Citation: Flinn, L., Braham, L., & **dasNair, R.** (2014). How Reliable are Case Formulations? A Systematic Literature Review. *British Journal of Clinical Psychology.* DOI: 10.1111/bjc.12073

How reliable are case formulations? A systematic literature review

Lucinda Flinn[1,2]*, Louise Braham[1,2] and Roshan das Nair[1,3,4]

[1]Trent Doctorate in Clinical Psychology, Division of Psychiatry and Applied Psychology, University of Nottingham, UK
[2]Nottinghamshire Healthcare NHS Trust, Nottingham, UK
[3]Department of Clinical Psychology & Neuropsychology, Nottingham University Hospitals NHS Trust, Nottingham, UK
[4]Division of Rehabilitation & Ageing, University of Nottingham, UK

Objectives. This systematic literature review investigated the inter-rater and test–retest reliability of case formulations. We considered the reliability of case formulations across a range of theoretical modalities and the general quality of the primary research studies.

Methods. A systematic search of five electronic databases was conducted in addition to reference list trawling to find studies that assessed the reliability of case formulation. This yielded 18 studies for review. A methodological quality assessment tool was developed to assess the quality of studies, which informed interpretation of the findings.

Results. Results indicated inter-rater reliability mainly ranging from slight (.1–.4) to substantial (.81–1.0). Some studies highlighted that training and increased experience led to higher levels of agreement. In general, psychodynamic formulations appeared to generate somewhat increased levels of reliability than cognitive or behavioural formulations; however, these studies also included methods that may have served to inflate reliability, for example, pooling the scores of judges. Only one study investigated the test–retest reliability of case formulations yielding support for the stability of formulations over a 3-month period.

Conclusions. Reliability of case formulations is varied across a range of theoretical modalities, but can be improved; however, further research is required to strengthen our conclusions.

Practitioner points

Clinical implications

- The findings from the review evidence some support for case formulation being congruent with the scientist-practitioner approach.
- The reliability of case formulation is likely to be improved through training and clinical experience.

Limitations

- The broad inclusion criteria may have introduced heterogeneity into the sample, which may have affected the results.

- Studies reviewed were limited to peer-reviewed journal articles written in the English language, which may represent a source of publication and selection bias.

Formulation, also referred to as case formulation and case conceptualization, has been identified as an important and generic skill for applied psychologists (British Psychological Society [BPS], 2008). For clinical psychologists in particular, formulation is seen to be a fundamental core skill (Division of Clinical Psychology [DCP], 2010). Although there are many definitions of 'formulation' (BPS, 2011), the DCP (2010) offer a succinct definition:

> Psychological formulation is the summation and integration of the knowledge that is acquired by this assessment process that may involve psychological, biological and systemic factors and procedures. The formulation will draw on psychological theory and research to provide a framework for describing a client's problem or needs, how it developed and is being maintained. (pp. 5-6)

Due to various schools within the profession of psychology, formulations inevitably focus on different aspects of a case depending on the theoretical orientation of the clinician. For example, a cognitive therapist is likely to focus on cognitive mechanisms, whereas a psychodynamic therapist may focus more on unconscious processes. Furthermore, formulations can be developed at the problem level or the case level. The former focuses on a specific issue whereas the latter takes account of all of the client's difficulties.

It has been purported that formulation follows the scientist-practitioner approach (Tarrier & Calam, 2002) by utilizing an evidence base to understand a concept. More specifically, formulation uses 'psychological science to help solve human problems' (DCP, 2010, p. 3). For the cognitive model (Beck, 1976) in particular, formulation has been described as 'the heart of evidence-based practice' (Kuyken, Fothergill, Musa, & Chadwick, 2005, p. 1188). However, for a skill considered so pertinent to the role of a clinical psychologist, there is a paucity of empirical research and formulation should be open to scientific examination.

One area of scientific investigation concerns reliability. Investigations into the reliability of formulation can be traced back to 1966, where Philip Seitz (1966) published 'the consensus problem in psychoanalytic research' (p. 206). This article detailed a 3-year research study involving a group of sixpsychoanalysts, which concluded that agreement was achieved in very few of the formulations. Seitz refers to one possible reason for this being the 'inadequacy of our interpretive methods' (p. 214) where participants demonstrated the tendency to develop complex inferences regarding the cases. Seitz also recognized that participants had the tendency to rely on intuitive impression without critically checking these. Theoverall value of Seitz' work was highlighting the 'consensus problem', however, in the years following, a range of researchers sought to improve reliability in case formulations. One key researcher and the first to achieve this was Lester Luborsky (1977) using his core conflictual relationship theme (CCRT) method. Whilst the majority of formulation methods were developed within a psychodynamic framework, other methods such as cognitive-behavioural, behavioural and integrative have also been proposed (Eells, 2007).

Formulations must not be wholly subjective, it is therefore important to understand and establish reliability. Bieling and Kuyken (2003) review of literature in relation to cognitive case formulation concluded that good levels of reliability have been obtained for descriptive aspects of a case, with reliability being somewhat compromised and subsequently decreasing for the more inferential and theory-driven aspects. In addition, they briefly reviewed the psychodynamic literature, which showed promising results for

reliability. Other reviews of case formulation literature have reported similar results (Aston, 2009; Mumma, 2011). Most research has focused on inter-rater reliability, that is, the rate of consistency between clinicians on aspects of a case. Test–retest reliability, whether formulations remain stable over time, has had much less of a focus (Bieling & Kuyken, 2003), with some research evident in relation to the psychodynamic model (e.g., Barber, Luborsky, Crits-Christoph, &Diguer, 1995) and none known to the cognitive model. Whilst the Bieling and Kuyken (2003), Aston (2009) and Mumma (2011) reviews considered some of the available research in the area, they were not systematic reviews and were by no means exhaustive.

Rationale and aim

Whilst it could be argued that some theoretical modalities place more emphasis on the relation of formulation to evidence-based practice, it is a skill central to the work of all clinical psychologists. It is therefore important to develop a scientific foundation for formulation, which includes reviewing the reliability. Clinical psychologist Gillian Butler (2006) suggests that low reliability is inevitable due to there being no one correct way to formulate. She argues that clinicians presented with the same information may well develop alternative formulations, even if they are formulating from the same psychological model. Whilst this may be the case, the literature suggests a tension between formulation being viewed as a 'science' (with its trappings of measurability, reliability, etc.) and an 'art' (with its emphasis on an ideographic approach that is beyond the realms of scientific scrutiny). We therefore felt that from a scientific perspective, a systematic literature review appears necessary to draw conclusions from the available literature about constructs related to reliability. To date, there has been no systematic literature review on any aspect of formulation. The overall aim of this systematic review is to answer the following question: What is the reliability of case formulations? In attempting to answer this question, we focus on the reliabilities of various theoretical modalities, and comment on the overall quality of the primary research.

Method

Due to previous reviews of case formulation (e.g., Aston, 2009; Bieling & Kuyken, 2003), we were aware that the number of studies that examined the reliability of case formulation would be limited. Therefore, no restriction in date was applied other than the start of the searched databases.

Inclusion criteria

Studies were eligible if they:

- Examined the inter-rater or test–retest reliability of case formulation. This required reporting the results of a reliability measure.
- Outlined the theoretical model of the formulation method, as psychology is a profession based on a variety of theoretical modalities.
- Included adult, child formulations, or both.
- Investigated the reliability of case formulations developed by any mental health professional, including studies that utilized a combination of clinicians and students.
- Were peer-reviewed journal articles, to control for quality.

- Were written in the English language.

Exclusion criteria

Studies were excluded if they:

- Had formulators recruited entirely from a student population, which would reduce the ecological validity of this review.
- Consisted of a review of previous research with no new research being undertaken.
- Focused on the assessment and reliability of measures that may serve to influence the process of formulation, for example, pre-therapeutic assessment tools.

Search overview

Studies were accessed through a range of databases in addition to reference list trawling. This included reference list trawling of reviews of formulation research such as those of Barber and Crits-Christoph (1993) and Bieling and Kuyken (2003). Five databases were searched in April 2014: PsycINFO (1806[1]); MEDLINE (1948); AMED (1985); CINAHL (1981); Web of Science (1900). These databases are similar to the ones utilized in a narrative review of the case formulation literature (Aston, 2009), and cover journal articles that relate to psychology. The following search terms were used: formulation OR case formulation OR case conceptualization OR case conceptualization AND statistical reliability OR reliability OR inter-rater reliability OR inter-rater reliability OR test reliability OR test–retest reliability.

These search terms yielded a total of 4,318 articles from all five databases. After applying the inclusion and exclusion criteria in addition to reference list trawling, 18 articles remained (see Figure 1 for the quorum diagram).

Data extraction

Specific data were extracted for each selected study. Table 1 details the data extracted.

Assessment of methodological quality

Several scales have been developed to assess methodological quality of studies and to standardize this process. However, due to the variability in research designs of the selected articles, none of the pre-existing tools could be applied to the current review. This follows from suggestions that authors can develop their own tool by adapting available tools (e.g., Parker, 2004). Therefore, our quality assessment tool was developed with reference to the Critical Appraisal Skills Programme (CASP, 2004) and the Newcastle-Ottowa Scale (Wells *et al.*, 2010). The resultant tool comprised five questions, each with a rating out of three and a total sum of 15. Furthermore, a separate section incorporated information regarding additional sources of bias to provide further information related to quality. This information is tabulated in Table 2.

For the current review, a hierarchy was decided upon for the reliability data. This was based on ecological validity and the real life experiences of formulators. Therefore, video recordings were assigned a score of 3, audio recordings were assigned a score of 2, and transcripts or written vignettes were assigned a score of 1. Furthermore, if reliability data were not outlined then studies were also assigned a score of 1. With regard to reliability

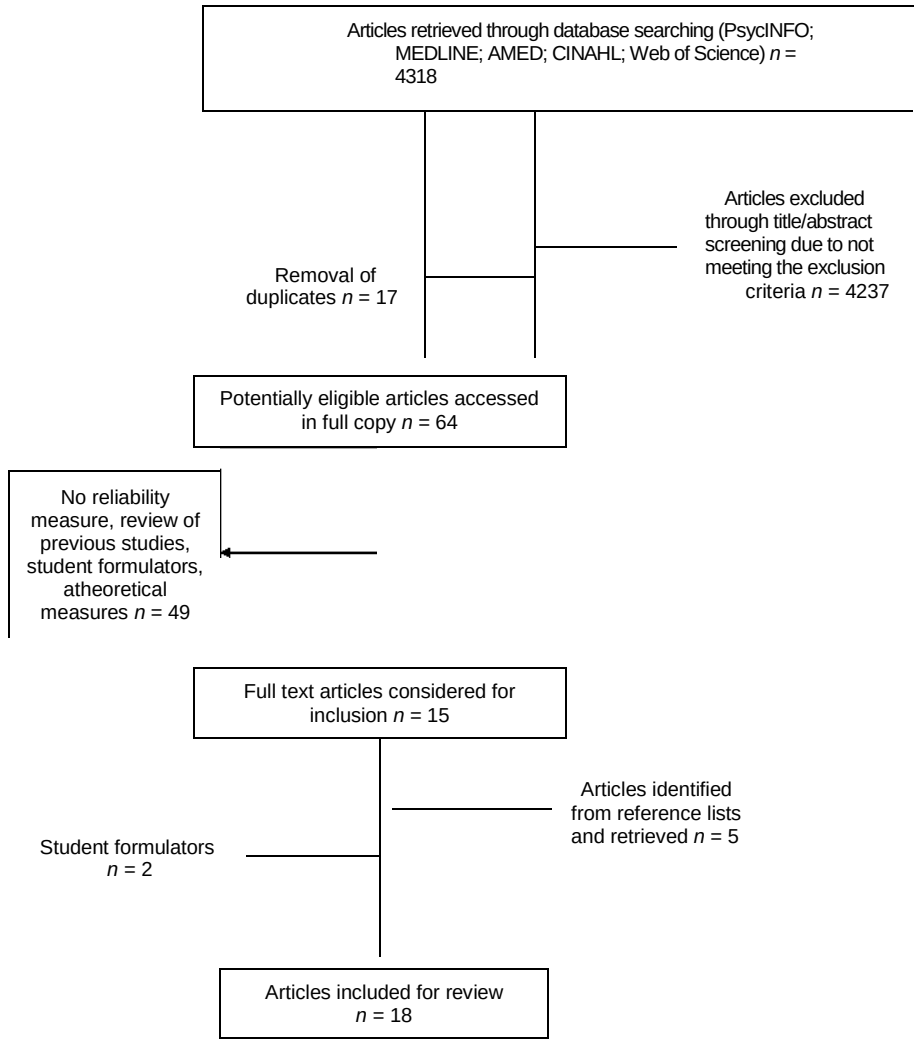

Figure 1. Quorum diagram detailing study selection.

measurement, studies were assigned a score of 3 if they used an appropriate statistical measure of reliability (that accounts for chance agreement) for all aspects of data analysis. To obtain a score of 3, studies could also incorporate percentage agreement but would be required to report statistical measures also for the same focus of agreement or reliability. A score of 2 was assigned if studies used statistical measures for some aspects and percentage agreement for others, and a score of 1 was assigned if studies used percentage agreement only.

Inter-rater reliability for quality assessment was assessed through the use of a second rater (LB) scoring over 20% of the studies. For this review, articles were not excluded through quality assessment. This was to avoid excluding potentially relevant studies. However, the quality assessment tool informed the interpretation of the findings. Quality rating was conducted by two reviewers (LF and LB) and inter-rater reliability between the

Table 1. Study details

Study ID, author(s)			
1. Kuyken *et al.* (2005) England	*Quantitative* Percentage agreement, formulations compared with each other and to benchmark *Materials* Benchmark formulation provided by J. Beck, video and assessment measures Theoretical modality – cognitive Trained participants – yes	Total (n = 115) Clinical psychologists (n = 35) Pre-qualification students (n = 29) Highest qualification (D.Clin.Psy./ Ph.D./M.D. (n = 29) Years post-qualification clinical experience (average = 7.31 years)	. Results indicated that participants could agree with each other and the benchmark on the descriptive aspects of the case. However, agreement decreased for the more inferential and theory-driven aspects . Higher percentage agreement was associated with accreditation status . The pre-qualified group were least likely to identify an important part of the benchmark formulation, however, on some aspects they demonstrated higher agreement that accredited practitioners, that is, the core belief 'I'm unlovable'
2. Persons, Mooney, and Padesky (1995) America	*Quantitative* Intraclass correlation coefficients and percentage agreement *Materials* 2 clients (depression and anxiety) Audiotapes/transcripts -12 min of session Participants could list up to 6 problems Multiple-choice questionnaire for underlying cognitive mechanisms. These were rated on a scale 0–10 for relevance to the case Theoretical modality – cognitive Trained participants – yes	Total (n = 46) Students (4.3%) Previous training in formulation (69.6%) Previous training in cognitive therapy (average = 89.1%)	. Inter-rater reliability for underlying mechanisms ranged from 0.07 to 0.70 for a randomly selected single judge and 0.27–0.92 for a random sample of five judges . Percentage agreement for the problem list ranged from 13% to 97.8% for the first case and 67.4–100% for the second case
3. Persons and Bertagnolli (1999)	*Quantitative* Intraclass correlation coefficients and percentage agreement	Total (n = 47) Students (12.8%) Previous training in formulation (63%)	. Inter-rater reliability for underlying mechanisms ranged from 0.44 to 0.91 for five judges and 0.13 to 0.66 for single judges

Continued

Table 1. (*Continued*)

Study ID, author(s)			
			averaged 67.46% . Providing participants with specific contexts did not increase reliability
4. Mumma and Smith (2001) America	*Quantitative* Intraclass correlation coefficients *Materials* 4 female patients (3 depression 1 generalized anxiety disorder) Raters rated Cognitive-Behavioural-Interpersonal Scenarios (CBIS) using a multidimensional rating scale Theoretical modality — cognitive Trained participants — no	Clinical formulators Total (n = 4) Clinical psychologists (n = 1) Students (n = 3) Clinical raters Total (n = 10) Clinical psychologists (n = 6) Counselling psychologists (n = 2)	. ICC ratings ranged from .83 to .94 averaged across 10 raters and ranged from .33 to .63 for single ratings . Results demonstrate reliability for clinical scenarios at a situation-level formulation
5. Dudley, Park, James, and Dodgson (2010) England	*Quantitative* Percentage agreement *Materials* 1 client (psychosis) role-played by a clinical psychologist	Total (n = 85) Clinical psychologists (n = 5) Students (n = 17) Other (n = 3)	. There was greater agreement for behavioural, physical symptoms and triggers and less agreement for more theory-driven and inferential aspects such as the

Continued

Table 1. (*Continued*)

Study ID, author(s)			
	30 min video recording of assessment session and timeline Benchmark provided by expert panel Theoretical modality — cognitive Trained participants — no	Highest qualification (D.Clin.Psy./ Ph.D./M.D. (n = 7)	identification of core beliefs and dysfunctional assumptions . Overall percentage agreement ranged from 8.9% to 95% but overt behaviour resulted in the most agreement (91.6%) . Greater clinical experience improved agreement with the benchmark
6. Muran, Segal, and Samstag (1994) America	*Quantitative* Intraclass correlation coefficients *Materials* 8 clients (5 with mood disorders 2 with anxiety disorders and 1 with a mood and anxiety disorder) Audiotapes of 2 assessment interviews 2—3 scenarios combined with 2—3 less or not relevant Rated for relevance by therapist, formulator and client on a 9 point Likert scale Theoretical modality — cognitive Trained participants — no	Therapists Total (n = 5) Experience in cognitive therapy (average = 3.1 years) Formulators Total (n = 2) Experience in cognitive therapy (1 = 1 year; 1 = 3 years)	therapists, formulators and clients were .92 for stimulus situation, .93 for cognitive, .90 for affective and .91 for motoric component
7. Wilson and Evans (1983) America	*Quantitative* Percentage agreement and concordance coefficients *Materials* 3 fictional child cases. Written narratives Theoretical modality — behavioural Trained participants — no	Total = 118 (65 in 3 problem experimental condition and 53 in 6 problem experimental condition) Primary profession Direct clinical service = 30% Administration = 11%	. Percentage agreement for the low complexity (3 problem) group ranged from 38% to 42% and 30% to 43% for the high complexity (6 problem) group . Kendall coefficient of concordance statistics ranged from .13 to .58 for the low

Continued

Table 1. (*Continued*)

Study ID, author(s)			
		All participants had either a Ph.D or Psy. D degree	complexity group and .13 to .52 for the high complexity group . When clinicians were asked to select target behaviours, reliability was low
8. Barber *et al.* (1995) America	*Quantitative* Percentage agreement and weighted kappa *Materials* 19 clients with depression Transcribed interview Standard categories of CCRT Theoretical modality — psychodynamic Trained participants — yes	RAP interviewer (n = 1 research assistant) Judges (n = 2 psychodynamic clinicians)	ranged from .60 to .68 for CCRT from RAP interviews, .64 to .81 for CCRT from sessions 3 and 5 and .40 to 1.0 when comparing CCRT from RAP interviews to CCRT from sessions . Results suggest agood level of agreement for deriving CCRT between two methods (RAP interviews and sessions) and that relationship themes that emerge during therapy are similar to the ones that emerge during a pre-therapy interview
9. Shefler and Tishby (1998) Israel	*Quantitative* Intraclass correlation coefficients and chi-square test *Materials* 15 clients (5 = no diagnosis, 5 = adjustment disorder, 3 = anxiety disorders, 2 = depressive disorders) Transcribed intake interviews The accuracy rating scale (developed for study) Central issue for each case and a second less	Formulators (n = 9 psychoanalytically orientated therapists) Clinical psychologists = 5 Psychiatrist = 1 Social workers = 3 Similarity judges (n = 15) Psychologists = 10 Social workers = 5	reliability yielded slight scores for three of the cases (R < .40) and fair to substantial for the remaining twelve, ranging from R = .46 to R = .85. The mean ICC for 15 cases was R = .54 . The chi-square test was calculated. In 10 out of 15 cases inter-rater agreement was greater than would be expected by chance. T values ranged from .30 to .74

Continued

Table 1. (*Continued*)

Study ID, author(s)			
	relevant but plausible central issue Theoretical modality – integrative psychoanalytical Trained participants – yes		
10. Rosenberg, Silberschatz, Curtis, Sampson, and Weiss (1986) America	*Quantitative* Intraclass correlation coefficients and Pearson correlations *Materials* 5 clients (neurotic and personality disorders) Transcripts of the first 2 hr of therapy Formulations combined with less relevant but plausible items Relevance rated on a 9 point Likert scale Theoretical modality – cognitive psychoanalytic Trained participants – not for the study but unclear about previous training	Formulators (n = 4 or 5 clinicians who shared the theoretical modality) Reliability judges (n = 4 who shared the theoretical modality)	. Intraclass correlation over the 5 patients ranged from .14 to .88 for the average judge and .39 to .97 for pooled judges . To assess the degree of overlap between the clinicians and reliability judges, Pearson correlations were compared. These ranged from .70 to .97 across all 5 patients
11. Curtis, Silberschatz, Weiss, Sampson, and Rosenberg (1988) America	*Quantitative* Intraclass correlation coefficients *Materials* 1 case vignette (unclear if fictional) Formulations reduced to concise statements and combined with less relevant but plausible items Relevance rated on 9 point Likert scale Theoretical modality – Cognitive psychoanalytic	Formulators (n = 3–5 clinicians) Reliability judges (n = 4–5)	. For plan component items, intraclass correlations for the mean of 5 judges' ratings were .96 for goals, .97 for obstructions, .94 for tests and .94 for insights . For the reliability of the formulation, intraclass correlation coefficients for the mean of the reliability judges were .90 for goals, .93 for obstructions, .72 for tests and .84 for insights

Continued

Table 1. (*Continued*)

Study ID, author(s)			
	Trained participants – not for the study but unclear about previous training		. Between the mean ratings of the formulation and reliability teams, interclass correlation coefficients were .88 for goals, .88 for obstructions, .62 for tests and .83 for insights
12. Collins and Messer (1991) America	*Quantitative* Intraclass correlation coefficients and percentage agreement *Materials* 2 clients (depression) A pre-existing narrative for Case B Theoretical modality – cognitive-psychoanalytic Trained participants – yes	Clinical judges Mount Zion Panel (n = 5 psychologists or psychiatrists with at least 3 years clinical work) Rutgers Panel (n = 5 clinicians, 3 who had between 2 and 5 years of clinical experience, 1 with 3 months experience and 1 with over 20 years of clinical experience)	. Intraclass correlation coefficients for pooled judges for the Rutgers Plans in time 1 ranged from .81 to .93. The Mount Zion coefficients fortime 1 were also substantial, rangingfrom .81 to .95 . Intraclass correlation coefficients for pooled judges for the Rutgers Plans for time 2 ranged from .75 to .96 . Pearson product-moment and percentage agreement ratings indicated high levels of stability. Between 85% and 97% of Case A items and between 90% and 96% of Case B items were retained at a 3-month follow-up
13.DeWitt, Kaltreider, Weiss, and Horowitz (1983) America	*Quantitative* Intraclass correlation coefficients *Materials* 18 clients with pathological grief reactions Audiotapes or videotapes of evaluation session Hypothesis and criteria set Theoretical modality – psychodynamic	Formulators (n = 9) Psychologist = 1 Psychiatrists = 6 Social workers = 2 3 teams of 3 judges 1 team formulated same case both at intake and follow-up Agreement raters (n = 4)	. Intraclass correlation coefficients for the identification of hypotheses were .78 using the pooled estimate and .74 for the single rater estimate . Intraclass correlation coefficients for the identification of criteria were .84 for the pooled estimate and .72 for the single rater estimate

Continued

Table 1. (*Continued*)

Study ID, author(s)			
	Trained participants – not for the study but unclear about previous training	Psychiatrist = 1 Psychologist = 3	
14. Caston and Martin (1993) America	*Quantitative* Spearman rho, chi-square (unreported) and *T* statistic *Materials* 1 client (low self-esteem and a lack of sexual responsiveness) Verbatim transcripts of the first 5 analytic hours 5 Psychoanalytical domains rated on 9-point Likert scales Theoretical modality – psychodynamic Trained participants – not for the study but unclear about previous training	Formulators (n = 2) Text-wise judges (n = 4 psychoanalysts) made independent ratings Mannequin 'text-less' judges (n = 4 psychoanalysts) made ratings on the same domains	Agreement on order and magnitude for the four text-wise judges yielded Spearman rho correlations of .61–.96 and *T* statistics of .15 – .65. For wishes, Spearman rho correlations ranged from .02 to .91 and *T* statistics ranged from .42 to .91 Agreement on order and magnitude for the four text-wise judges and the 4 text-less judges yielded Spearman rho correlations of .08–.77 and *T* statistics of .00–.15. For wishes, Spearman rho correlations ranged from .31 to .87 and *T* statistics ranged from .03 to .49
15. Crits-Christoph *et al.* (1988) America	*Quantitative* Intraclass correlation coefficients, weighted kappa and percentage agreement *Materials* 35 clients (variety of mental disorders) Transcripts for 2–3 sessions Assessed completeness of relationship episodes on scale of 0–5 Theoretical modality – psychodynamic Trained participants – yes	Relationship episode identifiers (n = 2) Psychiatrist (n = 1) Research assistant (n = 1) CCRT Judges (n = 2) Psychiatrist (n = 1) Clinical psychologist (n = 1) Similarity judges (n = 4) Research assistants (n = 4)	Interjudge agreement when picking standard categories was 95% Intraclass correlation coefficient of pooled 4 similarity judges for mismatched cases was .79 Using standard categories, weighted kappa for the wish and negative responses of self were .61. For negative response from other, weighted kappa was .70
16. Perry, Augusto, and	*Quantitative* Intraclass correlation coefficients and	Formulators (n = 2 clinicians with at least 5 years post-qualification	The intraclass reliability of the consensus ratings of similarity ranges from .54 to .75.

Continued

Table 1. (*Continued*)

Study ID, author(s)			
(1989) America	Analysis of Variance (ANOVA) *Materials* 20 clients who had a videotaped interview and written ICF (8 with borderline personality disorder 6 with antisocial personality disorder and 6 with bipolar type II) Theoretical modality – psychodynamic Trained participants – yes	experience) Similarity judges (n = 4 students)	The ICC was then recalculated, subtracting out the means for each type of comparison where ratings ranged from .51 to .68 . ANOVA of means for matched cases were significantly higher than the means for either type of mismatched comparison
17. Ee Ills *et al.* (1995)	*Quantitative* Intraclass correlation coefficients and t-tests *Materials* 2 clients (1 = pathological grief and 1 = social phobia) Transcripts of first 5 sessions Similarity ratings made on a 6-point Likert scale Theoretical modality – integrative Trained participants – not for the study but unclear about previous training	Formulators Team 1 (n = 12) Clinical psychologists (n = 2) Students (n = 10) Team 2 (n = 3) Clinical psychologists (n = 2) Psychiatrist (n = 1) Similarity judges (n = 20 students)	. Intraclass correlation coefficients for pooled scores of 20 similarity judges generated a mean of.74 for RRMC items and .89 for RRM items . Intraclass correlation coefficients for RRM quadrants yielded mean scores of .74 for the Desired RRM quadrant to .87 for the Dreaded RRM quadrant . Matched-group t-tests indicated that correctly matched pairings were more similar than incorrectly matched pairings
18. Popp*et al.* (1996) America	*Quantitative* Weighted kappa *Materials* 13 clients with a minimum of20 usable dreams Standard categories of CCRT Theoretical modality – psychodynamic Trained participants – not for the study but unclear about previous training	Judges (n = 2 with undefined demographics)	. Weighted kappa of two judges for dreams was 0.58 (wish) 0.70 (response from other) 0.83 (response of self) and for narratives was 0.67 (wish) 0.74 (response from other) 0.75 (response of self) . The results suggest the presence of a central relationship pattern that is expressed in waking narratives as well as dreams

Note. RRMC, Role-Relationship Model Configuration; RRM, Role-Relationship Model.

Table 2. Assessment of methodological quality

Study	Author(s)	Participant demographics	Sample representativeness	Formulation data	Blinding	Reliability measurement	Other potential sources of bias
1	[illegible]	*	*	*	*	*	[illegible] workshops with notable differences in content
2	[illegible]	*	*	*	*	*	[illegible]
5	Dudley, Park, James, and Dodgson (2010)	*	*	*	*	*	[illegible]
6	[illegible]		*	*	*	*	[illegible]
7	[illegible]	*	*	*	*	*	[illegible]
8	[illegible]	*	*	*	*	*	[illegible]
9	[illegible]	*	*	*	*	*	[illegible]
0	[illegible]	*	*	*	*	*	[illegible]
1	[illegible]	*	*	*	*	*	[illegible] Formulation items combined with less relevant but plausible [illegible]

Continued

Table 2. (*Continued*)

Study	Author(s)	Participant demographics	Sample representativeness	Formulation data	Blinding	Reliability measurement	Other potential sources of bias
5	[illegible]	*	*	*	*	*	[illegible]
6	[illegible]	*	*	*	*	*	[illegible]
7	[illegible]	*	*	*	*	*	[illegible]
8	[illegible]	*	*	*	*	*	[illegible]

Note. (1) Participant demographics (formulators/raters not clients): ***participant demographics are reported clearly **participant demographics are reported partially *participant demographics are reported inadequately. (2) Sample representativeness: ***participants consisted of clinicians **participants consisted of a range of clinicians and students *participants mainly consisted of students. (3) Formulation data: ***the study used video recordings **the study used audio recordings *the study used transcripts, written vignettes or does not define the formulation data. (4) Blinding: ***the study reports adequate blinding **the study reports partial blinding *there is no evidence of blinding. (5) Reliability measurement: ***the study used a statistical measure of reliability **the study uses a statistical measure of reliability in addition to percentage agreement *the study uses percentage agreement.

two raters resulted in 83% agreement. Discrepancies were resolved through discussion. The 18 studies included in the review yielded total quality scores ranging from 7 to 14.

Results

All 18 studies tested the inter-rater reliability of formulations with one study (12) also investigating test–retest reliability. This was achieved by investigating the stability of formulations over a 3-month period in the absence of new client information.

Study location

The majority of studies were conducted in the USA, two studies were conducted in England (1 and 5) and one study was conducted in Israel (9).

Theoretical modality

Six studies formulated from a cognitive modality (1, 2, 3, 4, 5 and 6) and six studies formulated using a psychodynamic modality (8, 13, 14, 15, 16 and 18). One study used a behavioural modality (7) and five studies used an integrative approach (9, 10,11, 12 and 17).

Participant demographics

In total, the studies used data from 152 client participants and between 550 and 553 formulators/raters. The exact number is not available as Studies 10 and 11 provided participant numbers within a range.

The majority of studies explicitly detailed the demographics of participants, including reference to amount of clinical experience and professional role. However, several studies (11, 14, 15 and 18) provided only minimal demographics, often referring to 'experienced clinicians' with no information as to how they defined experience.

Training

Some studies reported training that was offered to participants as part of their participation (1, 2, 3, 8, 9, 12, 15 and 16). Another study directly recruited from training courses (5). Some studies did not refer to training (4, 6, 7, 9, 10, 11, 14 and 18), but it is possible that participants had completed training as part of their professional role. A study where training was completed (12) demonstrated higher levels of agreement in comparison to a study that offered no training (7). However, for two studies where participants completed training as part of the research (2 and 3), intraclass coefficients ranged between .07 and .70 for underlying mechanisms and between 13% and 100% agreement of a client's presenting problems.

Sample

Some studies (6, 9, 10, 11, 12, 13 and 14) used clinicians to assess the reliability of case formulations. However, several studies used both clinicians and students (1, 2, 3, 4, 5, 16 and 17), although the students in study number 17 were recruited to assess the similarity of formulations as opposed to the formulators. Two studies (8 and 15) recruited both clinicians and research assistants.

Results from Studies 1 and 5 indicated that greater clinical experience was linked to increased agreement with the benchmark formulations. Study 1 found that the pre-qualified student participants were least likely to identify an important aspect of the benchmark formulation. However, it is of note that for some of the inferential aspects of the formulation, pre-qualification students actually demonstrated a higher rate of agreement in comparison to the accredited practitioners.

Formulation data

The majority of studies provided participants with one source of material in which to formulate clients' problems, including transcripts (2, 4, 8, 9, 10, 14 and 15), written vignettes (7 and 11), audio recordings (2, 3, 6 and 13), video recordings (1, 5, 13 and 16), or a combination (2 and 13). However, one study did not outline the full material used but referred to a written narrative for one of the two cases (12). Two studies (1 and 5) used multiple sources of information (video and assessment measures), which could provide more ecological validity with clinical practice. However, factors that may have served to decrease ecological validity were also present, including the use of a fictional vignette (1) and having an actor role play a client (5). In comparison to other studies, using more than one source of material did not appear to increase levels of agreement (1 and 5).

Some studies asked participants to combine their formulation items with those considered plausible but less relevant (6 and 9), which were then rated by separate participants. This could be a potential source of bias with the alternative formulation being 'straw men' and therefore easier to rate, inflating reliability. One study combined formulations of a different theoretical modality, which may have inflated reliability through theoretical bias (12).

Several studies provided participants with standard categories to choose from, for example, lists of wishes and fears for the CCRT method (8, 15 and 18) and lists of underlying cognitive mechanisms (2 and 3). Although this could serve to increase reliability, the results from the current review do not necessarily support this and results ranged from slight to substantial for cognitive formulations (2 and 3) and moderate to substantial for psychodynamic formulations (8, 15 and 18).

Blinding

The majority of studies evidenced no use of blinding (1, 2, 3, 4, 5, 6, 8, 10, 14, 17 and 18) which may have introduced bias into the research. Several studies, however, did evidence some or full use of blinding (7, 9, 11, 12, 13, 15 and 16). For example, study number 11 used reliability judges that adhered to the same theory as the formulators but were blind to the identities of the client, therapist and treatment outcome. Study number 15 used matched and mismatched formulations based on three possible types, that is, mismatched for diagnosis but matched for gender. Similarity judges in this study were blind to the hypothesis and comparison type.

Reliability measurement

Although percentage agreement can be used as a measure of reliability, it has been described as flawed in many respects (Hayes & Krippendorff, 2007) as it does not take into consideration agreement based on chance (McHugh, 2012). To account for this,

alternative measures of reliability were developed such as Cohen's kappa (Cohen, 1960) and the intraclass correlation coefficient (Shrout & Fleiss, 1979).

The studies used a range of reliability measurements, including percentage agreement (1,2,3,5,7,8, 12and15),intraclasscorrelationcoefficients(2,3,4,6,9, 10,11,12,13,15, 16 and 17), weighted kappa (8, 15 and 18), chi-square tests (9 and 14), spearman rho (14), concordance coefficients (7), Pearson correlations (10), analysis of variance (16), and t-tests (17).

Key findings

Graham, Milanowski, and Miller (2012) have suggested that there are no universal rules regarding the level of agreement required to ascertain reliability, however, there have been suggestions put forward for the various measurements of reliability. For percentage agreement, Luborsky and Diguer (1998) proposed that 70% or higher indicates good reliability, however others have suggested that anything over 75% demonstrates acceptable agreement (Hartmann, 1977; Stemler, 2004). For intraclass correlation coefficients, Shrout (1998) proposed that a score in the range of 0–.1 relates to virtually no reliability, .1–.4 relates to slight reliability, .4–.6 relates to fair reliability, .6–.8 relates to moderate reliability and .81–1.0 relates to substantial reliability. To answer the research question posed by the current review, the key results from the studies will be considered in relation to these levels of agreement.

Results from the 18 studies demonstrated agreement and reliability ranging from virtually none to substantial (14). Six studies yielded slight to substantial reliability (2, 3, 4, 7, 9 and 10). However, fair to moderate reliability was found in two studies (16 and 18), moderate reliability was demonstrated in two studies (13 and 15), moderate to substantial

Table 3. The range of reliability scores across studies included in the review

	Total number of studies and study numbers
Overall reliability range	
Virtually none to substantial	1 (14)
Slight to substantial	6 (2, 3, 4, 7, 9 and 10)
Fair to moderate	2 (16 and 18)
Moderate	2 (13 and 15)
Moderate to substantial	4 (8, 11, 12 and 17)
Substantial	1 (6)
Cognitive reliability range	
Virtually none to moderate	1 (2)
Slight to moderate	2 (3 and 4)
Substantial	1 (6)
Behavioural reliability range	
Slight to fair	1 (7)
Psychodynamic reliability range	
Slight to substantial	1 (14)
Fair to moderate	1 (16)
Moderate to substantial	5 (8, 13, 14, 15 and 18)
Integrative reliability range	
Slight/fair to substantial	2 (9 and 10)
Moderate to substantial	3 (11, 12 and 17)

reliability was shown in four studies (8, 11, 12 and 17), and substantial reliability was shown in one study (6). The full range of reliability across all studies included in the review (excluding 1 and 5 which used percentage agreement) is detailed in Table 3.

Cognitive formulations

The reliability of cognitive formulations ranged from virtually none to substantial. In general, there was higher agreement for the descriptive aspects of cognitive formulations, that is, overt difficulties (1, 2 and 3) and less agreement for the more theory-driven inferential aspects (1 and 5). Although not accounting for agreement based on chance, studies that employed purely percentage agreement for all aspects of the formulation demonstrated less than a third of items meeting the 70% threshold (1 and 5). With the limitations associated with percentage agreement, it is possible that agreement may be even less. Although these studies used both clinicians and students, these findings were maintained when levels of agreement were considered for the accredited clinicians only.

Case level formulations appeared to produce fairly low levels of reliability (1 and 5), with problem/situation specific formulations yielding substantial reliability (4). However, when study number 2 attempted to increase the reliability of a case level formulation by replicating the study but providing clinicians with specific contexts in which to rate a client's schemas, reliability was not increased (3).

Behavioural formulations

As there was only one study that used a behavioural formulation (7), comparisons with the same theoretical modality were not possible. However, it is of note that this study demonstrated low percentage agreement (30–43%) in the identification of overt problems in addition to low coefficients (.13–.58).

Psychodynamic formulations

Formulations developed from a psychodynamic theoretical modality mainly yielded moderate to substantial reliability (8, 13, 15 and 18). However, study number 16 generated fair to moderate levels of reliability. Clinicians in study 14 largely demonstrated moderate to substantial levels of reliability for order – whether items are ranked in a similar way. However, for agreement in relation to the magnitude of items, scores ranged between slight and substantial. It should be noted that scores in Studies 14 and 15 were pooled over four judges, which may have served to inflate reliability.

Integrative formulations

When correlations were pooled and averaged over several judges, results for integrative formulations demonstrated moderate to substantial reliability (10, 11, 12 and 17). However, when an average was taken for a single judge, reliability appeared to be in the slight/fair to substantial range (9 and 10), which demonstrates that pooling scores can serve to inflate reliability. There was only one study that assessed the test–retest reliability of formulations (12). Integrative formulations demonstrated good stability over a 3-month period through Pearson product-moment and percentage agreement ratings (85–97% for Case A and 90–96% for Case B; 12).

Discussion

How reliable are case formulations?

This review investigated the reliability of case formulations. Studies yielded mixed results, with reliability mainly ranging from slight to substantial. Reliability did not appear to increase when formulators were asked to identify discrete areas, for example, overt problems in behavioural formulations (7). However, results indicated the moderate agreement for the identification of underlying cognitive mechanisms and overt problems in cognitive formulations (2 and 3). When comparing different theoretical modalities, psychodynamic formulations appear to generate higher levels of reliability, however, these studies utilized methods that may have inflated reliability, such as using standard categories (8, 15 and 18) and pooling the scores of judges (14 and 15). In general, results indicate that reliability in case formulation can be achieved across all modalities. However, it is difficult to draw clear conclusions due to the dearth of literature, the varying methodologies employed and the limitations associated with these.

One methodological limitation concerns the use of students (1, 2, 3, 4 and 5). It is difficult to ascertain the standard at which a student can formulate. In clinical psychology doctoral training programmes, case formulation features heavily and is an essential skill that all trainees are required to develop (BPS, 2011). There is much less of an emphasis on case formulation in undergraduate or masters level psychology courses. It is therefore questionable at what level a psychology student or graduate could formulate. This is particularly relevant when considering the more inferential and theory-driven aspects of a case that may require advanced training and clinical experience. Although some studies incorporated training into their research (1, 2, 3, 8, 9, 12, 15 ad 16) or recruited from training programmes (5), it is questionable whether this can be comparable to the years of experience a clinician may have within a particular theoretical modality.

Another limitation concerns the use of transcripts (2, 8, 9, 10 and 14). It has been argued that using transcripts is likely to increase reliability due to the decreased chance of idiosyncrasies (Barber & Crits-Christoph, 1993). In general, these studies indicated a moderate level of reliability. However, as reliability ranged from slight to substantial it is difficult to draw clear conclusions regarding the impact that the use of transcripts has on inter-rater reliability. It could be argued that the use of transcripts, vignettes, and audio recordings decreases the ecological validity of case formulation research. In clinical practice, formulations are largely developed through engagement and exploration *with* the client and the use of such materials loses the collaboration that is often associated with formulation. Whilst this is unlikely to be overcome for research, it could be argued that using such materials prevents the formulator from noticing and interpreting non-verbal cues. Therefore, studies that incorporate video recordings (1, 5, 13 and 16) may be more ecologically valid than transcripts or vignettes.

The difficulties forming conclusions are further compounded by the different measures of reliability employed by the studies. The variability and levels of agreement required by measures make comparisons difficult (Bland & Altman, 1990). Furthermore, the use of certain statistical methods has been criticised. For example, Rankin and Stokes (1998) suggest that the Pearson correlation coefficient is inappropriate because the measurement responds to the linear association as opposed to agreement. It is of note that this measure was used to investigate the only known study of test–retest reliability of case formulation (12). In addition, there are several intraclass correlation coefficient equations available, which can result in different values being produced from the same data (Shrout & Fleiss, 1979). It has been suggested that Cohen's kappa is the most appropriate

statistical measure of reliability for nominal data, weighted kappa is most appropriate for ordinal data and the one-way analysis of variance is the most appropriate measure for continuous data (Haas, 1991).

Implications for clinical practice

Although results from the current review highlight that reliability in case formulation can be achieved, from a scientific perspective, the wide range in levels of inter-rater reliability provides modest support for formulation being at 'the heart of evidence-based practice' (Kuyken *et al.*, 2005, p. 1188). Results from the current review suggest that training may lead to higher levels of agreement and reliability, particularly with inferential and theory-driven aspects of a case. Therefore, in clinical practice, it is plausible to suggest that training and greater clinical experience may serve to increase reliability between clinicians. Although not linked to reliability, research suggests that an increase in training leads to higher quality case formulations (Kendjelic & Eells, 2007). Whilst several of the studies offered training to their participants, the amount of training varied and it is unlikely that a single workshop would be adequate to develop the skill of formulation.

A possible explanation for the limited inter-rater reliability for the inferential aspects of formulation concerns the potential of cognitive shortcuts that therapists may take, such as availability and anchoring heuristics (Corrie & Lane, 2010; Kuyken, Padesky, & Dudley, 2009). It is of note that study number 16 requested that participants include supporting evidence for their formulations and generated fair to moderate reliability. The authors suggest that this may have kept inferences at a low level. Whilst inferential aspects of a formulation are important, providing evidence may limit the amount of cognitive shortcuts that therapists take, potentially leading to higher reliability.

High levels of reliability do not necessarily imply validity and it is questionable whether validity could be scientifically evaluated, particularly with clients who are acquiescent with credible formulations. Butler (2006) suggests 'formulations, as hypotheses (are) a way of making theory-based guesses' (p. 9) and therefore may be very different but potentially equally valuable. However, this poses questions regarding the implications for treatment outcome if reliability between clinicians is low and different areas are being targeted through treatment. It should be noted that not everyone advocates the importance of case formulation reliability. For example, Wilson (1996) has likened the case formulation to clinical judgement, which he argues 'can be all too fallible' (p. 299). He has therefore placed more emphasis on treatment outcome, arguing that standardized manual-based treatments are no less effective than formulation-based individualized therapy (Wilson, 1996). Unfortunately, there is a paucity of research investigating the link between formulation and treatment outcome (BPS, 2011).

Future research on the reliability of case formulations

1. Future research should utilize reliability statistics that control for chance agreement, for example, the intraclass correlation coefficient.
2. Future research should use blinded raters to decrease the possibility of bias.
3. To increase ecological validity, future research should aim to recruit participants who use case formulation as part of their professional role. However, as students are often recruited in research, future research should separate professional groups in order to draw comparison between levels of qualification and experience.

4. With regard to formulation data, it has been argued that the use of transcripts may lead to scientific bias with researchers selecting particular cases or small samples (Barber & Crits-Christoph, 1993). As explained previously, this may also lead to important non-verbal cues being missed. To provide more ecological validity, future research should use audiovisual material.
5. Developing formulations in teams is likely to increase reliability. Although formulations can be developed as part of a multi-disciplinary team (BPS, 2011; Johnstone & Dallos, 2006, 2014), such research may have limited applicability to clinical practice where clinicians work alone. Therefore, future research should compare the reliability of two or more independent formulators.
6. The BPS (2011) highlights best practice guidelines for the use of formulation, which includes grounding formulation in an appropriate level of assessment. This is likely to include information from multiple sources, such as assessment measures and clinical interview. In this way, information can be triangulated to provide a comprehensive understanding of the client. Although two studies (1 and 5) used more than one source of material, that is, video recordings and assessment measures, most studies used only one. Future research may benefit from examining differences in reliability when participants are provided with more than one source of client information.

Limitations

One potential limitation of the current review concerns the broad inclusion criteria, which may have introduced heterogeneity into the sample, potentially affecting the results. This is one possible reason why the range of reliability was so varied, from virtually none to substantial. As can be seen from the data extraction table (Table 1), there were a variety of disorders within the client samples and a range of professions in the formulator and rater samples. It is possible that narrower inclusion criteria may have affected the levels of agreement and reliability, and subsequently the conclusions that can be made. Furthermore, we only included peer-reviewed journal articles written in the English language, which may represent a source of selection and publication bias.

Conclusion

This review has shed light on the reliability of case formulation and demonstrated that it can be achieved through a range of psychological modalities. However, this review has also highlighted a fairly under-researched area for a skill so pertinent to the profession of clinical psychology and requires further investigation.

References

Aston, R. (2009). A literature review exploring the efficacy of case formulation in clinical practice. What are the themes and pertinent issues? *The Cognitive Behaviour Therapist, 2*,63–74. doi:10.1017/S1754470X09000178

Barber, J. P., & Crits-Christoph, P. (1993). Advances in measures of psychodynamic formulations. *Journal ofConsulting and Clinical Psychology*, *61*,574–585. doi:10.1037/0022-006X.61.4.574

Barber, J. P., Luborsky, L., Crits-Christoph, P., & Diguer, L. (1995). A comparison of core conflictual relationship themes before psychotherapy and during early sessions. *Journal ofConsulting and Clinical Psychology*, *63*, 145–148. doi:10.1037/0022-006X.63.1.145

Beck, A. T. (1976). *Cognitive therapy and the emotional disorders*. NewYork, NY: Meridian Books.

Bieling, P. J., & Kuyken, W. (2003). Is cognitive case formulation science or science fiction? *Clinical Psychology: Science and Practice, 10*, 52–69. doi:10.1093/clipsy.10.1.52

Bland, J. M., & Altman, D. G. (1990). A note on the use of the intraclass correlation coefficient in the evaluation of agreement between two methods of measurement. *Computers in Biology and Medicine, 20*, 337–340. doi:10.1016/0010-4825(90)90013-F

British Psychological Society (2008). *Generic professional practice guidelines*. Leicester, UK: Author.

British Psychological Society (2011). *Good practice guidelines on the use of psychological formulation*. Leicester, UK: Author.

Butler, G. (2006). The value of formulation: A question for debate. *Clinical Psychology Forum, 160*, 9–12.

CASP (2004). *Critical Appraisal Skills Programme*. Retrieved from http://www.casp-uk.net/

Caston, J., & Martin, E. (1993). Can analysts agree? The problems of consensus and the psychoanalytic mannequin: II. Empirical tests. *Journal of the American Psychoanalytic Association, 41*, 513–548. doi:10.1177/000306519304100209

Cohen, J. (1960). A coefficient for agreement for nominal scales. *Educational and Psychological Measurement, 20*, 37–46. doi:10.1177/001316446002000104

Collins, W. D., & Messer, S. B. (1991). Extending the plan formulation method to an object relations perspective: Reliability, stability, and adaptability. *Journal of Consulting and Clinical Psychology, 3*, 75–81. doi:10.1037/1040-3590.3.1.75

Corrie, S., & Lane, D. A. (2010). *Constructing stories, telling tales: A guide to formulation in applied psychology*. London, UK: Karnac.

Crits-Christoph, P., Luborsky, L., Dahl, L., Popp, C., Mellon, J., & Mark, D. (1988). Clinicians can agree in assessing relationship patterns in psychotherapy: The core conflictual relationship theme method. *Archives of General Psychiatry, 45*, 1001–1004. doi:10.1001/archpsyc.1988.01800350035005

Curtis, J. T., Silberschatz, G., Weiss, J., Sampson, H., & Rosenberg, S. E. (1988). Developing reliable psychodynamic case formulations: An illustration of the plan diagnosis method. *Psychotherapy, 25*, 256–265. doi:10.1037/h0085340

DeWitt, K. N., Kaltreider, N. B., Weiss, D. S., & Horowitz, M. J. (1983). Judging change in psychotherapy: Reliability of clinical formulations. *Archives of General Psychiatry, 40*, 1121– 1128. doi:10.1001/archpsyc.1983.01790090083013

Division of Clinical Psychology (2010). *Clinical psychology: The core purpose and philosophy of the profession*. Leicester, UK: British Psychological Society.

Dudley, R., Park, I., James, I., & Dodgson, G. (2010). Rate of agreement between clinicians on the content of a cognitive formulation of delusional beliefs: The effect of qualifications and experience. *Behavioural and Cognitive Psychotherapy, 38*(2), 185–200. doi:10.1017/S1352465809990658

Eells, T. D. (2007). *Handbook ofpsychotherapy case formulation*. New York, NY: Guilford Press.

Eells, T. D., Horowitz, M. J., Singer, J., Salovey, P., Daigle, D., & Turvey, C. (1995). The role relationship models method: A comparison of independently derived case formulations. *Psychotherapy Research, 5*, 154–168. doi:10.1080/10503309512331331276

Graham, M., Milanowski, A., & Miller, J. (2012). *Measuring andpromoting inter-rater agreement of teacher andprincipalperformance ratings*. Retrieved from http://www.cecr.ed.gov/pdfs/ Inter_Rater.pdf

Haas, M. (1991). Statistical methodology for reliability studies. *Journal of Manipulative and Physiological Therapeutics, 14*, 119–132.

Hartmann, D. P. (1977). Considerations in the choice of interobserver reliability measures. *Journal ofApplied Behavior Analysis, 10*, 103–116. doi:10.1901/jaba.1977.10-103

Hayes, A. F., & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures, 1*, 77–89. Retrieved from http://www.afhayes.com/public/cmm2007.pdf

Johnstone, L., &Dallos, R. (2006). *Formulation inpsychology andpsychotherapy: Making sense of people's problems.* London, UK: Routledge.

Johnstone, L., &Dallos, R. (2014). *Formulation inpsychology andpsychotherapy: Making sense of people's problems* (2nd ed.). London, UK: Routledge.

Kendjelic, M., & Eells, T. (2007). Generic psychotherapy case formulation training improves formulation quality. *Psychotherapy: Theory Research, Practice, Training, 44,* 66–77. doi:10.1037/0033-3204.44.1.66

Kuyken, W., Fothergill, C. D., Musa, M., & Chadwick, P. (2005). The reliability and quality of cognitive case formulation. *Behaviour Research and Therapy, 43,* 1187–1201. doi:10.1016/j. brat.2004.08.007

Kuyken, W., Padesky, C. A., & Dudley, R. (2009). *Collaborative case conceptualisation.* New York, NY: Guilford Press.

Luborsky, L. (1977). Measuring a pervasive psychic structure in psychotherapy: The core conflictual relationship theme. In N. Freedman & S. Grand (Eds.), *Communicative structures andpsychic structures (pp.* 367–395). New York, NY: Plenum Press.

Luborsky, L., & Diguer, L. (1998). The reliability of the core conflictual relationship theme method measure: Results from eight samples. In L. Luborsky & P. Crits-Christoph (Eds.), *Understanding transference: The core conflictual relationship theme method* (2nd ed., pp. 97–107). New York, NY: Basic Books.

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica, 22*, 276–282. Retrieved from http://www.biochemia-medica.com/2012/22/276

Mumma, G. H. (2011). Current issues in case formulation. In P. Sturmey & M. McMurran (Eds.), *Forensic case formulation (pp.* 33–60). Chichester, UK: Wiley-Blackwell.

Mumma, G. H., & Smith, J. L. (2001). Cognitive-behavioral-interpersonal scenarios: Interformulator reliability and convergent validity. *Journal of Psychopathology and Behavioral Assessment, 23*, 203–221. doi:10.1023/A:1012738802126

Muran, J. C., Segal, Z. V., & Samstag, L. W. (1994). Self-scenarios as a repeated measures outcome measurement of self-schemas in short-term cognitive therapy. *Behavior Therapy, 25*, 255–274. doi:10.1016/S0005-7894(05)80287-4

Parker, I. (2004). Criteria for qualitative research in psychology. *Qualitative Research in Psychology, 1*, 95–106. doi:10.1191/1478088704qp010oa

Perry, J. C., Augusto, F., & Cooper, S. H. (1989). Assessing psychodynamic conflicts: I. Reliability of the idiographic conflict formulation method. *Psychiatry, 52*, 289–301.

Persons, J. B., & Bertagnolli, A. (1999). Inter-rater reliability of cognitive-behavioral case formulation of depression: A replication. *Cognitive Therapy and Research, 23*, 271–283. doi:10.1023/A:1018791531158

Persons, J. B., Mooney, K. A., & Padesky, C. A. (1995). Interrater reliability of cognitive-behavioral case formulation. *Cognitive Therapy and Research, 19,* 21–33. doi:10.1007/BF02229674

Popp, C. A., Diguer, L., Luborsky, L., Faude, J., Johnson, S., Morris, M.,...Schaffler, P. (1996). Repetitive relationship themes in waking narratives and dreams. *Journal of Consulting and Clinical Psychology, 64,* 1073–1078. doi:10.1037/0022-006X.64.5.1073

Rankin, G., & Stokes, M. S. (1998). Reliability of assessment tools in rehabilitation: An illustration of appropriate statistical analyses. *Clinical Rehabilitation, 12*, 187–199. doi:10.1191/026921598672178340

Rosenberg, S. E., Silberschatz, G., Curtis, J. T., Sampson, H., & Weiss, J. (1986). A method for establishing reliability of statement from psychodynamic case formulation. *AmericanJournal of Psychiatry, 143,* 1154–1456.

Seitz, P. F. (1966). The consensus problem in psychoanalytic research. In L. Gottschalk & L. Auerbach (Eds.), *Methods of research and psychotherapy (pp.* 209–225). New York, NY: Appleton-Century-Crofts.

Shefler, G., &Tishby, O. (1998). Interjudge reliability and agreement about the patient's central issue in time-limited psychotherapy (TLP) and its relation to TLP outcome. *Psychotherapy Research, 8*, 426–438. doi:10.1080/10503309812331332507

Shrout, P. E. (1998). Measurement reliability and agreement in psychiatry. *Statistical Methods in Medical Research, 7,* 301–317. doi:10.1177/096228029800700306

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86,* 420–428. doi:10.1037/0033-2909.86.2.420

Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research & Evaluation, 9,* 1–19. Retrieved from http://PAREonline.net/getvn.asp?v=9&n=4

Tarrier, N., & Calam, R. (2002). New developments in cognitive-behavioural case formulation. *Behavioural and Cognitive Psychotherapy, 30,* 311–328. doi:10.1017/S1352465802003065

Wells, G. A., Shea, B., O'Connell, D., Peterson, J., Welch, V., Losos, M., & Tugwell, P. (2010). *The Newcastle-Ottawa Scale (NOS) for assessing the quality of nonrandomized studies in meta-analyses.* Retrieved from http://www.ohri.ca/programs/clinical_epidemiology/oxford.asp

Wilson, G. T. (1996). Manual-based treatments: The clinical application of research findings. *Behaviour Research Therapy, 34,* 295–314. doi:10.1016/0005-7967(95)00084-4

Wilson, F. E., & Evans, I. A. (1983). The reliability of target-behavior selection in behavioral assessment. *Behavioral Assessment, 5,* 15–32.