Behavioural and Cognitive Psychotherapy, 2001, 29, 431–446
Cambridge University Press. Printed in the United Kingdom

THE REVISED COGNITIVE THERAPY SCALE (CTS-R): PSYCHOMETRIC PROPERTIES

Ivy-Marie Blackburn, Ian A. James, Derek L. Milne, Chris Baker, Sally Standart, Anne Garland & F. Katharina Reichelt

Newcastle Cognitive and Behavioural Therapies Centre, U.K.

Abstract. The existing scale for assessing competence in cognitive therapy (CTS) dates from 1988 and only the previous version of 1980 has been validated to any extent. A revised version, the CTS-R, was devised to improve on the CTS by: eliminating overlap between items, improving on the scaling system, and defining items more clearly. Kolb's well-known educational model was used as a guideline. In the new 14-item scale, three new items measure general therapeutic flair, the facilitation of emotional expression, and therapist's non-verbal behaviours (optional). We hypothesized that the CTS-R would prove more user friendly and demonstrate satisfactory reliability and validity. Twenty-one mental health professionals undergoing training in cognitive therapy provided 102 video-tapes of therapy with 34 patients, reflecting three stages of therapy. The tapes were rated by four expert raters, in a balanced design. The CTS-R showed high internal consistency and adequate average inter-rater reliability. Reliability for individual items varied widely among pairs of raters. Validity was demonstrated by improved ratings of competence for trainees who saw patients early and later during the course of training. Although raters found the CTS-R a more useful tool than the CTS and satisfactory reliability and validity were demonstrated, more refinement is needed in item definition. The study has led to modifications in the CTS-R, which are in the process of evaluation.

Keywords: Cognitive therapy scale, psychometrics, competence.

Introduction

Background

The growing demand for short-term, evidence based psychotherapies (Roth & Fonagy, 1996; Stern, 1993) means that the need for carefully designed training programmes in effective therapeutic methods has greatly increased recently (The Sainsbury Centre, 1997). Cognitive Therapy (CT) has been shown to be effective for a number of psychological disorders, in particular major depression, general anxiety, panic disorder, obsessive-compulsive disorder and bulimia (Blackburn, Twaddle, & Associates, 1996). A growing body of evidence is also accumulating for the efficacy of CT in several other disorders (for example, schizophrenia). It is therefore not surprising that demand for training in cognitive therapy continues to

Reprint requests to I.-M. Blackburn, Newcastle Cognitive and Behavioural Therapies Centre, Plummer Court, Carliol Place, Newcastle upon Tyne NE1 6UR, U.K.

increase (Freiheit & Overholser, 1997). In order to demonstrate the effectiveness of training, a measure of competence is required that assesses adherence to CT methods and levels of skilfulness in the application of these methods.

Published controlled trials of CT usually utilize highly trained cognitive therapists, for example the NIMH study of depression (Elkin et al., 1989), for which Shaw and Wilson-Smith (1988) described the process of training and monitoring of competence. Time in training for 10 therapists varied from 13 to 18.5 months and, even so, two therapists were not considered to have reached an adequate level of competence, as rated by the Cognitive Therapy Scale (CTS, Young & Beck, 1980). Studies of the relationship between level of experience in psychotherapy and outcome of therapy have not produced consistent results (Berman & Norton, 1985; Hattie, Sharpley, & Rogers, 1984; Orlinsky, Grawe, & Parks, 1994; Shapiro et al., 1994). However, there is some evidence that higher levels of competence in CT predict better outcome (Beckham, 1990; Burns & Nolen Hoeksema, 1992; Hollon, Shelton, & Davis, 1993; Shaw et al., 1999). It is reasonable to assume that a certain level of competence is necessary to achieve results in the clinic that are comparable to those found in the controlled trials. This relates to the effectiveness versus efficacy controversy in psychotherapy. Several studies have found that the Effect Size in outcomes in clinical studies is usually smaller than in randomized controlled trials (Weisz, Weiss, & Ham, 1995). This difference has been attributed to several possible causes, among others, the competence of the therapists and their adherence to the treatment protocol. Therefore, two key tasks for an empirically validated therapy such as CT are to establish how best to train cognitive therapists and how best to measure competence.

Several courses in CT have been set up to train therapists from varying mental health disciplines. These courses tend to vary in length and intensity. In this paper, we do not directly address the question of the effectiveness of different types of training, but rather we examine the measurement of competencies in the context of a well-established course of one-year duration, which has run successfully for 7 years (Milne, Baker, Blackburn, James, & Reichalt, 1999).

The Cognitive Therapy Scale (CTS)

The CTS is a widely used scale for measuring competence in CT. It was developed at the Cognitive Therapy Centre in Philadelphia (Young & Beck, 1980, 1988). The original scale consisted of 11 items and has been the one whose psychometric properties have been tested (Dobson, Shaw, & Vallis, 1985; Vallis, Shaw, & McCabe, 1988; Vallis, Shaw, & Dobson, 1986; Shaw et al., 1999). The most recent scale consists of 13 items and uses the same 7-point scale (0–6), with definitions at alternate points. The items are listed in three sections: *General interview procedure* (1. agenda setting; 2. eliciting feedback; 3. collaboration; 4. pacing and efficient use of time); *Interpersonal effectiveness* (5. empathic skills; 6. interpersonal effectiveness; 7. professionalism); and *Specific cognitive-behavioural techniques* (8. use of guided discovery; 9. case conceptualization; 10. focus on key cognitions; 11. application of cognitive techniques; 12. application of behavioural techniques; 13. homework). The CTS has a detailed rating manual that remains unpublished.

As part of the NIMH Treatment of Depression Collaborative Research Programme, Vallis et al. (1986) reported on the psychometric properties of the CTS (11 items). Inter-rater reliability was calculated from five of the seven raters who had rated the same 10 videotapes

from the pool of 94 tapes. The intraclass correlation coefficient (ICC) for a single rater was 0.59 ($p < .01$), while reliability for single items ranged from 0.27 (pacing) to 0.59 (empiricism, now renamed guided discovery). Dobson et al., (1985) reported a study from the same NIMH research programme. Four expert raters examined 21 videotapes provided by 21 trainee cognitive therapists. Using Pearson product moment correlations, they found an inter-rater correlation of 0.94 ($p < .001$) for total scores, while inter-rater reliability for individual items ranged from 0.54 (feedback, $p < .05$) to 0.87 (application of cognitive-behavioural techniques, $p < .001$). Both Vallis et al. (1986) and Dobson et al. (1985) found a range of competence among therapists. Both studies found the CTS to be internally reliable ranging from 0.59 (agenda) to 0.90 (implementation of strategy) in Vallis et al. 91986) and from 0.58 (questioning) to 0.92 (pacing) in Dobson et al. (1985), except for one item (homework), which obtained a correlation of 0.07 with total score. Shaw et al. (1999) found limited support for the relationship of level of competence, as measured by the CTS, and outcome in the NIMH depression study (Elkin et al., 1989). The CTS accounted for 15% of the variance in post-treatment level of depression on an observer rating scale (Hamilton Rating Scale for Depression, HRSD; Hamilton, 1960), after controlling for the effect of therapist adherence and facilitative conditions. No effect was found on self-rated measures. They found the Structure subscale of the CTS to be the best predictor of outcome.

Conceptual and practical problems with the CTS

Although the above work was promising, there are several problems with the CTS. First, the original 11-item version (Young & Beck, 1980) has been superseded by a 13-item version (Young & Beck, 1988) and the latter version has not been evaluated. Second, our own experience, as expert raters (and that of others as reported to us), indicates that ratings on the CTS do not discriminate well between different levels of competence. The different points on the scale are not adequately defined (only alternate points are defined) and require varying degrees of inference from the raters. Whisman (1993) argued that the CTS includes items that aggregate different constructs and he also commented on the high degree of inference needed by the raters. Shaw et al. (1999) agreed with these criticisms and added that some important aspects of CT competence may not be tapped by the CTS. Third, we considered that some items show major overlap, in particular in the General Interview Procedure, and that CT, as it is practised now, emphasizes the role of emotion far more than the CTS implies. Educationalists, for example Kolb (1984), have stressed the importance of the facilitation of emotional expression in learning. CT is an educational model and although the first CT manual (Beck, Rush, Shaw, & Emery, 1979) stresses the importance of emotions in therapy, this aspect seems to have been underplayed later. For example, Beck et al. (1979, p. 36) state: ‘‘In fact, since an essential part of the cognitive therapy of depression is to establish the connection between an unpleasant emotion and the antecedent cognitions or the prevailing attitudes, it is obviously essential to focus on and discriminate the patient’s emotional reaction’’. Recent developments in Cognitive Therapy, for example Teasdale (1996) in his interactive cognitive subsystems (ICS) model and Young (1990), stress the importance of eliciting, labelling and working with emotions, if change is to occur. The vast accumulating literature on the importance of the therapeutic alliance (for example, Raue & Goldfried, 1994), also emphasized by Beck et al. (1979), prompted us to review the relevant items from the General Interview Procedures of the CTS.

Methodological issues

There are numerous methodological problems relating to the measurement of competence in psychotherapy. First is the source of the material for ratings (audiotapes, videotapes, transcripts of supervisees' accounts of therapy). In this study, we opted for video-tapes as more accurate accounts of therapists' verbal and non-verbal behaviour. Second, is the decision to rate parts of tapes or whole tapes. We decided to rate whole tapes, again for more accuracy. Third is the problem of what criteria are used for comparison. Since CT is a well-defined mode of treatment, desirable therapist's behaviours have been extensively described, at least in the case of depression and anxiety (for example, Becket al., 1979; the unpublished manual of the CTS, Young & Beck, 1980; Blackburn & Davidson, 1995).

The fourth problem is the type of scales used to measure competence. Scales can be specific, measuring competence in a specific therapy, or general, measuring psychotherapeutic competence common to all therapies. It could also be argued that different scales should be devised to measure competence at different stages of therapy. Ratings could consist of a checklist of specific behaviours, rated as present or absent, or could measure the frequency and extensiveness of interventions, taking into consideration therapist errors. The former would be suitable for rating adherence to a particular therapy protocol. The CTS measures competence, while adherence is measured by a different scale devised for the NIMH study, the Collaborative Study Psychotherapy Rating Scale (CSPRS; Hollon et al., 1988). In the CTS-R, we combine the measurement of adherence and competence. The CTS-R represents a specific scale, which nonetheless includes general therapeutic skills. The same scale is used across different stages of therapy, using a checklist of specified behaviours which are rated for presence, frequency, extensiveness and skill. Finally, competence may vary according to context variables, for example the type of presenting problems and characteristics of the patient, which may be taken into consideration in the measurement scales. In the CTS-R, we attempt to take into consideration therapist errors and context variables.

Our specific aims were to:

1. Use a rating system that is defined at every point of the scale and is based on existing scales of skills acquisition.
2. Develop the CTS on the principles that competence consists not only of adherence to prescribed CT methods and of skilfulness in the application of these methods, but also of a pan-theoretical component, which is the therapeutic alliance.
3. Assess the psychometric status of the CTS-R.

Method

Revision of the CTS

Two independent raters, a clinical psychologist and an educationalist, both not CT practitioners, rated videotapes of therapy sessions by expert cognitive therapists to arrive at a definition of competence from a pantheoretical viewpoint. They used Kolb's (1984) model of experiential learning and Dreyfus's (1989) model of skill acquisition (see Milne, Claydon, Blackburn, James, & Sheikh, 2001, for details regarding Kolb's and Dreyfus's models) to arrive at the following suggestions: the scales were arbitrary and could be improved by

Table 1. Items of the CTS-R compared with the CTS

CTS-R	CTS
1. Agenda setting	1. Agenda setting
2. Feedback	2. Eliciting feedback
3. Collaboration	3. Collaboration
4. Pacing and efficient use of time	4. Pacing and efficient use of time
5. Interpersonal effectiveness	5. [+]Emphatic skills
6. *Charisma/flair	6. [+]Interpersonal efficiency
7. *Facilitation of emotional expression	7. [+]Professionalism
8. Guided discovery	8. Guided discovery
9. Conceptualization	9. Conceptualization
10. Identifying key cognitions	10. **Focus on key cognitions
11. Application of cognitive change methods	11. Application of cognitive techniques
12. Application of behavioural techniques	12. Application of behavioural techniques
13. Use of homework	13. Use of homework
14. *Non-verbal behaviour [appropriate eye-contact; expressive facial communication; expressive body movements; appropriate posture; uses humour appropriately; appropriate tone of voice; appropriate volume of voice; positioning of self and patient; appropriate silences; clarity of speech; facilitatory grunts and noises; professional demeanour (dress); professional demeanour (language)]	

* New items; [+] Collapsed into item 5 on CTS-R; ** Modified to new item 10, because of overlap with item 11.

using the NVCQ model of ratings of skills acquisition from ''novice'' to ''competent''; and not enough emphasis was put on elicitation and use of emotions in therapy.

Four expert cognitive therapists (all tutors and supervisors on the Newcastle Post-Qualification Certificate in Cognitive Therapy Course), with extensive experience of the CTS, then met on several occasions to revise the CTS (Young & Beck, 1988) in the light of their experience with that scale and to incorporate the recommendations of the independent raters. The CTS-R consists of 14 items, most of which have been retained from the CTS, as shown in Table 1.

Changes include three new items: charisma/flair was added in an attempt to measure an aspect of the therapeutic alliance that is recognizable in therapists but difficult to operationalize. Facilitation of emotional expression has been discussed above. Item 14, regarding non-verbal behaviours of the therapist, is optional, as it can only be rated from video-tapes. This item includes 13 desirable aspects of therapist's non-verbal behaviours rated as present or absent. These were intended to reflect ethological methods of observation, in order to operationalize what might be an important component of the facilitative component of the therapeutic alliance. An attempt to get rid of overlap in the General Therapy section was made by collapsing items 5, 6 and 7 of the CTS into one new item (interpersonal effectiveness). The CTS item 5, empathic skills, measuring ability to understand the patient's internal reality, seemed to us closely related to item 6, interpersonal effectiveness, measuring

warmth, concern and genuineness, in that one could not probably obtain a high score on one and a low score on the other. We have found item 7 ''Professionalism'' not to be discriminatory among mental health professionals and to be also legitimately part of ''Interpersonal Effectiveness'' as a therapist. The new item ''Interpersonal Effectiveness'' has as key features: ability to engage the patient, demonstration of empathic understanding of patient's implicit meanings, genuineness, warmth, and facilitation of disclosure. The CTS-R can be used as a 13- or 14-item scale, depending on the medium used for rating (audio- or video-tape). Results will, therefore, be reported for both, CTS-R 13 and CTS-R 14 versions. Apart from changes in the items themselves, as described above, the rating system was also modified. The 7-point scale was retained, but to increase reliability and discriminatory power, each point of the Likert scale is defined, using an adaptation of Dreyfus's (1989) 5-level distinction of skill acquisition. Both the level of competence in the execution of CT techniques and adherence to CT protocol are taken into consideration. A score of 0 indicates non-adherence to CT, and a score of 6 indicates extreme expertise in the face of difficulties. Table 2 indicates how the current scale has incorporated the Dreyfus's (1989) evaluation system, which ranges from ''incompetence'' to ''expert''. Finally, a detailed scoring manual has been developed, as an aid to raters and ratees (available from the authors).

Therapists

These were 21 therapists undergoing training in cognitive therapy at the Newcastle Cognitive and Behavioural Therapies Centre. The majority ($N = 18$) were the intake of one specific year of the one-year Post-Qualification Certificate Course in CT. Of these, there were seven men and eleven women, five psychiatrists, six clinical psychologists and seven specialist nurses or community psychiatric nurses. Additionally, there was one third-year clinical psychology trainee doing a one-year CT placement, female; one senior registrar in psychiatry undergoing specialized training in CT at the Centre, male; one senior nurse attending the training clinic in CT, male. This latter form of training is at a lower level and less intensive than the Post-Qualification Certificate Course.

Thus, of the 21 therapists, 12 were female and 9 were male. The average age was 39.0 (*SD* 7.0) with a range of 26 to 54. For the analyses, the third year clinical psychology trainee was dropped. This was done because, owing to recruitment problems, insufficient numbers of non-qualified staff were available to make sub-group comparisons possible.

Patients

Thirty-four patients received cognitive therapy from the 20 trainee therapists, some therapists seeing two patients concurrently or at different stages of training, and some only one in the course of the study. The average age of the patients was 37.0 (*SD* 12.5), with a range of 19 to 70. Twenty were married or cohabiting and 14 were single, separated or divorced. Twenty-one were in employment and 13 were unemployed. In terms of DSM-IV diagnostic groups, 16 patients suffered from depression (major or Not Otherwise Specified, NOS); 6 from social phobia; 5 from panic disorder; 4 from obsessive-compulsive disorder and 3 from generalized anxiety disorder.

Table 2. Example of item definition and scoring

Item 10—Application of cognitive change methods
Core function: Therapist skilfully uses, and helps the patient to use, appropriate cognitive techniques in line with the formulation. The therapist helps the patient devise appropriate cognitive methods to evaluate the key cognitions associated with distressing emotions, leading to major new perspectives and shifts in emotions and behaviours.

Three features need to be considered:
(i) the appropriateness and range of cognitive methods (e.g. cognitive change diaries, continua, distancing, responsibility charts, evaluating alternatives, examining pros and cons, determining meanings, imagery restructuring, etc.);
(ii) the skill in the application of the methods – however, skills such as feedback, interpersonal effectiveness, etc. should be rated separately under their appropriate items;
(iii) the suitability of the methods for the needs of the patient (i.e. neither too difficult nor complex).

NB: This item is *not* concerned with accessing or identifying thoughts, rather with their re-evaluation.

Competence level	Features
0	Therapist fails to use cognitive therapy methods.
1	Therapist applies insufficient or inappropriate methods.
2	Therapist applies either insufficient or inappropriate methods, and/or with limited skill and flexibility.
3	Therapist applies a number of methods in competent ways, although some of the interventions are incomplete.
4	Therapist applies a sufficient range of methods with skill and flexibility, enabling the patient to develop new perspectives. Some difficulties evident.
5	Therapist systematically applies an appropriate range of methods in a creative, resourceful and effective manner.
6	Excellent range and application, or successful application in the face of difficulties.

Design

All therapy sessions were video-taped and three tapes were obtained for each of the 34 patients, one at the beginning of therapy (within sessions 1–4); one towards the middle of therapy (within session 5–8); and one towards the end of therapy (within sessions 9–12). Session 1 was defined as the first session of therapy proper, after assessment and history had taken place (two sessions). Thus 102 tapes were obtained to give a relatively continuous measure, as opposed to one snapshot, view of individual competence. Training on the post-

https://doi.org/10.1017/S1352465801004040 Published online by Cambridge University Press

Table 3. Intraclass correlations for pairs of raters (N = 16–17) for total scores on CTS-R, 13 items and (14 items); Pearson's product moment correlations in italics.

Rater	Rater B	Rater C	Rater D
Rater A	.40 (.48)*	.63*** (.73)****	.86**** (.79)****
	*.40 (.53)**	*.68** (.72)***	*.87**** (.79)****
Rater B		.57*** (.34)*	.55** (.48)*
		*.61** (.52)***	*.54* (.48)*
Rater C			.64**** (.45)***
			*.72**** (.68)***

*p < .05; **p < .025; ***p < .01; ****p < .001.

qualification course includes weekly face-to-face supervision in pairs, as well as weekly didactic sessions and workshops. A proportion of the trainees saw their second patient later on in the course, providing an opportunity for comparing competency early and later in the course of training.

Each tape was double-rated by the four expert raters in a balanced design, so that raters were paired with each other an equal number of times. Each rater thus rated 51 therapy-sessions ($\frac{102\times 2}{4}$) on the CTS-R and each pair rated 17 tapes in common. Raters made their ratings independently and were not informed about the stage of therapy at which the recording was made. This attempt at keeping raters unaware of the stage of therapy was to counteract the expectation effect that therapists would become more competent as the course proceeds. However, it is appreciated that blindness could not be guaranteed, as expert raters would probably be able to guess the stage of therapy from the content of sessions.

Hypotheses

1. The CTS-R will have acceptable inter-rater reliability and adequate internal reliability.
2. Discriminant validity will be demonstrated by an increase in scores over the course of training, showing improved adherence to and skilfulness in CT methods. Only the scores of trainees who saw a second patient at a later stage in training will be compared.

Results

Reliability

Internal reliability. Cronbach's alpha coefficient was calculated for each rater for the 13-item and 14-item version of the CTS-R. For the 13-items version, alphas for the four raters were 0.92, 0.95, 0.97 and 0.95. For the 14-items version corresponding alphas were 0.75, 0.78, 0.85 and 0.86. Thus, both versions are highly internally consistent, with the 13-item version achieving higher internal reliability.

Inter-rater reliability for total scores. The four expert CT raters were paired with each other producing six inter-rater correlations for total scores, as shown in Table 3. Intraclass correlations (ICC) were calculated for 13-items total score and 14-items total score. ICC is considered to be a more suitable measure of reliability than product-moment correlation

https://doi.org/10.1017/S1352465801004040 Published online by Cambridge University Press

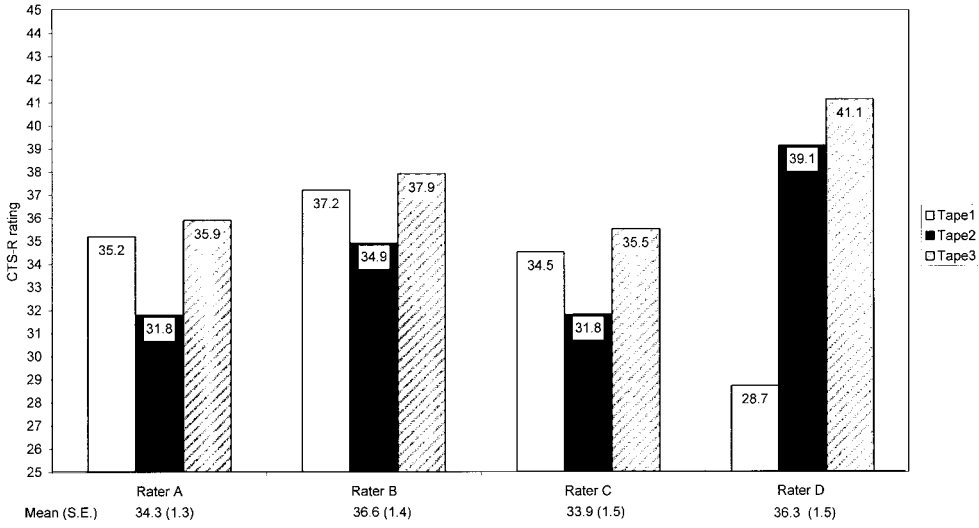

Figure 1. Average ratings of the 4 raters over time and overall mean (S.E.).

(Shrout & Fleiss, 1979; Lahey, Downey, & Saal, 1983), where several pairs of judges rate the same targets. Two-way analyses of variance were calculated for each pair of raters, which enabled the calculation of rater effects, ratee effects and their interaction. The significance levels are those of F ratios (*dfs* 15, 16; or 16, 16) of $\frac{\text{Tape effect}}{\text{Interaction of rater} \times \text{tape}}$. Full details of rater and ratee effects are reported in a complementary paper (James, Blackburn, Milne, & Reichelt, 2001).

For comparison with previous validation data (Dobson et al., 1985) on the 11 items version of the CTS, product moment correlations are reported in italics. ICCs ranged from 0.86, $p < .001$, to 0.40 (NS) for pairs of raters on the 13-item version and from 0.79, $p < .001$ to 0.34, $p < .05$ for the 14-item version. The average ICCs across raters were 0.63 (13-items) and 0.57 (14-items), both significant at $p < .01$. Rater B obtained the lowest correlations with each of the other three raters than these did with each other.

Pearson correlations ranged from 0.87, $p < .0001$ to 0.40 (NS) for the 13-items version and 0.79 ($p < .001$) to 0.48 (NS) for the 14-items version. The average Pearson correlations for four raters were 0.66 (13 items) and 0.63 (14 items), $p < .001$. For three raters, excluding Rater B, the corresponding Pearson rs were 0.77 and 0.73, $p < .0001$. Figure 1 shows the means for the four raters on each occasion of assessment and the overall means with standard errors.

Inter-rater reliability for individual items. Average product-moment correlations ranged from 0.42, *df* 49, $p < .01$ (conceptualization) to 0.67, *df* 49, $p < .001$ (guided discovery), indicating that all average correlations for single items were highly significant. However, there was a wide fluctuation between pairs of raters. All the lowest correlations were again due to rater B. Detailed analysis of individual raters indicated that rater B was on the whole more lenient than the other three raters (see Figure 1), her modal rating being 43–50 (13 items), whereas the three other raters' modal rating was 27–34.

https://doi.org/10.1017/S1352465801004040 Published online by Cambridge University Press

The ICCs between pairs of raters were also calculated and reflected the same picture. The highest ICC for a pair of raters was 0.84, $p < .01$ (Guided Discovery) and the lowest −0.14 (Collaboration).

Validity

Face validity. Validity of any new rating scale is difficult to establish and the evidence is usually circumstantial. In this case, concurrent validity could not be established as other measures (for example the Helper Behaviour Rating System, Shapiro, Barkham, & Irving, 1984) are not specific to the assessment of CT competence. Correlations with the CTS would have been desirable, but could not be done within the scope of this study.

The *face validity* of the CTS-R is good, as the expert raters all agreed that they found the scale easier and more meaningful to rate than the original CTS. All found it difficult to go back to using the CTS in their daily work, after rating 51 tapes each on the CTS-R.

Discriminant validity. If the scale is a valid measure of competence in CT, the expectations would be that trainee therapists would increase their competence over time, as they acquire more skills during the course of training. Eleven of the 20 trainees saw a second patient towards the end of training and their average scores for their first and second patient are presented in Figure 2, for the 13-items version. The mean average score for the first case was 35.1 (*SD* 7.2) and for the second case 38.9 (*SD* 5.9). Paired *t*-tests for total scores showed a significant improvement, $t = 2.68$; df 10; $p < .02$.

A two-way analysis of variance of competence on the three occasions of assessment for case 1 and case 2 produced a significant change for time ($F = 4.41$; df 2, 20; $p < .03$), a significant difference between competence for case 1 and case 2 ($F = 7.93$; df 1, 10; $p < .02$) and no interaction between time and case ($F = 0.86$; df 2, 20; $p = .44$). Thus, the 11 trainees improved their level of competence over time for both cases. The level of competence was higher for the second case, although the profiles of improvement were similar over time.

Univariate analyses indicated that the mean gain in competence was at the beginning of treatment (session 1–4), with means of 31.02, *SD* 7.27, on occasion 1 and 38.0, *SD* 7.28, on the second occasion, $t = 4.43$, df 10, $p < .001$. No significant differences were found at later stages of treatment.

Changes in competence from first tape to second tape for the 11 trainees on each CTS-R item were also examined. Table 4 shows the mean scores across raters for the 11 therapists. It can be seen that significant improvements were obtained on seven items: eliciting feedback ($p = .03$), pacing and efficient use of time ($p = .02$), interpersonal effectiveness ($p = .02$), charisma/flair ($p = .01$), guided discovery ($p = .04$), conceptualization ($p = .05$) and application of cognitive techniques ($p = .01$). Thus, only slightly more than 50% of the items showed a significant change.

This analysis indicates that improvement is not general during training, some aspects of CT being more responsive to learning. The average scores indicate that adherence was generally present (no scores of 0) and that at the end of training scores were generally at the competent level (scores of 3). When comparing individual scores for first tapes from the two cases (early therapy), second tapes (middle therapy) and third tapes (later stage of therapy), significant differences were obtained on seven items for the first tape, none on the second and one item (conceptualization) on the third tape. Thus, as for the differences in

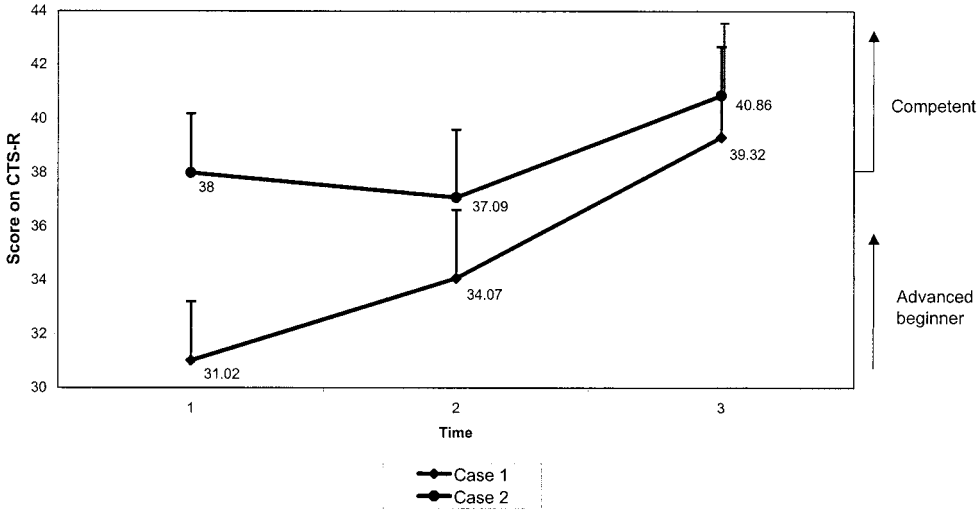

Figure 2. Means and S.E.s of levels of competence on CTS-R (13 items) for earlier (case 1) and later (case 2) cases during training

Table 4. Mean scores (*SD*s) of competence on individual items of CTS-R for the first case and second case for trainees who saw their second patient later in training ($N = 11$)

Items	Case 1	Case 2	*p*
1. Agenda setting	2.8 (0.7)	3.2 (0.4)	NS
2. Eliciting feedback	2.8 (0.5)	3.0 (0.5)	.03
3. Collaboration	3.2 (0.5)	3.3 (0.5)	NS
4. Pacing	2.6 (0.6)	3.1 (0.5)	.02
5. Interpersonal effectiveness	3.2 (0.5)	3.4 (0.5)	.02
6. Charisma & flair	2.8 (0.5)	3.2 (0.5)	.01
7. Facilitation of emotional expression	2.5 (0.5)	2.8 (0.4)	NS
8. Guided discovery	2.7 (0.6)	3.0 (0.5)	.04
9. Conceptualization	2.5 (0.7)	3.0 (0.5)	.05
10. Focus on key cognitions	2.7 (0.5)	3.0 (0.6)	.06
11. Application of cognitive technique	2.5 (0.6)	2.8 (0.6)	.01
12. Application of behavioural techniques	1.7 (0.9)	2.0 (0.8)	NS
13. Use of homework	3.0 (0.7)	3.1 (0.4)	NS

mean scores reported above, gains on individual items of the CTS-R were at the beginning of the course of treatment (seesions 1–4).

Discussion

This study is the first, since the pilot study of Williams, Moorey and Cobb (1991), to examine the acquisition of competence of cognitive therapists during formal training, not in the context of an outcome study, and involving unselected clinical patients suffering from

https://doi.org/10.1017/S1352465801004040 Published online by Cambridge University Press

a variety of clinical disorders. It was argued that the Cognitive Therapy Scale (CTS, Young & Beck, 1980, 1988) needed revision from a conceptual viewpoint and also pragmatically, because of the experience of trainers in cognitive therapy who find it difficult to discriminate between levels of competence on the existing scale. The 11-item CTS (Young & Beck, 1980) has been studied psychometrically, but the 13-item version (Young & Beck, 1988), which is now more commonly used, has not been examined for its psychometric robustness.

The CTS-R is an attempt to incorporate a well tried educational model (Milne et al., 2001) in a measure of clinical competence, to adapt a known scale of skill levels to every item of the competence scale and to revise the content and wording of each item to facilitate discriminatory rating by trainers. The underlying rationale of the CTS-R is that competence in cognitive therapy will consist in adherence to and skilfulness in cognitive therapy methods and in pan-theoretical therapeutic skills relating to the therapeutic alliance.

The results indicated that the CTS-R was internally consistent. The high internal consistency of the CTS-R, as of the CTS, may give support to the criticism that there is a high degree overlap among items (Whisman, 1993). Another possibility is that the raters are influenced by a ''halo'' effect, which leads them to rate all items similarly for individual trainees. More interestingly, the high internal consistency may reflect the way skills are learned and/or are taught in training. Trainers do not learn one aspect of cognitive therapy at a time, or at different times in training, but learn cognitive therapy skills globally; improvement in one aspect leading to changes in all other aspects. The only way this last possibility could be tested would be to teach cognitive therapy piecemeal, but this could not be done ethically as trainees treat real patients in ordinary Mental Health clinics.

Inter-rater reliability across raters was highly significant, but varied between pairs of raters, ranging from highly significant to non-significant levels. One of the raters tended to have a different pattern of rating from the other three, obtaining higher modal scores. When average correlations were calculated for three raters, excluding that latter rater, higher levels of average reliability were obtained. The difference among raters indicates that even with more detailed definitions of each level of rating, raters still have to make inferences, albeit perhaps less arbitrarily than on the CTS. The average ICCs in this study for 13 items are somewhat superior to that of Vallis et al. (1986) and for 14 items, they are nearly identical. These results indicate that high inter-rater reliability is difficult to achieve, even among expert therapists.

The average reliability of individual items varied from highly significant (Guided Discovery) to significant (Conceptualization) and is comparable with the results of Dobson et al. (1985), who reported average inter-rater reliability of 0.54 (feedback) to 0.87 (application of behavioural techniques). However, correlations for pairs of raters varied widely, with the one rater showing again less agreement with her peers. This variability in item reliability is worrying. Our results indicate that further work in item definition and in refining of the scale is necessary. Above all, however, more training of raters is indicated if a better consensus is to be expected. In a recent exercise to test the effect of training on raters using the CTS-R, correlations for 10 raters on 24 videotapes increased from 0.44 (p = .05) to 0.67 (p = .001), these correlations differing at the .01 level.

Hypothesis 1 was supported regarding the internal reliability of the CTS-R and the average inter-rater reliability for total score and individual items, but not with regard to the reliability of individual items for single pairs of raters. From a methodological view point,

https://doi.org/10.1017/S1352465801004040 Published online by Cambridge University Press

it has to be noted that the four expert raters were also trainers and supervisors on the course. This creates a contamination factor, as they had prior knowledge of the trainees. This situation could have been avoided by recruiting expert raters from another centre. However, apart from the practical problem in doing the latter, although prior knowledge of the trainees could have affected the level of ratings (say, for example, if a trainee was known to be good at essay writing), it would not have influenced the inter-rater reliability of the ratings.

A comment is also necessary on the fact that each trainee presented several tapes and 11 of the 20 trainees presented two tapes at each stage of assessment. These tapes could be seen as not independent items, thus affecting the reliability of the ratings. Several safeguards in the methodology temper this potential weakness; the tapes were double rated in a balanced design by two out of four raters, so that the same trainee was not always rated by the same raters; the raters were unaware of the stage of treatment; the trainees did not necessarily submit tapes of the same patients; and the raters, as they rated 50–51 tapes each, could not keep track of the ratings that they had given on individual tapes, unless they kept a record and checked. This was strictly proscribed, the raters being instructed to rate each tape as a separate entity.

The discriminant validity of the CTS-R was demonstrated by the scores of trainees increasing significantly from the beginning to later stages of training. As no average score of 0–1 was obtained, adherence to the cognitive therapy protocol was demonstrated. Improvement over time for individual items indicated that not all aspects of CT were equally amenable to learning. About 50% of the items showed a significant improvement. This increase in competence was seen particularly on competence levels achieved at the beginning of treatment (sessions 1–4). While improvements at later stages of treatment were achieved, these did not reach significance. Examination of individual scores at the three stages of treatment for the two cases also indicated that significant gains were achieved on several items of the CTS-R at the beginning of treatment and not at the later two stages, except for ‘‘conceptualization’’. These findings and the significant improvement in competence over time for both the early and later case indicate that increase in competence in both instances is likely to be a confound effect of stage of training and stage of treatment. As trainees get to know their patients better, they are more likely to show more competence.

Thus hypothesis 2, that discriminant validity of the CTS-R would be demonstrated by improved competence over the course of one academic year, was supported. While adherence was generally present, increase in competence was not found on all aspects of CT and levels improved from advanced beginner to competent level. Three non-mutually exclusive conclusions may be derived from these results: a one-year training in cognitive therapy, even for trainees with at least 2 years post-qualification experience, is not sufficient to reach proficient levels in competence; trainers need to pay more attention to certain aspects of training, and/or the CTS-R needs further refinements to increase its discriminatory power.

In conclusion, raters were satisfied with the face and conceptual validity of the CTS-R. Although adequate levels of internal consistency and of global inter-rater reliability were found, the variability in inter-rater reliability, for individual items, indicated that further work needs to be done on the scale, both in terms of clearer definitions of items and in discriminating between different points on the rating scale. The results also indicated that even expert raters need careful training in their use of the scale in order to increase consensus and harmonization. Significant improvements can be obtained over the course of a one-year training course, but it is evident that further experience should be sought through

expert supervision. The results reported in this study may be of help to trainers in pinpointing typical weaknesses in trainees and where more training may be needed. The final version of the CTS-R can be obtained from the authors. The item rating non-verbal behaviours has now been dropped, because of its non-generalizability to audio-tape ratings. The item rating charisma/flair has also been dropped, as it was introduced as a possible test of innate therapist general qualities. This was not demonstrated, as scores changed significantly with training.

Although it is evident that further work is needed, we believe that the CTS-R is an improvement on the CTS in its conceptual rationale and rating format. Barkham et al. (1998) recommended three discrete stages for the development of core outcome batteries. They call these Design, Implementation and Yield (the DIY model). Applying this model to the CTS-R as a measure of training outcome in cognitive therapy, the ''Design'' stage of development has been completed. The CTS-R has respectable reliability; it is sensitive to change; easy to score manually or by computer; and it can detect varying levels of skill. The ''Implementation'' stage is in progress, in that there is an existing manual, the scale is being refined and other services will be contributing to the data. The ''Yield'' stage, that is the use of the CTS-R as ''benchmarking'', can then follow.

Acknowledgements

This study was funded by the Mental Health Foundation, Grant No. PR1 595/3.

References

Barkham, M., Evans, C., Margison, F., McGrath, G., Mellor-Clark, J., Milne, D., & Connell, J. (1998). The rationale for developing and implementing core outcome batteries for routine use in service settings and psychotherapy outcome research. *Journal of Mental Health, 7*, 35–47.

Beck, A. T., Rush, A. J., Shaw, B. F., & Emery, G. (1979). *Cognitive therapy of depression: A treatment manual*. New York: Guilford Press.

Beckham, E. E. (1990). Psychotherapy of depression. Research at a cross-roads. *Clinical Psychology Review, 10*, 207–228.

Berman, J. S., & Norton, N. C. (1985). Does professional training make a therapist more effective? *Psychological Bulletin, 89*, 401–402.

Blackburn, I. M., & Davidson, K. M. (1995). *Cognitive therapy for depression and anxiety. A practitioner's guide*. Oxford: Blackwell Science.

Blackburn, I.-M., Twaddle, V., & Associates (1996). *Cognitive therapy in action: A practitioner's casebook*. London: Souvenir Press.

Burns, D. D., & Nolen-Hoeksema, S. (1992). Therapeutic empathy and recovery from depression in cognitive behavioural therapy: A structural equation model. *Journal of Consulting and Clinical Psychology, 60*, 441–449.

Dobson, K. S., Shaw, B. F., & Vallis, T. M. (1985). Reliability of a measure of the quality of cognitive therapy. *British Journal of Clinical Psychology, 24*, 295–300.

Dreyfus, H. L. (1989). The Dreyfus model of skill acquisition. In J. Burke (Ed.), *Competencey based education and training*. London: Falmer Press.

Elkin, I., Shea, M. T., Watkins, J. T., Imber, S. D., Sotsky, S. M., Collins, J. F., Glass, D. R., Pilkoniz, P. A., Leber, W. R., Docherty, J. P., Fiester, S. J., & Parloff, M. B. (1989). NIMH treatment of depression collaborative research program: General effectiveness of treatments. *Archives of General Psychiatry, 46*, 971–982.

Freiheit, S. R., & Overholser, J. C. (1997). Training issues in cognitive-behavioural psychotherapy. *Journal of Behaviour Therapy and Experimental Psychiatry*, *28*, 79–86.

Hamilton, M. A. (1960). A rating scale for depression. *Journal of Neurological and Neurosurgical Psychiatry*, *23*, 56–62.

Hattie, J. A., Sharpley, C. F., & Rogers, H. J. (1984). Comparative effectiveness of professional and paraprofessional helpers. *Psychological Bulletin*, *95*, 534–541.

Hollon, S. D., Shelton, R. C., & Davis, D. D. (1993). Cognitive therapy for depression: Conceptual issues and clinical efficiency. *Journal of Consulting and Clinical Psychology*, *61*, 270–275.

Hollon, S. D., Evans, M. D., Auerbach, A., DeRubeis, R. G., Elkin, I., Lowery, A., Kriss, M., Grove, W., Thason, V. B., & Piasecki, J. (1988). *Development of a system for rating therapies for depression: Differentiating cognitive therapy, interpersonal therapy, and clinical management pharmacotherapy*. Unpublished manuscript, University of Minnesota, Twin Cities Campus.

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

Kolb, D. A. (1984). *Experiential learning: Experience as the source of learning and development*. Englewood Cliffs: Prentice-Hall.

Lahey, M. A., Downey, R. E., & Saal, F. E. (1983). Intraclass correlations: There's more than meets the eye. *Psychological Bulletin*, *93*, 586–595.

Milne, D. L., Baker, C., Blackburn, I.-M., James, I. A., & Reichelt, F. K. (1999). Effectiveness of cognitive therapy training. *Journal of Behaviour Therapy and Experimental Psychiatry*, *30*, 81–92.

Milne, D. L., Claydon, T., Blackburn, I.-M., James, I. A., & Sheikh, A. (2001). Rationale for a new measure of competence in therapy. *Behavioural and Cognitive Psychotherapy*, *29*, 21–33.

Orlinsky, D., Grawe, K., & Parks, B. K. (1994). Process and outcome in psychotherapy. In A. E. Bergin & S. L. Garfield (Eds.), *Handbook of psychotherapy and behavior change* (4th ed.) (pp. 270–376). New York: Wiley.

Raue, P. J., & Goldfried, M. R. (1994). The therapeutic alliance in cognitive behavior therapy. In A. O. Horvath & L. S. Greenberg (Eds.), *The working alliance: Theory, research and practice* (pp. 131–152). New York: Wiley.

Roth, A., & Fonagy, P. (1996). *What works for whom? A critical review of psychotherapy research*. New York: Guilford Press.

Shapiro, D. A., Barkham, M., & Irving, D. L. (1984). The reliability of a modified helper behaviour system. *British Journal of Medical Psychology*, *57*, 45–48.

Shapiro, D. A., Harper, H., Startup, M., Reynolds, S., Bird, D., & Suskas, A. (1994). The high-water mark of the drug metaphor: A meta-analytic critique of process-outcome research. In R. L. Russell (Ed.), *Reassessing psychotherapy research* (pp. 1–35). New York: Guilford Press.

Shaw, B. F., Elkin, I., Yamaguchi, J., Olmsted, M., Vallis, T. M., Dobson, K. S., Lowery, A., Sotsky, S. M., Watkins, J. T., & Imber, S. D. (1999). Therapist competence ratings in relation to clinical outcome in cognitive therapy of depression. *Journal of Consulting and Clinical Psychology*, *67*, 837–846.

Shaw, B. F., & Wilson-Smith, D. (1988). Training therapists in cognitive-behaviour therapy. In C, Perris, I.-M. Blackburn, & H. Perris (Eds.), *Cognitive psychotherapy: Theory and practice*. Heidelberg: Springer Verlag.

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

Stern, R. (1993). Behavioural-cognitive psychotherapy training for psychiatrists. *Psychiatric Bulletin*, *17*, 1–4.

Teasdale, J. D. (1996). Clinically relevant theory: Integrating clinical insight with cognitive science. In P. M. Salkovskis (Ed.), *Frontiers of cognitive therapy*. New York: Guilford Press.

The Sainsbury Centre for Mental Health (1997). *Pulling together. The future roles and training of mental health staff*. London: The Sainsbury Centre for Mental Health.

Vallis, T. M., Shaw, B. F., & Dobson, K. S. (1986). The Cognitive Scale: Psychometric properties. *Journal of Consulting and Clinical Psychology*, *54*, 381–385.

Vallis, T. M., Shaw, B. F., & McCabe, S. B. (1988). The relationship between therapist competency in cognitive therapy and general therapy skill. *Journal of Cognitive Psychotherapy: An International Quarterly*, *2*, 237–250.

Weisz, J. R., Weiss, B., & Han, S. S. (1995). Bridging the gap between laboratory and clinic and adolescent psychotherapy. *Journal of Consulting and Clinical Psychology*, *63*, 688–701.

Whisman, M. A. (1993). Mediators and moderators of change in cognitive therapy of depression. *Psychological Bulletin*, *114*, 248–265.

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 Psychotherapy*, 373–376.

Young, J. E. (1990). *Cognitive therapy for personality disorders: A schema-focused approach*. Sarasota: Professional Resource Exchange.

Young, J. E., & Beck, A. T. (1980). *Cognitive Therapy Scale: Rating manual*. Unpublished Manuscript, University of Pennsylvania, Philadelphia, PA.

Young, J. E., & Beck, A. T. (1988). *Cognitive Therapy Scale*. Unpublished Manuscript, University of Pennsylvania, Philadelphia, PA.