The Cognitive Behaviour Therapist, 2011, **4**, 89–100
doi:10.1017/S1754470X11000067

REVIEW

Observing competence in CBT supervision: a systematic review of the available instruments

Derek L. Milne[1]* and Robert P. Reiser[2]

[1]*School of Psychology, Newcastle University, UK*
[2]*Palo Alto University, CA, USA*

Received 3 March 2011; Accepted 17 October 2011

Abstract. Government policy, like evaluations of clinical practice, indicates the growing importance of supervision in fostering practitioner development and in improving the fidelity of therapies. However, instruments with which to measure competent supervision are often problematic, thereby hampering these key activities (e.g. they are rare, rely on self-ratings by participants, and psychometric data can be limited). To contribute to progress, this paper reviews the current options for measuring competent clinical supervision by means of direct observation, a favoured approach within cognitive behaviour therapy (CBT). We systematically reviewed 10 existing instruments that were designed to observe and quantify competent supervision, focusing on three broad criteria for sound measurement (i.e. an instrument's Design, Implementation, and Yield: DIY). Suggestions for future research on instruments that can fulfil the functions that are provided distinctively through direct observation are outlined.

Key words: Clinical supervision, competence, instruments, observation.

Introduction

Clinical supervision is regarded as an essential component for delivering modern, patient-centred, high-quality mental health care. For example, within the British National Health Service (NHS), there is recognition that 'we need to develop and maintain the infrastructure comprising skilled mentors and supervisors' (Department of Health, 2001, p. ix). The same government department recognizes that 'regular clinical supervision will encourage reflective practice and needs to be available to all staff . . . the importance of staff training and support cannot be underestimated' (Department of Health, 2004, p. 35). Similarly, in the USA supervision has been viewed as vital to the fidelity of evidence-based treatments (Baer *et al*. 2006), consistent with the growing emphasis on evidence-based practice within the American Psychological Association (APA Presidential Task Force, 2006). Although there is consensus on the importance of clinical supervision, there is actually no definitive statement of what is meant by the term. Within the NHS, supervision is defined as

*Author for correspondence: D. L. Milne, Ph.D., School of Psychology, Ridley Building, Newcastle University, Newcastle upon Tyne NE1 7RU, UK (email: d.l.milne@ncl.ac.uk).

> an intervention provided by a more senior member of a profession to a more junior member or members of that same profession. This relationship is evaluative, extends over time, and has the simultaneous purposes of enhancing the professional functioning of the more junior person(s), monitoring the quality of professional services offered to the clients, she, he, or they see, and serving as a gatekeeper for those who are to enter the particular profession (Department of Health, 1993, p. xx).

The policy of supporting and developing competent supervisors is founded on research evidence which indicates that competent supervision is effective. For example, meta-analyses of controlled clinical outcome trials (examining the effectiveness of collaborative care for depressed patients in primary care) indicated that access to regular and planned supervision was related to more positive clinical outcomes (Gilbody *et al*. 2006). Such reviews are supported by randomized controlled evaluations of clinical supervision in relation to its role in developing therapeutic alliances and contributing to symptom reduction (Bambling *et al*. 2006), and in relation to improving the transfer of training into therapeutic practice (Gregoire *et al*. 1998; Heaven *et al*. 2005). Given its importance, it is comforting to know that supervision is also a highly acceptable approach to the supervisors and supervisees involved, ranking as one of the most influential factors in clinical practice (Lucock *et al*. 2006).

However, despite this endorsement of clinical supervision, there is a problem regarding its measurement: 'one of the most pernicious problems confronting supervision researchers is the dearth of psychometrically sound measures specific to a clinical supervision context' (Ellis & Ladany, 1997, p. 488). According to these authors, studies designed to develop new instruments 'should adhere to traditional standards of scientific research that include explicating the theoretical basis and research hypothesis' (p. 488) that underpin the development work. In a comprehensive review, Ellis & Ladany (1997) concluded that there were no instruments designed to measure competence in clinical supervision that they could recommend. The need for valid and reliable supervision instruments is also highlighted in a summary review of the state of psychotherapy supervision research, in which Watkins (1998) identifies ten key needs: 'Thus, one of the most pressing needs for psychotherapy supervision research in the next century remains the development and establishment of reliable, valid criterion measures to guide supervision research' (Watkins, 1998, p. 94). A decade on the situation appears not to have improved: when Ellis *et al.* (2008) replicated their earlier review to include seven new instruments, they again concluded that 'there continues to be a paucity of psychometrically valid and reliable instruments' (p. 492).

This lack of valid and reliable instruments for evaluating supervision and establishing competence in supervisory practices is a particularly serious deficiency. The concept of competence lies at the heart of modern professional training and licensing (Falender & Shafranske, 2008; Kenkel & Peterson, 2010). It also underpins the commissioning of training and of services, and affords the means to develop accountable, evidence-based clinical services that are based on the explicitly defined technical, cognitive and emotional attributes that constitute safe and effective mental health practice (Epstein & Hundert, 2002; Falender & Shafranske, 2004; Roth & Pilling, 2008); i.e. it has social validity. Epstein & Hundert (2002) define competence as: 'The habitual and judicious use of communication, knowledge, technical skills, clinical reasoning, emotions, values, and reflection in daily practice for the benefit of the individual and community being served' (p. 226). In this context, it is somewhat concerning that Ellis & Ladany (1997) were only able recommend two instruments within

the supervision literature, both of which were self-report questionnaires and neither of which directly assessed competence (i.e. The Relationship Inventory: Schacht *et al*. 1988; and The Role Conflict and Role Ambiguity Inventory: Olk & Friedlander, 1992). The situation had not altered at the time of their later review (Ellis *et al*. 2008).

We propose to address this situation by conducting a systematic review of available instruments measuring psychotherapy supervision, taking into account recent research and so updating the systematic reviews of Ellis & Ladany (1997) and Ellis *et al*. (2008). Second, we do so by means of a different but complementary method, the 'best-evidence synthesis' (BES) approach to the systematic review (Petticrew & Roberts, 2006). There are precedents for the BES approach, ones that have yielded comparatively optimistic and constructive accounts of the supervision literature, and therefore carry rather different research and service development implications from the exhaustive review approach (see e.g. Milne & James, 2002; Milne *et al*. 2010). Third, we focus purely on instruments that utilize direct observation. This is because direct observation is the dominant method of assessing competence within professional training programmes, and a focus on specific, observable behaviours was deemed 'most useful' in a review of the assessment options within clinical supervision (Lambert & Ogles, 1997, p. 440). Fourth, we extend the Ellis & Ladany (1997) and Ellis *et al*. (2008) reviews by extending beyond their traditional psychometric criteria, by adding an emphasis on the practicalities and benefits of alternative instruments (i.e. adding pragmatic criteria). To do this we adopt the 'design, implementation, yield' (DIY) dimensions advocated by Barkham *et al*. (1998). This more rounded appraisal is desirable as it takes due account of important practical determinants of the utilization of instruments (e.g. instrument availability, the utility of data), factors that might best be viewed alongside the traditional psychometric criteria as equally necessary conditions for sound measurement practice.

Rationale for direct observation

Direct observation has been widely accepted as an established method in supervision (e.g. integral to cognitive-behavioural supervision; Liese & Beck, 1997), one that is considered 'especially effective' in training therapists (Watkins, 1997, p. 337). It is also endorsed as a method for researching supervision (Lambert & Ogles, 1997). For instance, Bernard & Goodyear (1998) recognized the need for more rigorous evaluations of adherence and competence, while Falender & Shafranske (2004) stressed the need to assess competencies through means such as performance assessments. Finally, Falendar & Shafranske (2010) echoed this same sentiment in an update on competency-based education and training models:

> This work will be advanced by employing a competency-based model of supervision (Falender & Shafranske, 2004; Kaslow *et al.* 2004) in which the component aspects of competencies (i.e. knowledge, skills, attitudes/values) are operationally identified, approaches to self-assessment and evidence-based assessment formative assessment are developed . . . and a range of learning strategies are employed (p. 45).

A second reason for emphasizing direct observation is the methodological desirability of using a variety of assessment methods, ones that draw on different perspectives and operationalizations, so that we can better triangulate the selected phenomenon (Roth

& Fonagy, 1996; Kazdin, 1998). Fittingly, the conventional range of methods includes permanent products (archival data), self-report information (interviews, questionnaires), audits, and direct observation (Milne, 2007). Within the general field of staff training and development, one or more of these methods is normally applied in order to establish the effect of a trainer or supervisor on the learners' knowledge, skills or attitudes (Kraiger *et al*. 1993). In principle, such an instrument may also be used to quantify the resource aspects of supervision (i.e. the structures, resources or efforts that are applied), and/or the content, the procedures, the processes, or efficiency of supervision (Milne, 2007). We also rely on the conventional principle that multiple methods of measurement increase the likelihood that measurement will prove valid. There is also empirical evidence that different evaluation criteria (such as direct observation) make a complementary contribution to evaluation. For example, the meta-analysis conducted by Alliger *et al*. (1997) considered the degree to which the different evaluation methods were correlated across a sample of 34 training studies. Their findings indicated that 'at most, there are modest correlations between the various types of training criteria' (p. 350). They concluded that 'it might be best to use multiple criteria with minimal overlap, to get a more complete picture of the effect of training' (pp. 352–353). Based on this logic, direct observation is likely to prove a key method within a comprehensive or systematic evaluation of supervision, although of course measurement should ultimately be selected thoughtfully, as appropriate, to assess the critical features of a given study.

Finally, in addition to offering a socially valid, complementary and more objective perspective, direct observation has the advantage of providing an external demonstration that a practitioner is competent (Prescott *et al*. 2002). These authors noted the increasing salience of this issue, as the public, the media and government have become more concerned about the quality of clinical practice. This increases the onus on designing rigorous systems for assessing competence. Therefore, we conclude that there is a clear need to review existing instruments. Next, however, we address the question of how such instruments should be evaluated: which criteria are appropriate?

Psychometric and other criteria

A well-established criterion when evaluating an instrument is its rigour, as reflected in its psychometric properties (Ellis & Ladany, 1997; Kazdin, 1998). Based on these authors' accounts and on general texts on psychometric evaluation, the relevant criteria for judging an observational instrument can be said to include several forms of validity (hypothesis, face, content, predictive, concurrent) and reliability (inter-rater agreement, test–retest reliability). Of these, perhaps the least well-known is hypothesis validity (Wampold *et al*. 1990). This concerns the extent to which an instrument properly operationalizes the theoretical basis underpinning a tool, so that appropriate inferences can be drawn and tested. Logically, it is as necessary as any of the other psychometric criteria.

According to Barkham *et al*. (1998), these psychometric criteria are termed 'design' issues. They argue that one should also consider two other broad dimensions for the evaluation of instruments, namely their 'implementation' and 'yield'. Implementation concerns pragmatic issues, such as whether an instrument is available, whether it can be applied readily, whether or not extensive training is required, and whether the scoring and interpretation aspects are straightforward. For example, Watkins (1997, p. 440) commented that 'the observer rating

scales developed by Rogers's students . . . are time-consuming and expensive to employ properly'. In this sense, an instrument may satisfy psychometric criteria but be impractical to apply.

A third important dimension for judging an instrument is the extent to which it 'yields' information that has some utility. The emphasis on yield is shared by other authors, albeit using different terms, such as 'impact' and 'consequential validity' (e.g. Vleuten & Schuwirth, 2005). Barkham *et al*. (1998) pinpointed one kind of yield, namely that an instrument can provide the basis for outcome benchmarking. Logically, it seems to us that there are additional ways to judge the yield of an instrument. For example, following the fidelity framework (Bellg *et al*. 2004), one might also define yield in terms of the ability of an instrument to profile, audit, evaluate and assess systemic impact. To explain these functions, one might use observation to 'profile' supervision, i.e. to furnish data on the strengths and weaknesses of that supervision, in relation to a given theory or approach. Technically, it is a form of corrective feedback (see e.g. Iberg, 1991). By comparison, 'auditing' supervision through observation provides data on the extent to which a supervisor is adhering to the standards or criteria of a specific therapeutic approach (see e.g. Henggeler *et al*. 2002). In essence, the profiling function represents formative evaluation, while the auditing function equates to summative evaluation. The third potential yield that might be achieved through observation is that of 'evaluating' supervision. In this instance the observational data allow objective judgements to be made about the skill of the supervisor (e.g. on the continuum from novice to expert; Dreyfus & Dreyfus, 1986). This can also be calculated relative to other supervisors, as per the outcome benchmarking logic (Barkham *et al*. 1998). An example of such an assessment tool is the Clinical Teaching Effectiveness Instrument (Copeland & Hewson, 2000), which subsumes a number of supervisory behaviours (e.g. 'establishes a good learning environment'). The 15 items are rated on a proficiency scale, ranging from 'poor' to 'superb'. Last, we suggest that yield can be defined in terms of 'impacting'; assessment can provide data indicative of whether supervision had the intended effect on the supervisee and/or the client, or through other dimensions of a service system (e.g. enhancing quality of care within a specialist service). For example, Bambling *et al*. (2006) linked an observational measure of the supervisors' adherence to CBT or psychodynamic approaches to self-report measures of therapeutic alliance and symptom change.

In summary, direct observation plays a distinctive role in support of valid measurement, but we argue that, to be judged valuable, an observational instrument should be considered in relation to three dimensions of rigorous and relevant measurement, those of design, implementation and yield (DIY; Barkham *et al*. 1998). By this reasoning, we next conduct a review of observational instruments that have been designed to assess competence in clinical supervision, using this DIY framework.

Method

Review of existing instruments for measuring competence in supervision

We searched for scientific papers that met the following criteria for inclusion: published in the last 30 years (i.e. up to the end of 2010); designed to measure competent supervision [three or more 'technical skills' or other domains from the Epstein & Hundert (2002) definition]; utilizing direct observation as the method of data collection; based on a quantitative approach; and measuring observable competencies. Using these criteria, a search strategy was

undertaken which included an electronic search of databases. For the electronic search, the following databases were utilized: Web of Knowledge, PsycINFO, Ovid, Embase, Medline and Cochrane Review. We searched for the following terms and their combinations: 'clinical supervision', 'rating', 'competence', 'observational', 'self-report', 'instrument' and 'CBT'. In addition, we browsed in libraries, consulted with experts, and searched references from studies meeting our criteria. In this way, 10 instruments were located, as set out in Table 1. These papers were then systematically reviewed, using an *ad hoc* manual that operationalized the DIY criteria, following Ellis & Ladany (1997) and Barkham *et al*. (1998). The manual lists five validity criteria, two reliability criteria, four implementation criteria and four yield criteria (a copy of the instrument evaluation manual is available from the first author). Each criterion is defined within the manual, so that the papers could be coded as reliably as possible (although we did not attempt to assess inter-rater reliability directly in the present study). This systematic review approach is therefore scientific in style, by contrast with the traditional narrative review. Like a survey, it emphasizes a methodical, replicable, and relatively bias-free way of deriving the answer to a question (e.g. careful sampling and systematic analysis; Petticrew & Roberts, 2006). Our question was: How well do existing measures of competent clinical supervision stand up to a DIY evaluation?

These review inclusion criteria meant that studies that were otherwise suitable but narrow in their focus were excluded. For example, Parsons & Reid (1995) developed a direct observation tool with the focus purely on one supervision variable, that of providing feedback. Similarly, unpublished measures which did operationalize comprehensive versions of competence in supervision were also excluded (e.g. James *et al*. 2005; Sudak *et al*. 2001). Moreover, papers that described self-ratings, supervisee ratings, and other approaches that were not explicitly based on direct observation by an independent observer were excluded (e.g. Henggeler *et al*. 2002; Bambling *et al*. 2006). Following the application of these inclusion criteria, 10 instruments were located and became the focus of the present review.

The results of this systematic review procedure are summarized in Table 1. They indicate that the 10 surveyed tools met the majority of the 14 criteria (mean criteria satisfied: 9.1; range: 6–11), but that none of the tools meet all of these criteria, including the design (psychometric) criteria. In this sense, we have replicated Ellis & Ladany's (1997) pessimistic account of supervision instruments, for this sample of observational instruments. Also, like them we could identify the best instruments, both of which satisfied 11 criteria (i.e. Parsons & Reid, 1995; Milne *et al*. 2002). We note that all of the criteria were addressed by at least one instrument (range 1–10), giving some content validity to our evaluation system. In terms of the DIY dimensions, it can be seen that the first six 'design' criteria in Table 1 were variably satisfied, with only one study addressing criterion 5, convergent-divergent validity (White & Rudolph, 2000), whereas nine studies reported information on the instruments' content validity and reliability. The strongest overall performance of the 10 instruments was on the DIY dimension of 'implementation', with seven of these satisfying the four relevant criteria. In terms of the four 'yield' criteria, only one measure assessed 'impact' – the effect of supervision on the supervisee (or on other aspects of the service system); half of the instruments did 'evaluate', by appraising the supervisors' competence; seven addressed the 'audit' function of measurement; while all 10 instruments met the 'profile' criterion. This suggests that these criteria were themselves valid.

In conclusion, the data in Table 1 support the view that the available observational instruments are deficient, at least in terms of the design and yield dimensions. Specifically,

Table 1. *Review of observational measures of supervision competencies*

		Psychometric criteria (design features)						Implementation criteria							
		Validity						Reliability				Yield criteria			
Paper	Name of tool (title inferred)	1	2	3	4	5	6	1	2	3	4	1	2	3	4
1. Cherniss (1986)	'Supervisor Behaviour Observation System'	√	√	×	×	×	√	√	√	×	√	√	×	×	×
2. Clark *et al*. (1985)	(Identification of Supervisory Interactional Skills)	√	√	×	×	×	√	√	√	√	√	√	√	×	×
3. Ducharme *et al*. (2001)	(Teaching Skills of Supervisory Staff)	√	√	×	×	×	√	√	√	√	√	√	√	×	×
4. Fleming *et al*. (1996)	(Supervisory Performance Skills)	√	√	√	×	×	√	√	√	√	√	√	√	×	×
5. Komaki *et al*. (1986)	Operant Supervisory Taxonomy and Index	√	√	√	×	×	√	√	√	√	√	√	√	×	×
6. Milne *et al*. (2002)	'Teachers' PETS': Process Evaluation of Training & Supervision	√	√	×	√	×	√	√	×	√	√	√	√	√	√
7. Parsons & Reid (1995)	(Observation of Supervisors Feedback)	√	√	×	√	×	√	√	√	√	√	√	√	√	×
8. Reid *et al*. (2003)	(On The Job Observation & Skill Check)	√	√	×	×	×	√	√	√	√	√	√	√	√	×
9. Shanfield *et al*. (1989, 1992)	Psychotherapy Supervision Inventory	×	√	√	√	×	×	×	√	×	×	√	×	√	×
10. White & Rudolph (2000)	Group Supervisory Behaviour Scale (GSBS)	×	×	√	√	√	√	√	√	√	√	√	×	√	×

Design criteria: 1, Hypothesis validity; 2, Content validity; 3, Construct; 4, Criterion (Predictive) validity; 5, Convergent/Divergent validity; 6, Reliability (inter-rater: √ = correlation of 0.7/70% agreement, or greater).
Implementation criteria: 1, Availability; 2, Application; 3, Training required; 4, Scoring & interpretation.
Yield criteria: 1, Profiling; 2, Auditing; 3, Evaluating; 4, Impacting.
× = No data or fails to satisfy criterion; √ = satisfies criterion.

there appears to be no suitably rigorous and pragmatic instrument available with which to evaluate competent clinical supervision through direct observation.

Discussion

In this paper we conducted a systematic review of supervision instruments that were based on direct observation, updating and extending the review by Ellis & Ladany (1997) and Ellis *et al*. (2008) so as to include the DIY criteria advocated by Barkham *et al*. (1998). This allowed us to take into account the important practical determinants of the utilization of instruments (e.g. instrument availability and the utility of data). We located 10 instruments that met our search criteria, most of which failed to meet a majority of the 14 DIY criteria that we applied, indicating that the field is still deficient in supervision measures that are based on direct observation. Thus, we echo the pessimistic conclusions in the systematic reviews of Ellis & Ladany (1997) and Ellis *et al*. (2008), and underscore the implications for research (e.g. that the lack of valid and reliable supervision instruments is a key impediment to research on supervision; Watkins, 1998).

Limitations of the review and recommendations for future research

The decision to credit DIY criteria when vetting the 10 studies was based on a low threshold: if the authors cited or claimed any of these criteria, using any regular form of analysis (e.g. only assessing reliability by means of percent agreement, without clarifying how this was calculated), then they were credited with satisfying the criterion within Table 1. This relatively relaxed criterion may have inflated the profiles of the 10 reviewed instruments. On the other hand, it could be argued that our attention to validity is disproportionate, as it can be argued that 'validity . . . is not a large concern' (Barlow *et al*. 2009, p. 130) when direct observation is the method of data collection, as the data tend to require little inference, being a sample of a larger class of topographically or functionally equivalent behaviours. Contrast this with traditional self-report instruments that treat the data as a sign of something unobservable, and hence necessitate an inference about the putative inferred underlying variable (e.g. personality traits). Conversely, we should perhaps have accorded greater attention to face validity. Although this is not technically a form of validity (as it only concerns what an instrument appears to measure), it is a desirable aspect in terms of maximizing rapport or cooperation (Anastasi & Urbina, 1997).

In this review we did not assess our inter-rater reliability as reviewers, which is a weakness. However, as we had contemporaneously demonstrated good reliability in relation to a more taxing review, it was assumed that we could code the sampled papers without significant error.

In conclusion, it appears that there is a need for an improved observational instrument with which to measure CBT supervision. Other considerations include the current pressure to measure competence (Department of Health, 2001, 2004), and to assess the extent to which supervision is delivered with fidelity (Bellg *et al*. 2004). An additional pragmatic consideration is brevity, in that the existing observational instrument that has good psychometric qualities and has been most frequently used to measure CBT supervision is time-consuming to apply (it relies on the momentary-time sampling method) and does not provide a rating of competence (i.e. Teachers' PETS; Milne *et al*. 2002). It follows that future

research should seek to develop a brief, supervisory competence rating scale that addresses fidelity issues with improved design, implementation and yield features. This instrument can then contribute to research and practice on the important issue of clinical supervision, as 'advances in knowledge can be expected to increase with advances in criterion measurement' (Lambert & Ogles, 1997, p. 441). These guidelines on a new instrument can be illustrated by reference to one that is currently under development by the present authors (SAGE: Supervision Adherence, Guidance and Evaluation; Milne *et al*. unpublished data). In relation to the 'design' task, SAGE is a 23-item competence rating scale, addressing the supervisory alliance (e.g. interpersonally effective), supervision skills (e.g. agenda-setting), and the supervisee's initial experiential learning within the supervision session (e.g. reflecting). The 23 items were derived from systematic reviews of the supervision literature, to try to ensure content and face validity, and by reference to expert consensus. Four CBT experts in the UK rated SAGE as having good face and content validity. There are also some promising data on inter-rater reliability (e.g. $r = 0.815$, $p = 0.001$). In terms of 'implementation', each of the 23 items in SAGE is rated by an observer, using a seven-point competence rating scale ranging from 'incompetent' to 'expert'. Unlike PETS, this permits rapid scoring. Moreover, SAGE is available free and requires minimal training. Last, regarding the 'yield' criterion, the ratings from SAGE can be used to profile supervision, which can then allow the assessment of supervision fidelity within research, or identify strengths and weaknesses for supervisor development within routine clinical practice.

Acknowledgements

We are grateful to Thomas Cliffe and Rosamund Raine for their collaboration over the measurement of supervision.

Declaration of Interest

None.

Recommended follow-up reading

Wheeler S, Richards K (2007). The impact of clinical supervision on counsellors and therapists, their practice and their clients. A systematic review of the literature. *Counselling & Psychotherapy Research* **7**, 54–65.

References

Alliger GM, Tannenbaum SI, Bennett W, Traver H, Shotland A (1997). A meta-analysis of the relations among training criteria. *Personnel Psychology* **50**, 341–359.

Anastasi A, Urbina S (1997). *Psychological Testing*. NJ: Prentice-Hall.

APA Presidential Task Force (2006). American Psychological Association Presidential Task Force on evidence-based practice in psychology. *American Psychologist* **61**, 271–285.

Baer JS, Ball SA, Campbell BK, Miele GM, Schoener EP, Tracy K (2006). Training and fidelity monitoring of behavioural interventions in multisite addictions research. *Drug and Alcohol Dependence* **87**, 107–118.

Bambling M, King R, Raue P, Schweitzer R, Lambert W (2006). Clinical supervision: its influence on client-rated working alliance and client symptom reduction in the brief treatment of major depression. *Psychotherapy Research* **16**, 317–331.

Barkham M, Evans C, Margison F, McGrath G, Mellor-Clark J, Milne DL, 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.

Barlow DH, Nock MK, Hersen M (2009). *Single-Case Experimental Designs: Strategies for Studying Behavior Change*. Boston: Allyn & Bacon.

Bellg AJ, Borrelli B, Resnick B, Hecht J, Minicucci DS, Ory M, Ogedegbe G, Orwig D, Ernst D, Czajkowski S (2004). Enhancing treatment fidelity in health behaviour change studies: Best practises and recommendations from the NIH Behaviour Change Consortium. *Health Psychology* **23**, 443–451.

Bernard JM, Goodyear RK (1998). *Fundamentals of Clinical Supervision* (2nd edn). London: Pearson.

Cherniss C (1986). Instrument for observing supervisor behaviour in educational programs for mentally retarded children. *American Journal of Mental Deficiency* **91**, 18–21

Clark HB, Wood R, Kuehnel T, Flanagan S, Mosk M, Northup JT (1985). Preliminary validation and training of supervisory interactional skills. *Journal of Organizational Behavior Management* **7**, 95–115.

Copeland HL, Hewson MG (2000). Developing and testing an instrument to measure the effectiveness of clinical teaching in an academic medical centre. *Academic Medicine* **75**, 161–166

Department of Health (1993). A vision for the future. London: NHS Executive.

Department of Health (2001). Working together – learning together: a framework for lifelong learning for the NHS. London: Department of Health.

Department of Health (2004). Organising and delivering psychological therapies. London: Department of Health.

Dreyfus HL, Dreyfus SE (1986). *Mind over Machine: The Power of Human Intuition and Expertise in the Era of the Computer*. Oxford: Blackwell

Ducharme JM, Williams L, Cummings A, Murray P, Spencer T (2001). General case quasi-pyramidal staff training to promote generalization of teaching skills in supervisory and direct-care staff. *Behavior Modification* **25**, 233–254.

Ellis MV, Ladany N (1997). Inferences concerning supervisees and clients in clinical supervision. An integrative review. In: *Handbook of Psychotherapy Supervision* (ed. C. E. Watkins), pp. 447–507. New York: Wiley.

Ellis MV, D'Iuso N, Ladany N (2008). State of the art in the assessment, measurement, and evaluation of clinical supervision. In: *Psychotherapy Supervision: Theory, Research and Practice* (ed. A. K. Hess, K. D. Hess and T. H. Hess), pp. 473–499. Chichester: Wiley.

Epstein RM, Hundert EM (2002). Defining and assessing professional competence. *Journal of the American Medical Association* **287**, 226–235.

Falender CA, Shafranske E (2004). *Clinical Supervision: A Competency-based Approach.* Washington, DC: APA.

Falender CA, Shafranske EP (2008). *Casebook for Clinical Supervision: A Competency-Based Approach*. Washington, D.C: American Psychological Association.

Falender CA, Shafranske E (2010). Psychotherapy-based supervision models in an emerging competency-based era: a commentary. *Psychotherapy Theory, Research, Practice and, Training* **47**, pp. 45–50. Washington, DC: APA.

Fleming RK, Oliver JR, Bolton DM (1996). Training supervisors to train staff: a case study in a human service organization. *Journal of Organizational Behavior Management* **16**, 3–25.

Gilbody S, Bower P, Fletcher J, Richards D, Sutton AJ (2006). Collaborative care for depression: a meta analysis and review of longer-term outcomes. *Archives of Internal Medicine* **166**, 2314–2320.

Gregoire TK, Propp J, Poertner J (1998). Supervisors' role in the transfer of training. *Administration in Social Work* **22**, 1–17.

Heaven C, Clegg J, McGuire P (2005). Transfer of communication skills training from workshop to work place: the impact of clinical supervision. *Patient Education and Counselling* **60**, 313–325.

Henggeler SW, Schoenwald SK, Liao JG, Letourneau EJ, Edwards DL (2002). Transporting efficacious treatments to field settings: the link between supervisory practices and therapist fidelity in MST programmes. *Journal of Clinical Child Psychology* **31**, 155–167.

Iberg JR (1991). Applying statistical process control theory to bring together clinical supervision and psychotherapy research. *Journal of Consulting and Clinical Psychology* **59**, 575–586.

James IA, Blackburn IM, Milne DL, Freeston M (2005). Supervision training and rating scale for cognitive therapy (stars - CT). Newcastle Cognitive Therapy Centre, England.

Kaslow NJ, Borden KA, Collins FL, Forrest L, Illfelder-Kaye J, Nelson PD (2004). Competencies conference: future directions in education and credentialing in professional psychology. *Journal of Clinical Psychology* **60**, 699–712.

Kazdin AE (1998). *Research Design in Clinical Psychology*. Boston: Allyn Bacon.

Kenkel MB, Peterson R.L. (2010). *Competency-Based Education for Professional Psychology*. Washington, D.C.: American Psychological Association.

Komaki JL, Zlotnick S, Jensen M (1986). Development of an operant-based taxonomy and observational index of supervisory behaviour. *Journal of Applied Psychology* **71**, 260–269.

Kraiger K, Ford JK, Salas E (1993). Application of cognitive, skill based and affective theories of learning outcomes to new methods of training evaluation. *Journal of Applied Psychology* **78**, 311–328.

Lambert NJ, Ogles BM (1997). The effectiveness of psychotherapy supervision. In: *Handbook of Psychotherapy Supervision* (ed. C. E. Watkins), pp. 421–446. New York: Wiley.

Liese BS, Beck JS (1997). Cognitive therapy supervision. In: *Handbook of Psychotherapy Supervision* (ed. C. E. Watkins), pp. 114–133. New York: Wiley.

Lucock MP, Hall P, Noble R (2006). A survey of influences on the practice of psychotherapists and clinical psychologists in training in the UK. *Clinical Psychology and Psychotherapy* **13**, 123–130.

Milne DL (2007). Evaluation of staff development: the essential 'SCOPPE'. *Journal of Mental Health* **16**, 389–400.

Milne DL, James IA (2002). The observed impact of training on competence in clinical supervision. *British Journal of Clinical Psychology* **41**, 55–72.

Milne DL, James IA, Keegan D, Dudley M (2002). Teachers PETS: a new observational measure of experiential training interactions. *Clinical Psychology and Psychotherapy* **9**, 187–199.

Milne DL, Reiser R, Aylott H, Dunkerley C, Fitzpatrick H, Wharton S (2010). The systematic review as an empirical approach to improving CBT supervision. *International Journal for Cognitive Psychotherapy* **3**, 278–293.

Olk ME, Friedlander ML (1992). Trainees experience of role conflict and role ambiguity in supervisor relationships. *Journal of Counselling Psychology* **39**, 389–397.

Parsons MB, Reid DH (1995). Training residential supervisors to provide feedback for maintaining staff teaching skills with people who have severe disabilities. *Journal of Applied Behaviour Analysis* **28**, 317–322.

Petticrew M, Roberts H (2006). *Systematic Reviews in the Social Sciences: A Practical Guide*. Oxford: Blackwell.

Prescott LE, Nocini JJ, McKinlay P, Rennie JS (2002). Facing the challenges of competency based assessment of post-graduate dental training: longitudinal evaluation of performance (LEP). *Medical Education* **36**, 92–97.

Reid DH, Rotholz DA, Parsons MB, Morris L, Braswell BA, Green C, Schell RM (2003). Training human service supervisors in aspects of PBS: Evaluation of a state-wide, performance based program. *Journal of Positive Behavior Interventions* **5**, 35–46.

Roth A, Fonagy P (1996). *What Works for Whom? A Critical Review of Psychotherapy Research*. New York: Guilford Press.

Roth AD, Pilling S (2008). Using an evidence-based methodology to identify the competencies required to deliver effective cognitive and behavioural therapy for depression and anxiety disorders. *Behavioural and Cognitive Psychotherapy* **36**, 129–147.

Schacht AJ, Howe HE junior, Berman JJ (1988). A short form of the Barrett–Lennard relationship inventory for supervisor relationships. *Psychological Reports* **63**, 699–706.

Shanfield SB, Mohl PC, Matthews KL, Hetherly V (1989). A reliability assessment of the psychotherapy supervisory inventory *American Journal of Psychiatry* **146**, 1447–1450.

Shanfield SB, Mohl PC, Matthews KL, Hetherly V (1992). Quantitative assessment of the behaviour of psychotherapy supervisors. *American Journal of Psychiatry* **149**, 352–357.

Sudak D, Wright J, Bienenfeld D, Beck J (2001). Cognitive behaviour therapy supervision checklist. Philadelphia: Cognitive Therapy Centre.

Vleuten CPM, Schuwirth LWT (2005). Assessing professional competence: from methods to programmes. *Medical Education* **39**, 309–317.

Wampold BE, David B, Good RH (1990). Hypothesis validity of clinical research. *Journal of Consulting and Clinical Psychology* **58**, 360–367.

Watkins CE (ed.) (1997). *Handbook of Psychotherapy Supervision*. New York: Wiley.

Watkins CE (1998). Psychotherapy supervision in the 21st century. *Journal of Psychotherapy Practice & Research* **7**, 93–101.

White JHD, Rudolph BA (2000). A pilot investigation of the reliability and validity of the Group Supervisory Behavior Scale (GSBS). *The Clinical Supervisor* **19**, 161–171.

Learning objectives

By studying this paper carefully, readers will be able to:
(1) Summarize the argument for direct observation.
(2) Discuss the 'DIY' criteria.
(3) Outline an approach to the systematic review.

https://doi.org/10.1017/S1754470X11000067 Published online by Cambridge University Press