both scales assess CBT competence, it was expected that scores from the two scales would be positively correlated.

## Results

### Descriptive Statistics

**Assessor-rated ACCS.** The mean total score within OCTC was 58.41 (*SD* = 8.13, range 37 to 74) and within First Step was 58.14 (*SD* = 10.52, range 27 to 76). Items in the measuring change domain ('choosing suitable measures' and 'implementing measures') were clustered around the lower range of the scale in OCTC and First Step. Two items ('interpersonal style' and 'empathic understanding') were clustered around the upper range of the scale within OCTC, although this was not replicated in First Step.

**Self-rated ACCS.** The mean total score within OCTC was 53.12 (*SD* = 6.88, range 36 to 66) and within First Step was 54.40 (*SD* = 10.95, range 23 to 76). Therapists in OCTC tended not to assign the upper limit of the scale and therapists in First Step tended not to assign the lower limit of the scale. Within OCTC, items in the measuring change domain ('choosing suitable measures' and 'implementing measures') were clustered around the lower range of the scale, although this finding was not replicated in First Step. The 'empathic understanding' item, was also clustered around the upper range of the scale within First Step, although this was not replicated within OCTC.

**Internal Consistency**

The range of item-total correlations considered 'acceptable' is 0.30 to 0.80 (Loewenthal, 2001; Streiner & Norman, 2003). Items below this range [<.3] indicate that the item is not measuring the same construct as other items in the scale and items above this range [>.8] indicate item overlap and thus may be redundant. Nearly all of the items (86.36 %) fell within the acceptable range for both the assessor and self-rated versions of the ACCS across First Step and OCTC. Two items fell just below this range ('reviewing homework' and 'rationale for interventions') and one item fell just above this range ('collaboration'). However, as these items only narrowly missed the threshold and the results were not consistent across sites, no items were removed from the scale. Corrected item-total correlations for the assessor-rated CTS-R ranged from .41 to.72 (*n* = 76 from OCTC), and for the self-rated CTS-R ranged from .55 to .82 (*n* = 41 from OCTC), and thus also fell within the acceptable range.

Cronbach's alpha (α) ranges from 0 = *items independent* to 1 = *items identical*. Cronbach's alpha for the assessor-rated version of the ACCS was .90 in OCTC and .94 in First Step, which is comparable to the Cronbach's alpha for the assessor-rated CTS-R in this sample (α = .90, OCTC data only, *n* = 76). For the self-rated version of the ACCS Cronbach's alpha was .88 in OCTC and .88 in First Step, which is comparable to the Cronbach's alpha for the self-rated CTS-R in this sample (α = .90, OCTC data only, *n* = 41). Thus, there was more than satisfactory agreement between scale items for the self- and assessor-rated version of the ACCS. The α if item deleted fell within the range of .86 and .95 for the self- and assessor-rated ACCS within First Step and OCTC, indicating that none of the ACCS items would significantly increase the scale α if they were deleted.

**Inter-rater Reliability**

Table 4 shows the intra-class correlations between assessors. The following benchmarking scale was used to interpret agreement coefficients: < 0.20 = *poor,* 0.21 - 0.40 = *fair*, 0.41 - 0.60 = *moderate*, 0.61 - 0.80 = *good*, and 0.81 – 1.0 = *very good* (Gwet, 2010). Agreement between raters for the ACCS total score was good in OCTC (ICC = .74) and in First Step (ICC = .73). ICCs for individual items ranged from .79, $p < .001$ to .27, NS in OCTC and from .83, $p < .001$ to .28, NS in First Step.

*Insert Table 4 about here*

**Discriminant Validity**

The mean total scores for the four recordings submitted in the first two terms of OCTC'S PGDip in CBT ($N = 17$) are presented in Figure 4. A repeated measures ANOVA indicated a significant increase over time in ACCS total scores ($F[3,48] = 5.50$, $p < .01$) and CTS-R total scores ($F[3,48] = 6.35$, $p < .01$), with significant increases in ACCS and CTS-R scores from recordings one to three ($p > .05$) and from recordings one to four ($p > .01$). Thus, as therapists progressed through the course, their scores on the ACCS and CTS-R increased, reflecting increased CBT competence.

*Insert Figure 4 about here*

**Convergent Validity**

There was a strong positive correlation ($r = .65$, $p = >.001$) between total scores on the ACCS and the CTS-R assigned by assessors within OCTC ($n = 76$) and between total self-rated scores on the ACCS and CTS-R ($r = .59$, $p = >.001$, $n = 40$ within OCTC).

## Discussion

Results indicate that scores on the Assessment of Core CBT Skills (ACCS) rating scale demonstrated good reliability and validity, both when used as a self-assessment and as an assessor rated tool. Descriptive statistics show that the majority of ACCS items did not demonstrate a strong positive or negative skew, were able to capture different levels of competence, and were not limited by floor or ceiling effects, both in the assessor- and self-rated samples. However, items in the measuring change domain ('choosing suitable measures' and 'implementing measures') clustered around the lower range of the scale, whilst two items ('interpersonal style' and 'empathic understanding') clustered around the upper range of the scale in the assessor-rated sample. This could be explained by an

inability of the ACCS to discriminate between levels of performance in these domains. Alternatively, these findings could be due to a lack of variability among the sample. This may be likely within the context of generic therapeutic skills such as impersonal style and empathic understanding, given that the therapist participants were predominantly NHS professionals with a number of years of clinical training and experience outside of a CBT framework. It is also possible that those completing the ACCS did not give sufficient consideration to the supporting information relating to measures employed. Descriptive statistics were broadly comparable across sites. The exception to this was the distribution of the self-rated ACCS scores: therapists in OCTC tended not to assign the upper limits of the scale, whilst therapists in First Step tended not to assign the lower limits of the scale. This may by reflective of an inability of the ACCS to discriminate between levels of performance in these domains, a perceived difference in skill (i.e. the therapists undergoing CBT training may have been more likely to underrate their skills), or a genuine difference in skill displayed relating to differences in the level of CBT experience between the two samples.

Items in the self- and assessor-rated ACCS were highly intercorrelated, but did not indicate excessive item overlap or redundancy. Cronbach's alphas for the assessor-rated ACCS ($\alpha$ = .90 in OCTC and .94 in First Step) were also comparable to those reported elsewhere for assessor-ratings on the Revised Cognitive Therapy Scale (CTS-R: $\alpha$ range = .75 - .97; Blackburn et al., 2001; James, Blackburn, Milne, & Reichfelt, 2001; Reichelt, James, & Blackburn, 2003), as well as within the current sample. As therapists progressed through the Postgraduate Diploma in CBT, their level of competence on the ACCS improved significantly as did their level of competence on the CTS-R. Thus, scores on the ACCS appear to compare well with scores on the CTS-R in terms of discriminant validity and could provide a useful scale for measuring therapists' progress within CBT training. There was a strong positive correlation between total scores on the ACCS and the CTS-R for both the assessor and self-rated versions of the scales, indicating that the scales measure the same underlying construct (CBT competence) but include distinct content.

Inter-rater reliability for individual items ranged from fair to good (ICCs ranged from .27 to .83), with none of the individual items falling in the range of poor agreement. This is an improvement on the inter-rater reliability reported for individual CTS-R items, which shows poor agreement for a number of items (ICC = -.14 to .84 [Blackburn et al., 2001] pre-training $r$ = .07 to .59, post-training $r$ = .26 - .62 [Reichelt et al., 2003]).

However, within First Step, agreement did fall in the fair range for three items relating to the provision of homework ('choosing suitable homework', 'rationale for homework', and 'planning homework') and for 'pace'. Within OCTC, five different items fell in the fair range ('feasible agenda', 'coherent formulation', 'empathic understanding', 'collaboration', and 'patient feedback'). These differential results across sites indicate that it may be more difficult to establish agreement within individual items across different services or training settings than between assessors who work within the same setting.

Agreement between assessors for the ACCS total score was good across both sites (ICC = .74 in OCTC and .73 in First Step). This is comparable or higher than inter-rater reliability achieved with the CTS-R (ICC range = .40 to .87, average ICC = .63 [Blackburn et al., 2001], *r* = .67 with training and .44 without training [Reichelt et al., 2003], ICC = .38 [Gordon, 2006]). These results are encouraging given that good inter-rater reliability is often difficult to achieve when assessing CBT competence. Previous research has shown that a large amount of assessor training is necessary to achieve adequate inter-rater reliability on the CTS-R (Gordon, 2006; Reichelt et al., 2003). Within the current study, OCTC assessors received no training, whilst assessors in First Step attended a one-day training session in how to use the ACCS. It was, therefore, surprising that inter-rater reliability for the total score was good within both sites. Although this suggests that assessors may not require training in order to achieve good inter-rater reliability on the ACCS, it is important to recognise that the sample within OCTC may not be representative of assessors who would typically use the scale as they all had a great deal of prior experience in assessing CBT competence using other rating scales. Further research is therefore needed to establish whether adequate inter-rater agreement can be achieved on the ACCS without assessor training.

## Overall Discussion

### Limitations and Future Directions

The studies described have several limitations. It is possible that participants in Studies 1 and 2 were invested in the development of a novel CBT competence scale and thus viewed the ACCS in a favourable light. Additionally, the recruitment process for the focus group may have resulted in a relatively homogenous sample, as all participants worked within the same organisation. Participants within the initial feedback study were also asked about perceived usability after having read the scale, rather than having used it

in practice. The focus group study was, however, able to gain feedback on usability from individuals with experience of using the scale in practice, although the sample size was relatively small ($n$ = 9). Despite being appropriate sample size for conducting in-depth focus groups (Stewart et al., 2007), this does limit the generalisability of the findings.

Although the sample size used in study 3 is comparable to or larger than those used to evaluate other competence rating scales, it remains relatively small and thus is a limitation. It is also possible that the sample may not be representative of the population for which the scale is intended. To minimise any idiosyncrasies, two very different sample populations were recruited: novice CBT therapists taking part in a CBT training course (OCTC) and accredited therapists delivering CBT within a National Health Service routine practice setting (First Step). This strategy also meant that the evaluation of the scale was conducted within a treatment centre (First Step) with no prior knowledge of or affiliation with the ACCS.

Ratings within study 3 were completed by assessors who knew the therapists, which is realistic given that many training courses and routine practice settings use supervisors to rate competence. However, assessors' prior knowledge of the therapists may have influenced their ratings. Assessors also rated the same therapist more than once, meaning that rater confounds could have influenced the results. Patients in the rated recordings primarily presented with an anxiety disorder and/or depression and were largely judged to be somewhat straightforward cases. It is, therefore, not possible to draw conclusions regarding the validity or reliability of scores on the ACCS when assessing the delivery of CBT to patients from other populations (e.g. acute settings or severe and enduring disorders such as psychosis or personality disorders).

Further examination of the psychometric properties of the ACCS within the context of more severe and complex patient presentations and a more diverse therapist group will be an important extension of the current study. Some participants in the evaluation studies also felt that assessor training may have improved usability of the ACCS and previous research has shown that assessor training can yield improved inter-rater reliability (Gordon, 2006; Reichelt et al., 2003). Thus, another useful avenue for further exploration is to examine whether assessor training is necessary in order to achieve adequate usability and inter-rater reliability using the ACCS or whether use of the manual alone is sufficient. It will also be important for future studies to examine whether the aspects of competence included in the ACCS are, in practice, necessary to achieve good patient outcomes.

Finally it is important to recognise the limitations in terms of the scope of the scale. The ACCS is designed to assess whether a therapist has demonstrated the core generic and CBT specific skills required to deliver effective CBT within an active treatment session. Thus a number of aspects of competence are not assessed by the ACCS. First, the ACCS does not assess therapists' knowledge or understanding of CBT, aspects of competence which can instead be assessed using multiple choice questionnaires, essays, case reports or clinical vignettes (Muse & McManus, 2013). Second, as the ACCS focuses on 'intervention competence', broader professional skills (e.g. ethical practice, effective use of supervision) are not covered. Third, the scale focuses on competences which are transdiagnostic, rather than skills which are specific to a particular disorder or treatment protocol. Fourth, as the ACCS assesses core CBT skills evident during active, mid-treatment therapy sessions, it does not assess therapists' assessment or relapse-prevention skills. Hence it is recommended that, where the scale is used for summative assessment purposes, the ACCS should not be used as a stand-alone measure of competence. Instead the ACCS should form part of a multi-method competence assessment programme.

**Concluding remarks**

The current paper reports on three studies involved in developing the ACCS. These include a large-scale feedback study examining content validity, face validity and usability; an in-depth focus group evaluating usability and utility; and an investigation of psychometric properties of the ACCS in 'real world' CBT training and routine practice contexts. The results of these studies indicate that the ACCS is comprehensive and includes items that are relevant, well operationalised, detailed and specific, and clear. The ACCS was found to be a user-friendly tool with good style, appearance and layout and an appropriate scoring system. Thus the ACCS was found to have good face validity, content validity, and usability. The ACCS also appears to provide a useful tool for promoting self-reflection and providing formative feedback. Scores on both the self-rated and assessor-rated ACCS demonstrated good internal consistency, inter-rater reliability, and discriminant validity. In addition, scores on the ACCS were found to be correlated with but distinct from the CTS-R and were comparable to the CTS-R in terms of internal consistency and discriminant validity. Additionally, the ACCS may have advantages over the CTS-R in terms of inter-rater reliability. Taken together, these results indicate that the ACCS provides an appropriate and useful measure of CBT competence and is a useful additional tool for self-reflection and providing formative and summative feedback. As

such, the ACCS appears to be suitable for use in clinical practice, training settings and research studies and can be used as a self-rating tool as well as an assessor-rated tool.

## References

Barber, J. P., Sharpless, B. A., Klostermann, S., & McCarthy, K. S. (2007). Assessing intervention competence and its relation to therapy outcome: A selected review derived from the outcome literature. *Professional Psychology-Research and Practice, 38*(5), 493-500. doi: 10.1037/0735-7028.38.5.493

Blackburn, I. M., James, I. A., Milne, D. L., Baker, C., Standart, S., Garland, A., & Reichelt, F. K. (2001). The revised cognitive therapy scale (CTS-R): Psychometric properties. *Behavioural and Cognitive Psychotherapy, 29*(4), 431-446. doi: 10.1017/S1352465801004040

Barber, J. P., Sharpless, B. A., Klostermann, S., & McCarthy, K. S. (2007). Assessing intervention competence and its relation to therapy outcome: A selected review derived from the outcome literature. *Professional Psychology-Research and Practice, 38*(5), 493-500. doi: 10.1037/0735-7028.38.5.493

Blackburn, I. M., James, I. A., Milne, D. L., Baker, C., Standart, S., Garland, A., & Reichelt, F. K. (2001). The revised cognitive therapy scale (CTS-R): Psychometric properties. *Behavioural and Cognitive Psychotherapy, 29*(4), 431-446. doi: 10.1017/S1352465801004040

Beck, J. S. (1995). *Cognitive therapy: Basics and beyond*. New York: Guildford Publications.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology, 3*, 77 - 101. doi: 10.1191/1478088706qp063oa

Brewer, J., & Hunter, A. (2005). *Foundations of multimethod research: Synthesizing styles*. Thousand Oaks, CA: Sage Publications.

Burisch, M. (1984). Approaches to personality inventory construction: A comparison of merits. *American Psychologist, 39*(3), 214-227. doi: 10.1037/0003-066x.39.3.214

Campanelli, P. C., Martin, E. A., & Rothgeb, J. M. (1991). The Use of Respondent and Interviewer Debriefing Studies as a Way to Study Response Error in Survey Data. *Journal of the Royal Statistical Society. Series D (The Statistician), 40*(3), 253-264. doi: 10.2307/2348278

Department of Health. (2008). IAPT implementation plan: National guidelines for regional delivery. Retrieved from http://www.iapt.nhs.uk

Dobson, K. S., & Singer, A. R. (2005). Definitional and practical issues in the assessment of treatment integrity. *Clinical Psychology: Science and Practice, 12*(4), 384-387. doi:10.1093/clipsy.bpi046

Farchione, T. J., Fairholme, C. P., Ellard, K. K., Boisseau, C. L., Thompson-Hollands, J., Carl, J. R., Gallagher, M. W., & Barlow, D. H. (2012). Unified protocol for transdiagnostic treatment of emotional disorders: A rangomised controlled trial. *Behaviour Therapy, 43(3),* 666 – 678. doi: 10.1016/j.beth.2012.01.001

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

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

Govaerts, M., van der Vleuten, C., Schuwirth, L., & Muijtjens, A. (2007). Broadening perspectives on clinical performance assessment: Rethinking the nature of in-training assessment. *Advances in Health Sciences Education, 12*(2), 239-260. doi: 10.1007/s10459-006-9043-1

Gwet, K. L. (2010). *Handbook of inter-rater reliability (2nd ed.): The definitive guide to measuring the extent of agreement among multiple raters*. Gaithersburg: Advanced Analytics.

Horvath, A. O., & Symonds, B. D. (1991). Relation between working alliance and outcome in psychotherapy: A meta-analysis. *Journal of Counseling Psychology, 38*(2), 139-149. doi: 10.1037/0022-0167.38.2.139

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

Krosnick, J., & Fabrigar, L. R. (1997). Designing rating scales for effective measurement in surveys. In L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz & D. Trewin (Eds.), *Survey measurement and process quality* (pp. 141-164). New York: Wiley.

Kaslow, N. J. (2004). Competencies in professional psychology. *American Psychologist, 59*(8), 774-781. doi: 10.1037/0003-066x.59.8.774

Kazantzis, N., Deane, F. P., & Ronan, K. R. (2000). Homework Assignments in Cognitive and Behavioral Therapy: A Meta-Analysis. *Clinical Psychology: Science and Practice, 7*(2), 189-202.

Kazantzis, N., & Lampropoulos, G. K. (2002). Reflecting on homework in psychotherapy: What can we conclude from research and experience? *Journal of Clinical Psychology, 58*(5), 577-585. doi: 10.1002/jclp.10034

Keijsers, G. P. J., Schaap, C. P. D. R., & Hoogduin, C. A. L. (2000). The impact of interpersonal patient and therapist behavior on outcome in cognitive-behavior therapy: A review of empirical studies. *Behavior Modification, 24*(2), 264-297. doi: 10.1177/0145445500242006

Kirk, J. (1998). Cognitive-behavioural assessment. In K. Hawton, P. M. Salkovskis, J. Kirk & D. M. Clark (Eds.), *Cognitive behaviour therapy for psychiatric problems: A practical guide*. Oxford: Oxford University Press.

Kuyken, W., Padesky, C. A., & Dudley, R. (2011). *Collaborative case conceptualisation: Working effectively with clients in cognitive-behavioral therapy*. New York: Guildford Press.

Kvale, S. (1996). *Interviews: An introduction to qualitative research interviewing.* London: Sage Publications.

Laireiter, A. R., & Willutzki, U. (2003). Self-reflection and self-practice in training of cognitive behaviour therapy: An overview. *Clinical Psychology and Psychotherapy, 10*(1), 19–30. Doi:10.1002/cpp.348.

Loewenthal, K. M. (2001). *An introduction to psychological tests and scales (2nd ed.).* Philadelphia: Taylor & Francis Inc.

Lynn, M. R. (1986). Determination and quantification of content validity. *Nursing Research, 35*(6), 382-385. 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*(3), 438-450. doi: 10.1037/0022-006x.68.3.438

McHugh, R. K., & Barlow, D. H. (2010). The dissemination and implementation of evidence-based psychological treatments: A review of current efforts. *American Psychologist, 65*(2), 73-84. doi: 10.1037/a0018121

McManus, F., Westbrook, D., Vazquez-Montes, M., Fennell, M., & Kennerley, H. (2010). An evaluation of the effectiveness of Diploma-level training in cognitive behaviour

therapy. *Behaviour Research and Therapy*, *48*(11), 1123-1132. doi: 10.1016/j.brat.2010.08.002

Miller, G. E. (1990). The assessment of clinical skills/competence/performance. *Academic Medicine, 65*, 63-67.

Milne, D. (2007). Evaluation of staff developoment: The essential 'SCOPPE'. *Research and Evaluation, 16(3),* 389 – 400. doi: 10.1080/09638230701382818

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

Muse, K., & McManus, F. (2015). Expert insight into the assessment of CBT competence: A qualitative exploration of experts' experiences, opinions and recommendations. *Clinical Psychology and Psychotherapy, Advance online publication.*

Newman, C. F. (2013). *Core competencies in cogitive-behavioral therapy*. Sussex: Routledge.

Pearsons, J. B. (1993). Case conceptualisation in cognitive-behavior therapy. In K. T. Kuehlwein & H. Rosen (Eds.), *Cognitive therapies in action: Evolving innovative practice* (pp. 33 - 53). San Fransisco: Jossey-Bass.

Reichelt, F., James, I. A., & Blackburn, I.-M. (2003). Impact of training on rating competence in cognitive therapy. *Journal of Behavior Therapy and Experimental Psychiatry*, *34*(2), 87-99. doi: 10.1016/S0005-7916(03)00022-3. ISSN: 0005-7916.

Ritchie, J., & Spencer, L. (2002). Qualitative data analysis for applied policy research. In A. M. Huberman & M. B. Miles (Eds.), *The qualitative researcher's companion* (pp. 305 - 329). London: Sage Publications.

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

Schwarz, N., Knauper, B., Hippler, H.-J., Noelle-Neumann, E., & Clark, L. (1991). Rating scales numberic values may change the meaning of scale labels. *Public Opinion Quarterly, 55*(4), 570-582. doi: 10.1086/269282

Stewart, D. W., Shamdasani, P. N., & Rook, D. W. (2007). *Focus groups: Theory and practice* (2nd ed.). London: Sage Publications.

Streiner, D. L., & Norman, G. R. (2003). *Health measurement scales: A practical guide to their development and use (3rd ed.)*. Oxford: Oxford University Press

Van der Vleuten, C., Schuwirth, L. W. T., Scheele, F., Driessen, E. W., & Hodges, B. (2010). The assessment of professional competence: Building blocks for theory development. *Best practice and research clinical obstetrics and Gyneacology, 24*, 703 – 71. 10.1016/j.bpobgyn.2010.04.001

Van Vaerenbergh, Y., & Thomas, T. D. (2012). Response Styles in Survey Research: A Literature Review of Antecedents, Consequences, and Remedies. *International Journal of Public Opinion Research.* doi: 10.1093/ijpor/eds021

Waddington, L. (2002). The therapy relationship in cognitive therapy: A review. *Behavioural and Cognitive Psychotherapy, 30*(02), 179-191. doi: 10.1017/S1352465802002059

Wallsten, T. S., Budescu, D. V., Zwick, R., & Kemp, S. M. (1993). Preferences and reasons for communicating probabilistic information in verbal or numerical terms. *Bulletin of the Psychonomic Society, 31*(2), 135-138.

Weck, F., Bohn, C., Ginzburg, D. M., & Ulrich, S. (2011). Treatment integrity: Implementation, assessment, evaluation and correlations with outcome. *Verhaltenstherapie, 21*, 99-107. doi: 10.1159/000328840

Westbrook, D., Kennerley, H., & Kirk, J. (2007). *An introduction to Cognitive Behaviour Therapy: Skills and Applications*. London: Sage Publications Ltd.

Williams, R. M., Moorey, S., & Cobb, J. (1991). Training in cognitive-behaviour therapy: Pilot evaluation of a training course using the cognitive therapy scale. *Behavioural and Cognitive Psychotherapy, 19*(04), 373-376. doi: 10.1017/S0141347300014075

Yagmale, F. (2003). Content validity and its estimation. *Journal of Medical Education, 3*(1), 25 - 27.

Table 1. Demographic characteristics for participants in study 1.

| | **Novices** *n* = 25 | **Experts** *n* = 41 |
|---|---|---|
| **Gender** - % female | 85.00% | 58.54% |
| **Age** – Mean (*SD*) | 36.97 (10.00) | 47.10 (11.07) |
| [range] | [25 – 55] | [30 – 71] |
| **Years practicing CBT** | 2.03 (2.23) | 19.09 (9.67) |
| Mean (*SD*) [range] | [0.00 – 6.00] | [5.00 – 38.00] |
| **Number of CBT cases:** | | |
| % treated 0 cases | 44.00 % | 0.00 % |
| % treated 1 - 50 cases | 44.00 % | 0.00 % |
| % treated 50 – 200 cases | 12.00 % | 24.39 % |
| % treated > 200 cases | 0.00 % | 75.61 % |

Table 2. Content validity results for each domain in the Assessment of Core CBT Skills (ACCS) [1]

| | **CVI** [1] | | **Relevance (1 - 4)** | | | **Clarity (1 - 4)** | | |
|---|---|---|---|---|---|---|---|---|
| **Domain** | Novices *n* = 25 % | Experts *n* = 41 % | Novices *n* = 25 Mean (*SD*) | Experts *n* = 41 Mean (*SD*) | Mann-Whitney *U* | Novices *n* = 25 Mean (*SD*) | Experts *n* = 41 Mean (*SD*) | Mann-Whitney *U* |
| Agenda Setting | 96.0 | 85.4 | 3.96 (.20) | 3.83 (.50) | 469.50 *p* = .25 | 3.68 (.56) | 3.46 (.67) | 426.00 *p* = .18 |
| Formulation | 92.0 | 92.7 | 4.00 (.00) | 3.98 (.16) | 500.00 *p* = .44 | 3.72 (.74) | 3.61 (.63) | 440.50 *p* = .21 |
| CBT Interventions | 96.0 | 87.8 | 3.96 (.20) | 3.83 (.44) | 457.50 *p* = .17 | 3.68 (.56) | 3.41 (.71) | 410.50 *p* = .19 |
| Homework | 96.0 | 100 | 3.96 (.20) | 3.93 (.26) | 495.50 *p* = .59 | 3.68 (.56) | 3.76 (.44) | 489.00 *p* = .68 |
| Assessing Change | 84.0 | 82.9 | 3.76 (.52) | 3.68 (.57) | 478.00 *p* = .54 | 3.68 (.80) | 3.59 (.71) | 455.00 *p* = .31 |
| Effective Use of Time | 92.0 | 95.1 | 3.88 (.44) | 3.93 (.35) | 496.50 *p* = .61 | 3.72 (.54) | 3.59 (.55) | 440.50 *p* = .25 |
| Fostering Therapeutic Relationship | 92.0 | 95.1 | 3.76 (.60) | 3.90 (.44) | 457.00 *p* = .15 | 3.88 (.33) | 3.71 (.46) | 424.00 *p* = .10 |
| Effective Two-way Communication | 96.0 | 97.6 | 4.00 (.00) | 3.95 (.22) | 487.00 *p* = .27 | 3.84 (.47) | 3.71 (.51) | 440.50 *p* = .18 |

[1] Feedback was obtained for the *original* version of the ACCS. Following evaluation of this initial draft, further refinements were made resulting in a *final* scale, as outlined this paper.

[2] CVI = Content Validity Index, the percentage of participants who rated item as ≥ three for **both** relevance and clarity

Table 3

*Demographic characteristics for participants in Study 3*

| | First Step | | OCTC [1] | |
|---|---|---|---|---|
| | **Therapists** *n* = 25 [2] | **Assessors** *n* = 12 | **Therapists** *n* = 17 [2] | **Assessors** *n* = 11 |
| **Gender** - % female | 80.00 % | 75.00 % | 76.00 % | 100 % |
| **Years practicing CBT** Mean (*SD*) [range] | 5.60 (2.69) [2 – 15] | 8.13 (3.66) [5 – 15] | 3.00 (1.95) [0 – 10] | 19.17 (5.70) [13 – 30] |
| **Number of CBT cases:** | | | | |
| % treated 0 cases | 0.00 % | 0.00 % | 58.82 % | 0.00 % |
| % treated 1 - 50 cases | 3.33 % | 0.00 % | 29.41 % | 0.00 % |
| % treated 50 – 200 cases | 16.67 % | 16.67 % | 11.76 % | 0.00 % |
| % treated > 200 cases | 80.00 % | 83.33 % | 0.00 % | 100 % |
| **BABCP accredited CBT therapists %** | 100 % | 100 % | 0 % | 100 % |

[1] OCTC = Oxford Cognitive Therapy Centre.

[2] Five therapists within First Step and three therapists within OCTC did not complete demographics.

Table 4

*Intra-class correlations between raters for the Assessment of Core CBT Skills (ACCS) in the Oxford Cognitive Therapy Centre (OCTC) and First Step*

| Domain | Item | OCTC *n* = 20 | First Step *n* = 35 |
|---|---|---|---|
| Agenda setting | Suitable items | .69 * | .66 *** |
| | Feasible agenda | .37 | .75 *** |
| Formulation | Coherent and dynamic formulation | .27 | .82 *** |
| CBT Interventions | Appropriate intervention targets | .54 | .67 *** |
| | Choosing suitable interventions | .71** | .68 *** |
| | Rationale for interventions | .78 ** | .71 *** |
| | Implementing interventions | .79 ** | .68 *** |
| | Reviewing interventions | .71 ** | .66 *** |
| Homework | Reviewing homework | .71 ** | .47 * |
| | Choosing suitable homework | .46 * | .28 |
| | Rationale for homework | .67* | .37 |
| | Planning homework | .61* | .32 |
| Appropriate tracking of progress | Choosing suitable measures | .46 | .78*** |
| | Implementing measures | .54 * | .75*** |
| Effective use of time | Pace | .52 | .37 |
| | Time management | .62 * | .42 |
| | Maintained focus | .64 * | .50 * |
| Fostering therapeutic relationship | Interpersonal style | .67 ** | .69 *** |
| | Empathic understanding | .40 | .83 *** |
| | Collaboration | .40 | .71 *** |
| Effective two way communication | Patient feedback | .38 | .57 ** |
| | Reflective summaries | .54* | .61 *** |
| | **Total Score** | .74 ** | .73 *** |

*$p$ < .05; **$p$ < .01; ***$p$ < .001.

Coding for agreement coefficients: < 0.20 = *poor*, 0.21 - 0.40 = fair, 0.41 - 0.60 = *moderate*, 0.61 - 0.80 = *good*, and 0.81 – 1.0 = *very good.*

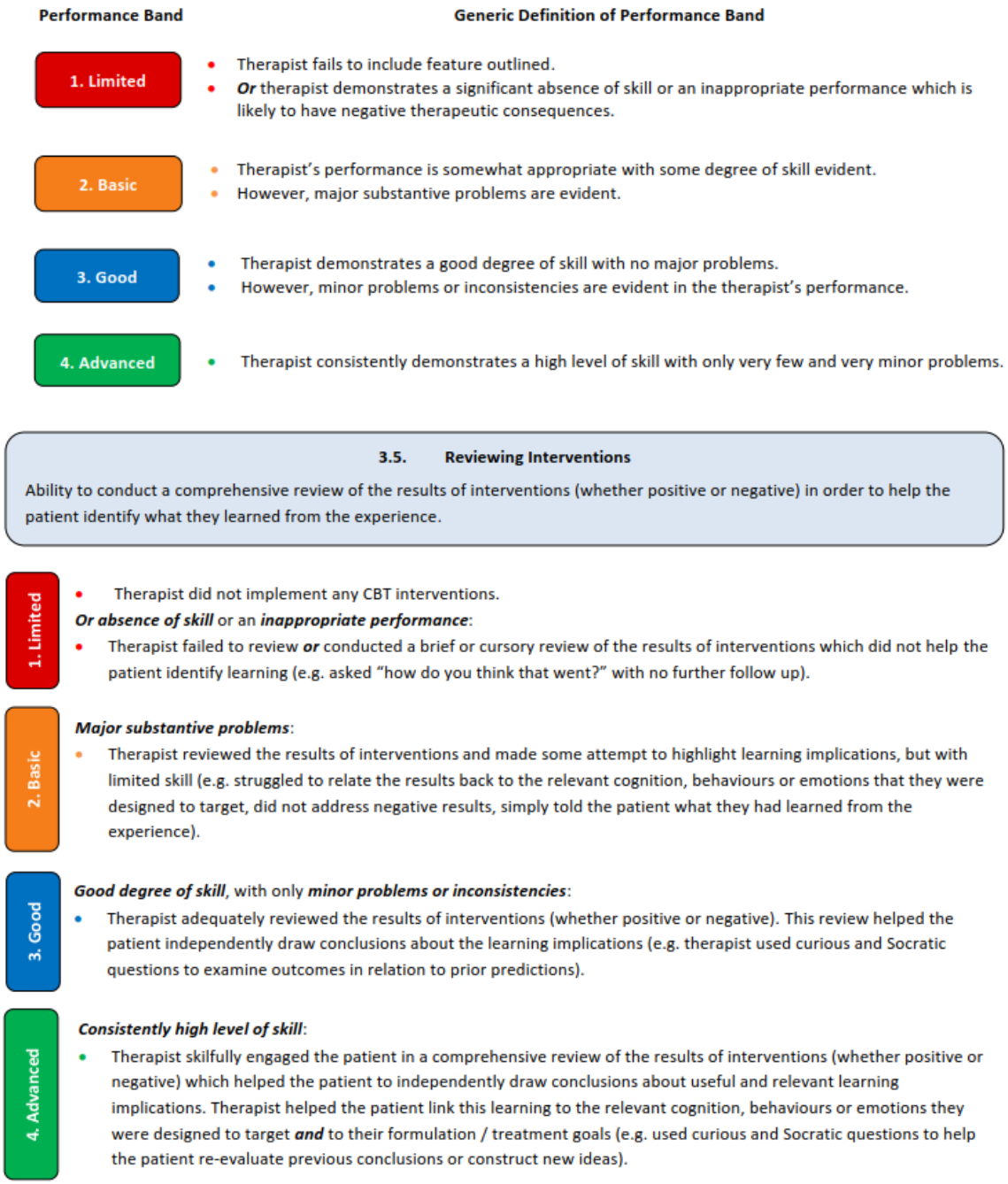


*Figure 1.* The generic performance bandings used for the ACCS scale ratings and an example of item-specific exemplar therapist behaviours for the reviewing interventions item

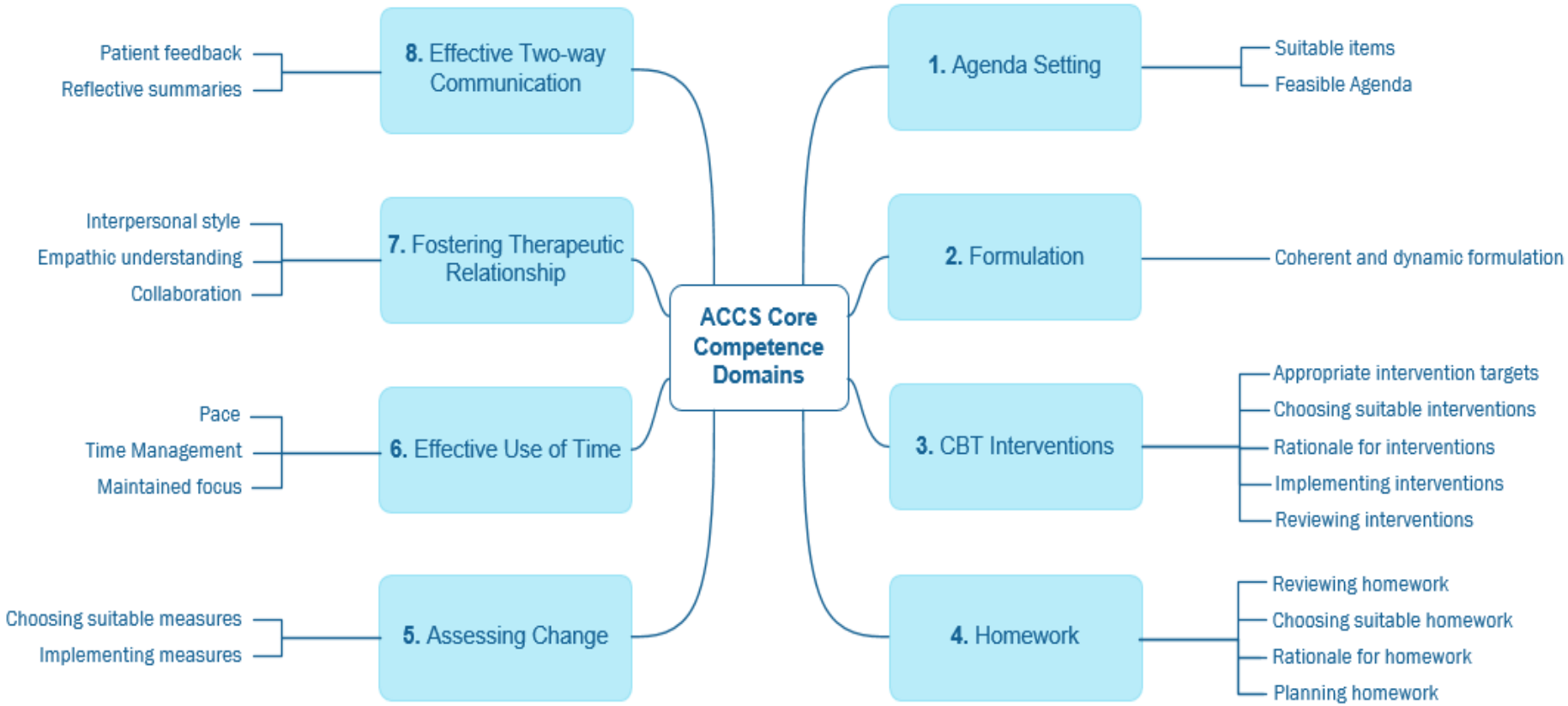


*Figure 2.* Final items included in the Assessment of Core CBT Skills (ACCS) rating scale

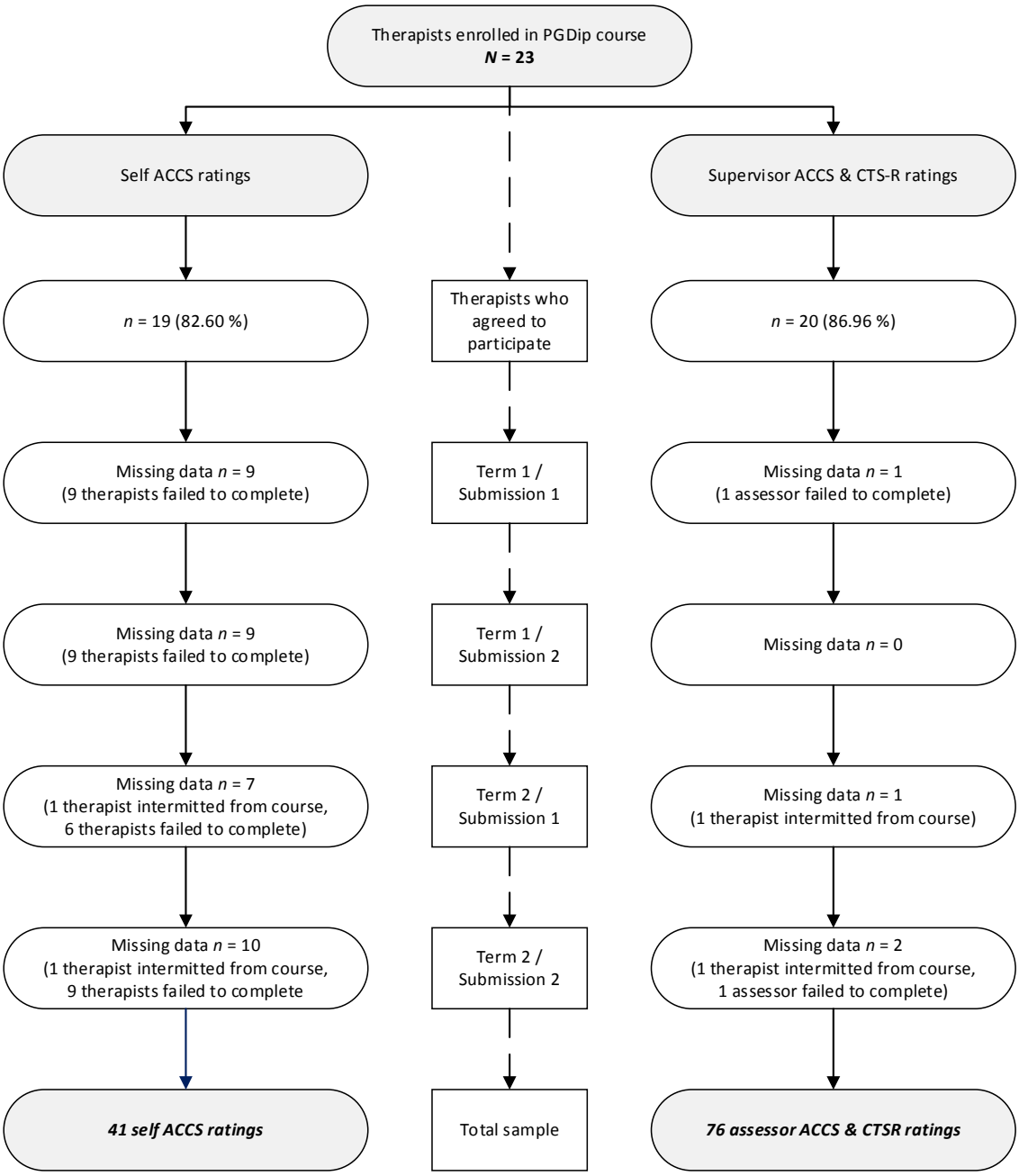


*Figure 3.* Flowchart outlining data collection for Study 3 at the Oxford Cognitive Therapy Centre Site

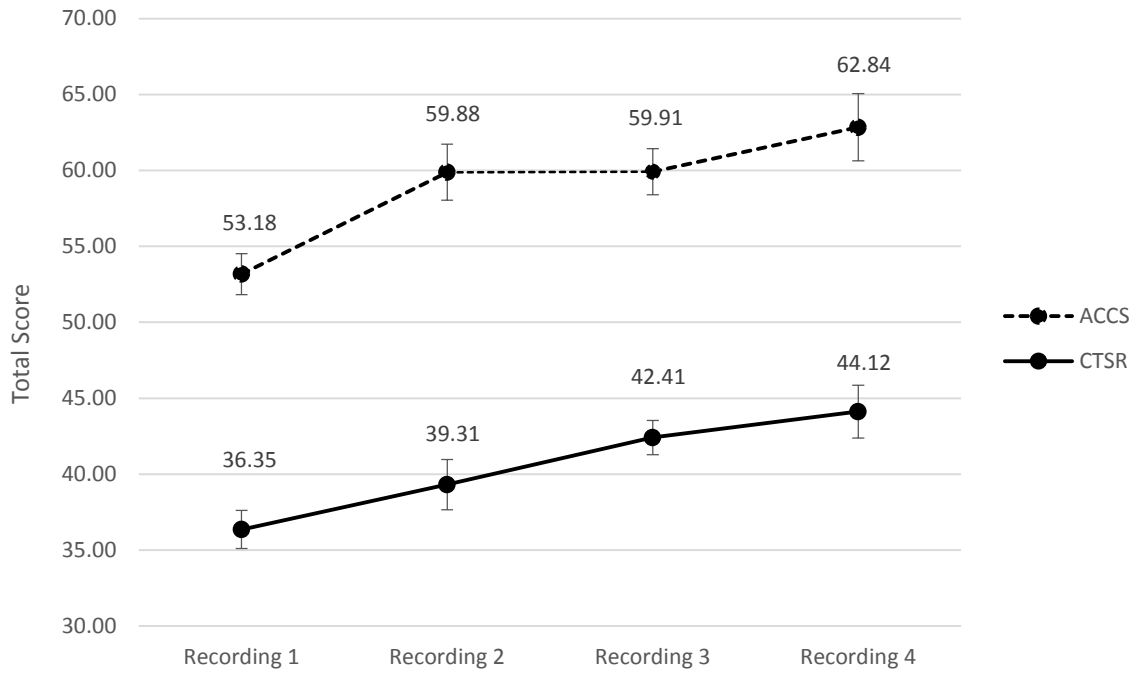


*Figure 4*. Means and standard errors for total Assessment of Core CBT Skills (ACCS: total score range 22 to 88) and Cognitive Therapy Scale-Revised (CTS-R: total score range 0 to 72) scores for recordings submitted in Term I and II of the Oxford Cognitive Therapy Centre Postgraduate Diploma in CBT ($N$ = 17 therapists).