Checklist for the Appraisal of Therapeutic Competence Scale Studies (CATCS)

	Criteria	Description	Excellent 2	Fair 1	Poor 0
1	**Generalisability**				
1.1	**Study purpose**	The purpose of the study is clearly defined and aims and objectives are clear	Study is clearly defined and clear aims and objectives	Study is clearly defined but no specific aims and objectives stated	Purpose of study unclear and no aims or objectives.
1.2	**Protocol for scale**	The scale is described and there a standardised protocol for administration and scoring which is fully described or reference to protocol is provided	Described in study or reference provided to protocol	Mentioned but not in sufficient detail	No reference to protocol
1.3	**Therapy/patients /setting**	Type of disorders treated (severity and/or disorder), stage of therapy and demographics of patients/service setting provided	All or most of details are provided	Brief details are provided	No information is provided
1.4	**Recordings**	There was a clear explanation of how tapes were selected for analysis. For examples see ** below	Clear explanation of tape sampling included	Brief explanation of sampling	No explanation of sampling included
1.5	**Number of raters**	Identify the number of raters used	7 or more raters	Between 3 and 6 raters	Less than 3 raters
1.6	**Raters**	There was at least some raters who were independent from the research team, well experienced and sufficiently trained	There was multiple raters who were independent from research team, well experienced and trained	There was at least one rater who was independent from the research team, experienced and trained	No information is provided or raters were not independent from the study or the rater was not trained and experienced
1.7	**Number of therapists**	Identify the number of therapists used	10 or more therapists	Between 5 and 10 therapists	Less than 5 therapists
1.8	**Therapists**	Therapists were independent from the research team; experience of therapist and their training is described and demographics of therapists is described	All or most of details are provided	About half of the details are provided	None or little information is provided

2	**Reliability**				
2.1	**Inter-rater reliability**	Appropriate statistical measures been used to assess agreement between two or more different raters. For excellent both total scale and individual items should be reported	Agreement is reported by Kappa or ICC agreement (with confidence limits reported). Total scale AND individual items both reported AND sample size ≥100	Statistical analysis is provided but only total scale without individual items OR Pearson correlation coefficient calculated OR sample size 30-99	Not best practice e.g. absolute percentages reported OR sample size <30
2.2	**Test-retest reliability**	Appropriate statistical measures been used to assess agreement between two or more occasions using the same rater. For excellent both total scale and individual items should be reported	Agreement is reported by Kappa or ICC agreement (with confidence limits reported). Total scale AND individual items both reported AND sample size ≥100	Statistical analysis is provided but only total scale without individual items OR Pearson correlation coefficient calculated OR sample size 30-99	Not best practice e.g. absolute percentages reported OR sample size <30
2.3	**Measurement error**	There were two measurements available to calculate measurement error, with an appropriate time interval and using appropriate statistical measures	Time interval described AND Standard error of Measurement (SEM), Smallest Detectable Change (SDC) or Limits of Agreement (LoA) reported AND sample size ≥100	Time interval not provided or data provided but not calculated OR sample size 30-99	SEM calculated based on Cronbach's alpha or SDC from another population OR sample size <30
2.4	**Internal consistency**	An internally consistent (homogeneous or unidimensional) scale is achieved through good construct definitions, good items, then principal component analysis or exploratory factor analysis, followed by confirmatory factor analysis	Factor analyses performed AND Cronbach's alpha(s) calculated per dimension AND sample size ≥100	No factor analysis OR doubtful design or method OR sample size 30-99	Analysis not calculated for each subscale OR sample size <30

3	**Validity**				
3.1	**Structural validity**	Structural validity should be assessed to determine or confirm existing subscales, for multi-item instruments	Exploratory or confirmatory factor analysis performed OR Item Response Theory tests for determining (uni)dimensionality performed AND sample size ≥100	A method was reported but alternative would have been more suitable OR sample size 30-99	N/A
3.2	**Hypothesis testing***	Specific hypothesis made that relate to convergent or divergent/discriminant validity	Specific hypotheses were formulated before data collection AND reported usually with correlation coefficients AND sample size ≥100	Doubtful design or method OR sample size 30-99	N/A
3.4	**Criterion validity**	This can be assessed if a study has identified a gold standard, and describes predictive validity when measured in the future, and concurrent validity when measured in the present	Convincing arguments that comparable measure is gold standard/or prominent measure AND correlation reported AND sample size ≥100	No convincing arguments that comparable measure is gold standard/or prominent measure OR doubtful design OR sample size 30-99	Criterion used can NOT be considered the gold standard OR sample size <30
3.5	**Content Validity**	Either opinion or consensus on usefulness of scale/measure was gathered	Reported in study sufficiently	Briefly mentioned	N/A
4	**Responsiveness***	Scale measures improvement in competence over time. Floor or ceiling effects are presented if more than 15% of respondents achieved the lowest or highest possible score	Effect size reported AND floor or ceiling effects presented if relevant AND sample size ≥100	Only effect size reported or doubtful design OR sample size 30-99	Not longitudinal design or time interval not described OR sample size <30

*It is important to clearly distinguish between hypothesis testing and responsiveness. Responsiveness refers to the ability of a scale to detect changes longitudinally/over time. So, in the case of competence scales this refers to therapists improving over time because of experience or training. Hypothesis testing is done to determine if scores of a scale are consistent with hypotheses (for instance regarding internal relationships, relationships to scores of other instruments, or differences between relevant groups). Good convergent validity would mean constructs on a scale that should be related are related. Good discriminant validity would mean constructs on a scale that should not be related are not related (Mokkink et al., 2010c). For example, if a scale has good discriminant validity it would be able to detect differences between novice and experienced therapists.

****Examples of how to rate Recordings**

Excellent: Clear explanation of tape sampling included
Example "As part of their training, therapists submitted six recordings of CBT sessions with patients. Data was collected from the first two terms (providing up to four recordings per therapist). Recordings were selected by therapists who completed a self-rating of their performance within the recorded session. In addition, 20 session recordings (26.32%) were selected at random and blind double rated by one of the authors."

Fair: Brief explanation of sampling
Example "Between two and four of 41 total videotaped sessions for each of the eight patients were randomly selected, including one of the 12 core CPT sessions and at least one of the additional individualized sessions. The first four sessions were not assessed. In total, the study comprised 30 videotapes."
While this study provided good information, it did not describe what random sampling method was used.

Poor: No explanation of sampling included