Assessing competence in CBT using structured observation

Prof Tony Roth
Pam Myles-Hooton
Dr Henry Clements
Today’s talk and workshop

• challenges of supervising
  - the argument for structured observation and supervision

• introducing the UCL competence scales
  - why a new scale?
  - what the scales look like
  - issues over reliability

• applying scales in the context of a programme

• workshop: a chance to do some live rating with the scales
Judging competence

• being a supervisor is a tricky business!
  - specially when it comes to judging competence

• competence involves the deployment of:
  - attitudes/ stance to the work
  - application of knowledge
  - application of relevant skills

• there is a lot that supervisors need to hold in mind
  - support tools for supervision - the focus of this talk and the workshop
Questions to hold in mind

• how competent does a supervisee need to be in order to judge them as competent?

• what competences are most relevant?

• how do you assess competence?

• how do you facilitate competence?

• how do you respond to unsatisfactory performance?
Structuring supervision

• what follows assumes that focused, structuring observation and supervision is more likely to be effective

• what are some of the issues that lead to this conclusion?
Some challenges facing supervisors

• clinical practice (and so supervision) is a ‘soft’ technology
  - links between what we do and what is effective are not always clear

• judgments about practice may not be as well anchored as we’d like them to be
  - and so may be subject to bias

• worth reviewing what we know about ways in which judgments can be swayed
Setting a consistent benchmark

• what are the reference points for making judgments?

• supervisors often adjust competence level to account for stage of training and/or experience
  - setting relative rather than absolute standards

• leads to variations in anchoring that relate to the person rather than their performance
Setting the benchmark – influence of client difficulty

• a challenging client can confound judgments

• evaluations of competence covary with client difficulty
  - supervisees with the most difficult clients tend to receive the lowest ratings

• it is harder for supervisees to do well if the clinical work is hard
Setting the benchmark – interpersonal issues

• relationship between supervisor and supervisee
  - trainees who rate supervisors as being more interpersonally satisfactory were in turn rated as more effective by supervisors

• although interpersonal effectiveness is important, it can be conflated with competence
Systematic bias

• leniency – a tendency for supervisors to under-use average and below average ‘grades’

• halo effects
  - which can be positive and negative
  - judgments across clients/ settings are consistent (though not necessarily accurate)
Tensions between facilitation and ‘gatekeeping’

• supervisors are only human
  - supervisee failure is experienced as supervisor failure
  - need to be liked vs. need to be candid
  - want to act later rather than sooner

• feedback is less accurate than it could/should be

• failure maybe not as common as it might be
Supervisees aren't always open with supervisors

Ladany

• 97% of trainees report withholding information at least once

• some areas related to their relationship to supervisor, but some to clinical issues:
  - clinical ‘mistakes’ (44%)
  - negative reactions to clients (36%)
  - counter transference (22%)
  - attraction to client (9%)
With the best of intentions, reportage is not reality

• what supervisees say they did is not the same as what they actually did

• supervisees don’t always know what they don't know, and so can’t report it

• memory is fallible and ‘reconstruction’ subject to bias (both intentional and unintentional)
Structured observation keeps things focused

- evidence is that supervision is more effective if:
  - focuses on specific learning tasks
  - comments on/ validates specific activities rather than global performance
  - asks specific questions and systematically reviews core concepts: "what were you aiming to achieve when you asked that?"
  - focuses on moment-to-moment aspects of the therapy rather than abstract issues
Using structured observation

- observation of live clinical material that is structured and anchored is less prone to bias

- and more likely to reflect what actually happens, rather than what is reported
Structured observation

- structure identifies the ‘lie of the land’ - reminds supervisees and supervisors
  - what areas/ techniques should be present
  - what might have been left out
  - how these should be ‘delivered’
Using scales to structure observation

• scales are best seen as ‘support-tools’
  - used to link a model of therapy to its practise

• used by supervisor and supervisee to structure
  - the observation of material
  - discussion of sessions
What scales to use?

- Cognitive Therapy Scale - Revised (CTS-R)
  - widely used
  - seen as setting a standard for competence
  - many people trained in its use

- but…questions have been raised about the CTS-R

- development of UCL CBT competence framework raised issues about scope of the CTS-R
Limitations of the CTS-R

• limited specification of change methods

• no specification of specific techniques in relation to disorder

• no provision for the appropriate absence of an area of skill

• rating from ‘novice’ to ‘expert’
  - limits the range of endorsement across the scale

• evidence of poor reliability across groups of raters (as opposed to within them)
Basis for a new scale: the UCL CBT competence framework
www.ucl.ac.uk/CORE/

- commissioned by the Improving Access to Psychological Therapies (IAPT) programme

- aimed to describe the skills and knowledge needed to deliver good-quality CBT
  - and so specify a curriculum for training

- original framework focused only on anxiety and depression
  - but subsequent frameworks have expanded range of conditions and clinical populations
Areas of knowledge and skills the framework
www.ucl.ac.uk/CORE/

• generic competences
  - used in all psychological therapies

• basic CBT skills
  - used in all CBT interventions

• specific CBT skills
  - used in some, but not all, CBT interventions

• specific evidence-based approaches for specific presentations

• metacompetences
  - procedural rules used to make judgments about when, whether and how to implement technique
Things we wanted to get right

• a scale that is easy to comprehend

• behavioural anchoring
  - based on the content of the competence framework

• specification of change methods

• specification of specific techniques in relation to disorder

• allow for the appropriate absence of an area of skill
Two scales

• Generic scale
  - generic therapeutic skills, applicable to most therapies

• CBT scale
  - focused on CBT skills
<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Establishing the context for the intervention (in initial session(s))</td>
</tr>
<tr>
<td>2</td>
<td>Appropriate non-verbal behaviour</td>
</tr>
<tr>
<td>3</td>
<td>Working with difference* [where significant areas of difference are apparent and/or where the client raises difference as an issue]</td>
</tr>
<tr>
<td>4</td>
<td>Structure and pacing</td>
</tr>
<tr>
<td>5</td>
<td>Active listening and empathy</td>
</tr>
<tr>
<td>6</td>
<td>Undertaking a generic initial assessment</td>
</tr>
<tr>
<td>7</td>
<td>Communicating a formulation</td>
</tr>
<tr>
<td>8</td>
<td>Discussing the intervention plan</td>
</tr>
<tr>
<td>9</td>
<td>Responding to emotional content</td>
</tr>
<tr>
<td>10</td>
<td>Collaboration</td>
</tr>
<tr>
<td>11</td>
<td>Developing and fostering the therapeutic alliance</td>
</tr>
<tr>
<td>13</td>
<td>Using measures</td>
</tr>
<tr>
<td>14</td>
<td>Ending the session</td>
</tr>
</tbody>
</table>
# CBT scale

**Section 1: Underpinning CBT techniques**

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Agenda setting and structuring sessions</td>
</tr>
<tr>
<td>2</td>
<td>Using summaries and feedback</td>
</tr>
<tr>
<td>3</td>
<td>Guided discovery and Socratic questioning</td>
</tr>
<tr>
<td>4</td>
<td>Identifying maintenance cycles (i.e. factors that feed into one another so as to maintain difficulties)</td>
</tr>
<tr>
<td>5</td>
<td>Sharing a longitudinal CBT formulation</td>
</tr>
</tbody>
</table>
# Section 2: Change techniques based on discussion & experiential methods

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>Using a thought record</td>
</tr>
<tr>
<td>7</td>
<td>Working with safety behaviours</td>
</tr>
<tr>
<td>8</td>
<td>Detecting, examining and helping clients reality test automatic thoughts and images</td>
</tr>
<tr>
<td>9</td>
<td>Identifying and modifying assumptions</td>
</tr>
<tr>
<td>10</td>
<td>Working with beliefs</td>
</tr>
<tr>
<td>11</td>
<td>Working with imagery</td>
</tr>
<tr>
<td>12</td>
<td>Planning and reviewing practice assignments</td>
</tr>
<tr>
<td>13</td>
<td>Planning and conducting behavioural experiments (designed to test a prediction)</td>
</tr>
<tr>
<td>14</td>
<td>Activity monitoring and scheduling</td>
</tr>
<tr>
<td>15</td>
<td>Problem solving (as an explicit strategy)</td>
</tr>
<tr>
<td>16</td>
<td>Conducting exposure (planned, repeated, prolonged exposures to the same situation(s))</td>
</tr>
<tr>
<td>17</td>
<td>Working with endings</td>
</tr>
</tbody>
</table>
### Section 3: Change techniques deployed for specific conditions

<table>
<thead>
<tr>
<th></th>
<th>Specific change techniques for working with panic</th>
</tr>
</thead>
<tbody>
<tr>
<td>18</td>
<td>Specific change techniques for working with GAD</td>
</tr>
<tr>
<td>19</td>
<td>Specific change techniques for working with OCD</td>
</tr>
<tr>
<td>20</td>
<td>Specific change techniques for working with social anxiety</td>
</tr>
<tr>
<td>21</td>
<td>Specific change techniques for working with trauma</td>
</tr>
</tbody>
</table>
Section 4: Considering the session as a whole:

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>23</td>
<td>Implementing CBT using a collaborative approach</td>
</tr>
<tr>
<td>24</td>
<td>Using measures</td>
</tr>
<tr>
<td>25</td>
<td>Using change techniques appropriate to the client’s presentation and problems</td>
</tr>
<tr>
<td>26</td>
<td>Metacompetences</td>
</tr>
</tbody>
</table>
## Anchoring

<table>
<thead>
<tr>
<th>Agenda setting and structuring sessions</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Does the therapist share responsibility for session structure and content with the client, by negotiating an explicit agenda?</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Does the therapist structure and pace the session in relation to an agenda, holding in mind the client’s needs and learning speed?</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Does the therapist strike the right balance between maintaining structure and being flexible in response to session material that emerges?</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
# The rating scale

<table>
<thead>
<tr>
<th></th>
<th>Competence not demonstrated or requires major development</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Relevant technique or process not present, but should be</td>
</tr>
<tr>
<td></td>
<td>Relevant technique or process barely present and/or applied in a manner that is ineffective*</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>2</th>
<th>Competence only partially and/or poorly demonstrated and requires significant development</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Only some aspects of technique apparent, and/or applied in a manner that is only marginally effective*</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>3</th>
<th>Competence demonstrated but requires further development</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Relevant technique present but delivered in a manner that is partial and so not as effective* as it could be, with a number of aspects requiring development</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>4</th>
<th>Competence demonstrated well but requires some specific development</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Relevant technique or process applied well and delivered in a manner that is effective*; however some specific (but not critical) areas for development</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>5</th>
<th>Competence demonstrated very well and requires no substantive development</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Relevant technique or process is applied fluently and coherently, in a manner that is demonstrably effective*</td>
</tr>
</tbody>
</table>
Using the scale - adherence and competence

• therapy is more than procedure – *how* it is done as much as *what* is done
  - being adherent to a model is not the same as being competent

• competence is doing the right thing in the right way

• rote application of technique in the absence of broader clinical judgment may be ineffective at best (or harmful at worst)
How reliable are the scales when used in routine practice?

- reliability is less of an issue if the scales are being used formatively

- but an important issue where the scales are used ‘summatively’ - making judgments about:
  - how well an individual has performed
  - whether they are above or below a threshold/benchmark level

- reliability is a test of whether independent raters agree/disagree about the ratings for a session
How reliable are the scales when used in routine practice?

• evidence is that rating scales can achieve very high reliability *within* a cohort of raters who work closely together
  - but also evidence of low reliability *between* groups of raters

• in routine practice, raters will confer minimally
  - examining reliability under these circumstances gives a better estimate of inherent reliability
The reliability study

• 14 therapists
  - ‘High-intensity’ CBT trainees on IAPT training programme

• 25 recordings

• 6 raters, working independently
  - restricted opportunities for conferring/consensus meetings

• every recording rated on:
  - CTS-R
  - UCL Generic and CBT scales
### How the scales fare – reliability

- modest inter-rater reliability on both CTS-R and UCL scales

<table>
<thead>
<tr>
<th>Scale</th>
<th>ICC for all raters (95% confidence intervals)</th>
<th>ICC with outlier removed (95% confidence intervals)</th>
</tr>
</thead>
<tbody>
<tr>
<td>UCL Generic Scale</td>
<td>0.272 (0.126 – 0.478)</td>
<td>0.346 (0.174 – 0.562)</td>
</tr>
<tr>
<td>UCL CBT scale</td>
<td>0.394 (0.228 – 0.598)</td>
<td>0.476 (0.294 - 0.657)</td>
</tr>
<tr>
<td>CTS-R</td>
<td>0.424 (0.260 - 0.621)</td>
<td>0.516 (0.339 – 0.702)</td>
</tr>
</tbody>
</table>
Qualitative review of reliability

• in addition to making a rating, raters explained the rationale for their scores

• enables us to examine some of the reasons for variation
Understandable unreliability: Variation within sessions

- within a session competence may vary
  - a specific skill could be applied well or poorly at different points

- raters:
  - awarded an averaged score
  - rated in line with the best examples
  - rated in line with the poorest examples
Understandable unreliability: Balancing structure and content

- good quality CBT requires attention to both:
  - structure (how something is set up)
  - content (identifying and working with material that is salient)

- therapists sometimes employed a technique (such as 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

- some raters:
  - awarded a low rating (the content was misjudged)
  - awarded a high rating (the therapist set-up the technique in a skilful manner)
Understandable unreliability: Missed themes

• how raters ‘read’ a session can be critical

• sometimes significant clinical themes are noticed by some (but not all) raters

• in this context raters appraise the use of specific techniques differently
  - because the context alters their appraisal of the relevance of certain techniques
What do we conclude?

- reliability of scales may be overestimated
  - especially *between* groups of raters

- ‘unreliability’ may reflect the complexity of the task

- if using scales summatively, caution is appropriate
  - triangulation of assessment methods makes sense
Using scales and structured observation

Pam Myles-Hooton
Structured Observation in Supervision

- Using the whole scale
  - Full length recordings
  - Self-rated
  - Feedback from supervisor and peers (with or without scores)
  - Seen as helpful by students

- Using part of a scale
  - Middle third (see Weck et al, 2014)
  - Short section of session recording
  - Generic/therapy specific, (e.g. active listening and empathy/setting up a behavioural experiment)
  - Group or individual setting
  - With or without ratings – with ratings may be less reliable for short sections
  - Popular with supervisees
When marks matter

What can we do to help us be confident in formal marks of students’ recordings?

• Marker training

• Inter-rater checks

• Moderation

• Double marking

• External Examining
Markers

Ensure markers are high calibre:

- Experienced clinicians
- Expert knowledge of CBT
- Experienced supervisors
- BABCP accredited
Marker Training

At the beginning of every academic year:

• All markers rate the same recording independently

• Come together to discuss marks and feedback

• Calibrate to agree a final mark and relevant feedback
  - If there is an errant marker who cannot calibrate, do not use
Inter-Rater Checks

For each video recording submission:

• All markers receive the same recording prior to marking their allocation

• Each marker rates independently and returns mark sheet

• The Moderator (usually Course Director/person who led the marker training)
  - reviews the recording and all mark sheets
  - decides on the final mark and feedback for student
  - sends anonymised mark sheets to markers with the final moderated mark sheet
  - provides individualised feedback to each marker with advice on how the marker should revise their marking for their remaining recordings
  - Any errant marking – further discussion
Moderation

- The Moderator reviews the marks and feedback for:
  - All fails (below 50%)
  - All borderline passes (50-52%)
  - All high marks (68%+)
  - A sample from each marker if do not fall in the above
  - Any recordings that the marker flags as problematic (e.g. unsure of how to mark certain items, fitness to practice etc)
  - Any marks that seem odd or out of kilter from what we know about previous performance

- Any proposed changes to marks are discussed and agreed with the marker (if agreement cannot be reached – blind double marked – likely to be close to marker or moderator – if still an issue, send to External Examiner)
Double Marking

• Any fail on second attempt (which would result in a student failing the programme) is blind double marked and moderated then sent to the External Examiner for ratification
External Examiner

• Plays an important role in confirming that marking is fair

• Sent a sample of all coursework and mark sheets including at least one piece of coursework for each student

• Reviews all fails on second attempt
Who makes the best markers?

- Supervisors who know the student and the cases or independent markers who know neither?

- The supervisor has a context for the student and the case which can be helpful BUT risk of halo effect or opposite!
  - May be best placed for formative marking.
  - Should undertake marker training.
  - Should engage in regular supervision of supervision.

- The independent marker is neutral and therefore no halo effect BUT important context can be lost.
  - Mitigate by requiring the student to record a short (up to 3 min) preamble and a short post amble to the recording along with a case report of the case to provide context.
Despite best efforts…

We can still end up with:

• Repeat offenders of errant marking
  - Further training
  - Close monitoring/moderation
  - If all else fails, drop

• Decisions that seem out of kilter with what we know about the student from past performance/supervision reports
  - Always check: moderate
Best practice

• Multiple recordings reviewed
  - Around 19 needed (Keen & Freeston, 2008)
  - Approx 50 samples reviewed in course supervision

• Random selection of recordings
  - Likely to be the best gauge of students’ routine performance
  - Deemed to be the most terrifying by students!
Using the scales outside training

- Recommend students continue using observation tools post training.

- Can be difficult for supervisors to judge if scores are appropriate without training – can be easier to use for feedback rather than scoring.

- Supervisor training recommended in using the competence scale.

- Training to include how to give feedback, including being prepared to ‘fail’.

- Supervision of supervision to include review of recordings.
In conclusion

• As we know that reliability of scales may be overestimated:

  - Training for marker and safety measures as described here should be in place when marking summatively.

  - Supervisors may be best placed to provide formative feedback and are likely to benefit from training in using the measure and supervision of supervision.