Judging clinical competence using structured observation tools: a cautionary tale

Article

Accepted Version

Judging clinical competence using structured observation tools: a cautionary tale

Abstract

One method for appraising the competence with which psychological therapy is delivered is to use a structured assessment tool that rates audio or video recordings of therapist performance against a standard set of criteria. The present study examines the inter-rater reliability of a well-established instrument (the Cognitive Therapy Scale – Revised; Blackburn et al, 2001) and a newly developed scale for assessing competence in CBT, using six experienced raters working independently and blind to each other's ratings. Inter-rater reliability was found to be low on both instruments, but it is argued that this represents a realistic appraisal of the accuracy of such scales, and that the figures often cited for inter-rater reliability are unlikely to be generalizable outside the specific context in which they were achieved. This raises concerns about the use of these scales for making summative judgments of competence in both educational and research contexts.

Judging clinical competence using structured observation tools: a cautionary tale

Introduction

There are many reasons for developing scales to assess the competence with which a psychological therapy is delivered. For example, researchers may need to establish whether therapists in a clinical trial are adherent to a particular method, and competent in its delivery. In a training context scales can be used to conduct summative assessments of trainee progression, or are used formatively as part of supervision.

A number of scales have been developed to gauge adherence and competence in the delivery of Cognitive Behaviour Therapy (CBT; Muse & McManus, 2013). Of these, the most extensively researched is the Cognitive Therapy Scale (CTS: Young & Beck, 1980) and its later revision (CTS-R (Blackburn, James, Milne, Baker, Standart, et al., 2001). These measures typically rely on the judgment of raters in the structured assessment of audio or video recordings of psychotherapy sessions, so inter-rater reliability and scale validity are key requirements.

Loades and Armstrong (2016) report a systematic review of 20 studies that have investigated the inter-rater reliability of the CTS and its variants. Some had a primary aim of investigating the metrics of the scale, but most employed the CTS in the service of a relevant research question (for example, in studies relating therapist competence to outcome) and reported on inter-rater reliability as part of the study design. Of the 20 studies, nine reported on the use of the CTS or CTS-R applied to clinical work with adults with anxiety or depression, while the remainder reported on adaptations of the CTS intended to make it more applicable to specific client populations (such as people with psychosis, with social anxiety disorder or to children). Intra-class correlation coefficients (ICCs) varied widely across studies, from 0.40 to 0.98,

with a median of 0.65. This variation is also seen within studies: McManus, Rakovshik, Kennerley, Fennell & Westbrook, (2012) report separate evaluations of recordings of trainees early and late in their training; ICCs for these two time-points were 0.47 and 0.71 respectively.

This wide variation in reliability estimates merits further exploration. Differences in estimates of inter-rater reliability may reflect factors such as the degree to which raters were trained and relatedly the extent to which raters have improved concordance by discussing any differences in their interpretation of the scale to achieve consensus. While it seems that groups of raters working together in this way can achieve very high levels of reliability (Loades & Armstrong, *op cit)*, where they are working more independently there seems to be poorer agreement. For example, in Dimidjian, Hollon, Dobson, Schmaling, Kohlenberg, et al. (2006) three raters appraised recordings; two were 'in-house' to the research team and one was an external expert. The overall ICC for the raters working together was 0.94, but this reduced to 0.47 with the inclusion of the 'external' rater, this despite the fact that all the raters were highly expert in CBT, both as trainers and developers. A similar (if more numerically extreme) picture is reported by Jacobson and Gortner (2000) where the ratings of two 'external' assessors (selected to be both expert and independent) were contrasted to each other and to an internal rater, yielding ICCs between 0.01 and 0.08.

It is clear (and not altogether surprising) that groups of raters can work towards a consensual position in which their ratings are closely calibrated, and so achieve good inter-rater reliability. However, *consistency* in ratings does not speak to the 'accuracy' of the judgments being made; reliability does not equate to validity. Scores from different groups of raters may be at a different level, within-group ratings being concordant, but between-group ratings

being discrepant. This observation is particularly pertinent if the rating scale is being used to make summative assessments, for example appraising a trainee's competence to practise. At issue is the level of concordance achieved by raters who are working independently with no or minimal training or active coordination; in other words the reliability and validity of the instrument when used in the field. As such a key aim of this paper is to establish the extent to which measures are appropriately used in routine circumstances for formative and summative evaluation of competence.

A new scale has been developed for structured observation of CBT; its development is described in Roth (2016). This is rooted in the competence framework for CBT (Roth & Pilling, 2007), developed as part of the English Improving Access to Psychological Therapy (IAPT) programme, and used to generate the IAPT CBT curriculum for working with people presenting with anxiety and depression. The framework organises the delivery of CBT into discrete areas of activity, and identifies the knowledge and skills that underpin all variants of CBT as well as specific CBT skills that are applied when working with specific conditions or presentations. A distinctive aspect of the UCL CBT scale is its identification of intervention methods that are present in almost all sessions along with those which characterise evidence-based interventions for specific disorders.

The framework also includes a domain of Generic Therapeutic Competences, knowledge and skills that are common across therapy modalities (for example, relational competences such as alliance building and repair) and skills associated with the management of sessions (for example, using measures, responding to emotional expression, or ending sessions). Although generic competences are necessary skills for the effective delivery of therapy, it is helpful to separate them from CBT-specific skills; by definition they are non-specific, and so do not test

how well a therapist is applying CBT. As such, two parallel scales were developed, both of which would usually be administered, focusing on generic and CBT competences respectively. Unlike measures developed for research use, the UCL scales are intended to be used in routine service contexts without extensive training, on the basis that each item is anchored with descriptions of specific therapist behaviours.

The present study has two aims. First, to benchmark the psychometric properties of the UCL scales against the CTS-R, to check if it has similar levels of reliability in a context where raters are effectively working independently (as would be the case in most real-world settings). Second (because raters employed the Cognitive Therapy Scale – Revised (CTS-R) in parallel with the new scale), to establish the extent to which it is appropriate to use therapy competence measures for formative and summative evaluation of competence in routine settings.

Method

Ethics

Ethical approval for this study was obtained from the University Research Ethics Committee. All clients gave written informed consent for their recordings to be used as part of this research study. Clients whose recordings were included in the trial were informed that their recordings could be used for educational research at the same time as their consent was obtained for recording sessions for training purposes.

Setting

The study was conducted at a university offering training in Cognitive Behaviour Therapy for people with depression and anxiety presentations (as part of the IAPT programme).

Rating scales: Each rater evaluated the whole sample of recordings using a) the Cognitive Therapy Scale-Revised (Blackburn et al., 2001) and b) the UCL generic and CBT competence scales (Roth, 2016).

Therapists: Fourteen therapists contributed recordings to the study; all were registered on a one-year Postgraduate Diploma offering training in CBT as part of the IAPT programme. This means that therapists were of different professional backgrounds, varied in relation to their experience of mental health presentations, and had varying levels of prior exposure to CBT (though all had at least two years of clinical experience, and some had experience of self-help CBT programmes).

Trainees on this programme routinely submit session recordings for evaluation, and the sessions for this study were selected from this corpus. There was no attempt to select recordings systematically in relation to the therapists' prior experience, or their stage of training.

Clients: All clients were seen in the setting of the IAPT services in which their therapists were employed. All were adults referred with a primary diagnosis of depression or with an anxiety disorder (phobia, panic disorder, generalised anxiety disorder or social anxiety).

Session recordings: Twenty-five session recordings (each approximately 50 minutes in duration) were identified for rating by a research assistant, selected from 76 recordings submitted as part of the standard schedule of assessments on the training programme. All were early or mid-treatment sessions (with assessments and final sessions excluded on the grounds that these (by definition) will have a restricted focus). As the study progressed the range of presenting problems was balanced, so as to ensure that there was reasonable representation of different disorders (see Table 1). Most recordings were taken from the middle stage of therapy, with only a minority from the initial or final stages of the intervention (Table 2). Eight therapists contributed a single recording, one contributed two recordings and five contributed three recordings

INSERT TABLE 1 ABOUT HERE

INSERT TABLE 2 ABOUT HERE

Raters: Six raters contributed to the study; all were employed as tutors on an IAPT Postgraduate Diploma programme, and so routinely appraised the work of trainees as part of the examination process. All were female Clinical Psychologists accredited as CBT therapists with the British Association for Behavioural and Cognitive Psychotherapies (BABCP), and had considerable experience both as clinicians (range 6 to 14 years, mean 7.6 years) and as tutors with the programme (range 3 to 6 years; mean 4 years). Raters reviewed all 25 recordings independently, and so were blind as to the ratings of their colleagues.

Training of raters: a) CTS-R: All six raters had received extensive training in the use of the CTS-R as part of their work with the programme, including annual consensus and review meetings aimed at ensuring consistency in their scoring. As such, they were not offered

further training on this instrument. b) UCL competence scales: There was limited training in the use of the UCL scales. As noted earlier, the intent was to approximate 'real-world' application of the scale, and so rely on the instruction materials accompanying the scale and the scale itself. Training comprised a meeting with all six raters focused on an initial session rating, allowing the opportunity for feedback on the scale itself and identifying any areas which were ambiguous or required clarification. A further mid-point consensus meeting was held after 10 session ratings had been completed; a previously-rated recording (subsequently excluded from the study) was reviewed, and clarification of the rating system discussed. This was followed by a 'live' rating of a further session (again, excluded from the study), which allowed for group discussion of reasons for any variation in scoring.

Controlling for order effects: The order in which the CTS-R or the UCL scales were applied was balanced both across recordings and across raters, so as to mitigate the risk that rating on one scale could influence ratings on the other.

Results

In this analysis the Intra-Class Correlation Coefficient (ICC) is computed for absolute agreement and for single raters using a two-way mixed effects model; results are displayed in Table 3.

INSERT TABLE 3 ABOUT HERE

INSERT TABLE 4 ABOUT HERE

a) UCL CBT scale: Across the six raters the ICC for the scale total was poor to moderate (ICC= 0.394: 95% confidence interval 0.228 – 0.598).The mean correlation between raters was 0.45 (range 0.26 to 0.74) As can be seen from Table 4, one rater had consistently low correlations with the other raters: removing this individual from the analysis increased the ICC to 0.476 (95% confidence interval 0.294 – 0.675), and the mean correlation between raters to 0.52.

b) UCL Generic scale: The ICC for the total scale was poor (ICC= 0.272: 95% confidence interval 0.126 – 0.478),with a mean correlation between raters of 0.32 (range -0.11 to 0.57). The same rater had consistently low correlations with the other raters: removing them from the analysis increased the ICC to 0.346 (95% confidence interval 0.174 – 0.562), and the mean correlation between raters to 0.43.

c) CTS-R: The ICC for the total scale was poor to moderate (ICC = 0.424: 95% confidence interval 0.260-0.621), with a mean correlation between raters of 0.44 (range 0.12 to 0.67). Once again the same rater had consistently low correlations with their colleagues: removing this individual from the analysis increased the ICC to 0.516 (95% confidence interval 0.339 – 0.702), and the mean correlation between raters to 0.56.

Discussion

In summary, the inter-rater reliability of the Generic and the CBT UCL scales was poor (with ICCs of 0.272 and 0.394 respectively). Removing the rater with a consistently low level of agreement with their colleagues improved the ICCs for the CBT scale closer to a moderate level (0.476). While the ICC for the generic scale also improved, it remained poor at 0.346. Scores for the CTS-R were broadly comparable to that on the UCL CBT scale; based on ratings from all raters the ICC was poor (ICC = 0.424) , with a moderate level of reliability if the outlier rater is removed (ICC=0.516).

These estimates for the reliability of the UCL scales are low, though they are comparable to the lower end of the range of ICCs found in studies of the CTS-R (Loades & Armstrong, op cit). The raters all had considerable experience using the CTS-R to evaluate session recordings of trainees on the same training programme, with periodic meetings aimed at checking the consistency of their ratings. In contrast, training on the UCL scales was minimal and restricted to an initial meeting at which the scales were discussed, rating of an initial recording, and a concordance meeting after 10 recordings had been evaluated. The figures for the CTS-R therefore represent a benchmark, against which the UCL CBT scale performs equivalently.

These findings confirm that there are significant difficulties achieving high reliability in whole-session structured assessment of therapist competence. Disparities in ratings could be attributable to any number of causes, among the most basic being deficiencies in the way the scale is structured or ambiguity in scale descriptors. But rating therapist competence is inherently challenging: manuals can attempt to anchor ratings, but unless scale items are very

specific and straightforward, raters will inevitably apply idiosyncratic clinical judgments regarding the 'meaning' of a scale item, and so arrive at different (but legitimate) ratings. In the present study raters were asked to explain each of their rating decisions, making it possible to explore the reasons for discrepant ratings and highlighting some of the dilemmas that raters attempted to resolve. For example:

a) Faced with significant intra-session variation in competence (with a specific skill being applied well or poorly at different points) raters sometimes awarded an averaged score or rated in line with the best or poorest examples.

b) Applying CBT techniques requires attention both to structure (how something is set up) as well as content (identifying and working with material that is salient). Therapists sometimes employed a technique (such as evaluating negative automatic thoughts, or setting up a behavioural experiment) in a way that was appropriately structured but focused on content that was not central to the client's issues (so reducing its potential value). Some raters responded by awarding a low rating (noting that the content was misjudged), whereas others awarded a high rating on the basis that the therapist had set-up the technique in a skilful manner,

c) On occasion raters identified significant clinical themes that had not been picked-up by their colleagues (for example, noting that the therapist had missed an important issue), and this led to their appraising specific techniques differently from other raters.

Each example illustrates a challenge to interpretation of scale items, despite the fact that each item was anchored with several examples of the behaviours and actions associated with each area of competence. However, not every eventuality can be anticipated, and at points raters will inevitably fall back on idiosyncratic conceptions.

One solution to this difficulty is to recognise that rating specific competences is an inherently complex and potentially unstable task because of the number of variables that need to be accounted for. Recognising this, some authors (e.g. Elkin, 1999) have suggested that there may be advantages to global rather than specific rating scales. Kuyken and Tsivrikos (2009) developed a four-item scale that rated therapists overall competence, overall skills in CBT, their flexibility, and their general skills, finding that all the scale items were highly inter-correlated, and significantly associated with outcome. The risk with this approach is that it is impossible to know how each rater arrives at a judgment; as such even if their overall ratings are congruent they might be based on different criteria. A workable compromise might be to combine the two approaches, with a scale that asks for global judgments based on detailed descriptions of specific therapist behaviours (as exemplified by the Competence in Cognitive Analytic Therapy scale (CCAT: Bennett & Parry, 2004).

Conclusions

Results from this study raise doubt about the capacity of raters to score structured competence items reliably in contexts where there is minimal opportunity for them to calibrate their scores through a training programme that aims to achieve consensus. This level of uncertainty matters less when using the scales for formative assessments, as the feedback on areas requiring improvement would still be useful. However, their use as a summative evaluation of competence is not supported without additional measures that can help to triangulate the assessment (such as extensive reliability checks, blind double-marking, moderation and external examiners and reports of direct observation from supervisors).

References

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

Blackburn, I.M., James, I.A., Milne, D.L., Baker, C., Standart, S., Garland, A., & Reichelt, F.K. (2001). The Revised Cognitive Therapy Scale (CTS-R): psychometric properties. *Behavioural and Cognitive Psychotherapy, 29,* 431–446. doi.org/10.1017/S1352465801004040

Dimidjian, S., Hollon, S.D., Dobson, K.S., Schmaling, K.B., Kohlenberg, R.J., Addis, M.E., Gallop, R., McGlinchey, J.B., Markley, D.K., Gollan, J.K., Atkins, D.C., Dunner, D.L., & Jacobson, N.S. (2006). Randomized trial of behavioral activation, cognitive therapy, and antidepressant medication in the acute treatment of adults with major depression. *Journal of Consulting and Clinical Psychology 74,* 658–670. doi: 10.1037/0022-006X.74.4.658

Elkin, I. (1999). A major dilemma in psychotherapy outcome research: Disentangling therapists from therapies. *Clinical Psychology: Science and Practice,* 6, 10-32. doi: 10.1093/clipsy.6.1.10

Jacobson, N. S., & Gortner, E. T. (2000). Can depression be de-medicalized in the 21st century: Scientific revolutions, counter-revolutions and the magnetic field of normal science. *Behaviour Research and Therapy, 38*, 103-117.

doi: 10.1016/S0005-7967(99)00029-7

Kuyken, W. and Tsivrikos, D. (2004). Therapist competence, comorbidity and Cognitive-Behavioral Therapy for Depression. *Psychotherapy and Psychosomatics*, 78, 42–48. doi: 10.1159/000172619

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

McManus, F., Rakovshik, S., Kennerley, H., Fennell, M. & Westbrook, D. (2012). An investigation of the accuracy of therapists' self-assessment of cognitive-behaviour therapy skills. *British Journal of Clinical Psychology 51*, 292–306. doi: 10.1111/j.2044-8260.2011.02028.x

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

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.org/10.1017/S1352465816000011

Roth, A.D. & Pilling, S. (2008). Using an evidence-based methodology to identify the competences required to deliver effective Cognitive and Behavioural Therapy for depression

and anxiety disorders. *Behavioural and Cognitive Psychotherapy, 36,* 129-147. doi.org/10.1017/S1352465808004141

Young, J. & Beck, A.T. (1980). Cognitive Therapy Scale: Rating manual (unpublished manuscript).

Table 1

Range of presentations

Presenting problem	Number of recordings
Depression	9
Panic Disorder	4
Phobia	3
GAD	3
OCD	3
Social Anxiety	2
Health Anxiety	1

Table 2

Therapy sessions from which recordings were rated

Session number	No of occurrences
3	2
4	6
5	1
6	7
7	4
8	3
9	1
10	1

Table 3

Intraclass correlation coefficients on the UCL Generic and CBT scales and the Cognitive Therapy Rating Scale

	ICC for all raters (95% confidence intervals)	ICC with outlier removed (95% confidence intervals)
UCL CBT scale	0.394 (0.228 – 0.598	0.476 (0.294 - 0.657)
UCL Generic Scale	0.272 (0.126 – 0.478)	0.346 (0.174 – 0.562)
CTSR	0.424 (0.260 - 0.621)	0.516 (0.339 – 0.702)

Table 4 Correlations between raters on each scale

CTSR

	Rater1	Rater2	Rater3	Rater 4	Rater 5
Rater 2	0.52				
Rater 3	0.29	0.13			
Rater 4	0.50	0.50	0.36		
Rater 5	0.52	0.61	0.12	0.59	
Rater 6	0.49	0.45	0.14	0.66	0.44

UCL Generic scale

	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5
Rater 2	0.42				
Rater 3	0.26	-0.10			
Rater 4	0.30	0.29	0.38		
Rater 5	0.40	0.57	-0.11	0.21	
Rater 6	0.49	0.40	0.31	0.59	0.15

UCL CBT scale

	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5
Rater 2	0.46				
Rater 3	0.35	0.09			
Rater 4	0.63	0.51	0.46		
Rater 5	0.61	0.58	0.25	0.54	
Rater 6	0.48	0.53	0.36	0.74	0.25