The Cognitive Behaviour Therapist, 2011, **4**, 123–138
doi:10.1017/S1754470X11000079

SAGE: preliminary evaluation of an instrument for observing competence in CBT supervision

Derek L. Milne[1*], Robert P. Reiser[2], Tom Cliffe[1] and Rosamund Raine[1]

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

Received 13 September 2011; Accepted 1 November 2011; First published online 8 December 2011

Abstract. Clinical supervision plays a recognized role in facilitating practitioner development and in promoting therapeutic fidelity, but instruments that can support such activities by measuring competence in supervision are rare and often psychometrically compromised. As part of a research programme intended to raise the empirical status of supervision, we describe the initial psychometric development of a new instrument for observing competence (Supervision: Adherence and Guidance Evaluation; SAGE). This instrument is suitable for measuring CBT supervision, and can be administered by self-rating or by an independent observer. Preliminary tests of the reliability and validity of SAGE suggest that it is a promising tool for evaluating supervision. In addition, SAGE can be applied readily and has good utility. In these respects, SAGE appears to have some advantages over existing instruments, and may therefore provide a basis for enhancing research and practice in CBT supervision. Suggestions for future research on SAGE are outlined, particularly the need for a generalizability analysis.

Key words: Clinical supervision, competence, direct observation, instrument.

Introduction

Clinical supervision is now widely acknowledged to play an essential role in developing and supporting mental health practitioners. For example, in the latest major policy statement from the UK's National Health Service (NHS), it is noted that: 'In high quality services . . . staff receive ongoing training and supervision' (DoH, 2009, p. 26). This emphasis is reflected in the requirements for initial professional training and registration, where supervisors are expected to be registered and thereby to adhere to certain basic standards relating to their supervision (e.g. HPC, 2010). Additionally, following initial training some professions (e.g. clinical psychology; BPS, 2009) and some therapeutic approaches have specified further standards. For instance, The British Association for Behavioural and Cognitive Psychotherapy (BABCP) has Standards of Conduct (BABCP, 2009) which note that members' 'professional development should be facilitated through the receipt of relevant regular supervision' (p. 7).

[*]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).

This policy of developing competent supervisors is supported by the research evidence, which typically 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 (Gregorie *et al.* 1998; Heaven *et al.* 2005). In-depth single-subject ($N = 1$) studies also consistently indicate the effectiveness of supervision (e.g. Milne & Westerman, 2001; Milne & James, 2002). Furthermore, given its importance, it is comforting to know that supervision is 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, the growing acknowledgement of supervision has not been matched by progress in related research. This has been illustrated through a systematic review by Wheeler & Richards (2007), who concluded that only two of the 18 studies that they scrutinized ‘met the criteria to be classified as very good’ (p. 63). More specifically, these 18 studies were only included in their review ‘if a valid and reliable instrument was used’ (p. 55). Their rather damning review is entirely consistent with other reviews, as regards both the general calibre of research and the specific need for improved measurement. To illustrate, in his recommendations on supervision research, Watkins (1998) noted that: ‘Advances in supervision research have occurred over the last two decades, but the need for more rigour in future experimental efforts remains paramount’ (p. 95). He also noted that few studies had used an instrument with any established reliability and validity. Similarly, in a comprehensive methodological review, Ellis & Ladany (1997) concluded that there were no instruments designed to measure competence in clinical supervision that they could recommend: ‘one of the most pernicious problems confronting supervision researchers is the dearth of psychometrically sound measures specific to a clinical supervision context’ (p. 488).

Even though more than a decade has passed since reviewers have noted these deficiencies in the supervision research literature, there have been scant signs of progress. In their concluding remarks to a recent review of psychotherapy-based supervision competencies, Falender & Shafranske (2010) pointed to a continuing deficiency, proposing a competence-based approach as a way forward:

> There is need for revisioning psychotherapy-based supervision in terms of competencies. . . . This work will be advanced by employing a competency-based model of supervision . . . 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 (p. 49).

Although sympathetic to the competence approach, in their introduction to development of a competency-based framework for supervision in the UK, Roth & Pilling (2008) admit to the relative lack of supporting empirical data, suggesting a pragmatic solution:

> Realistically . . . it seems clear that any competence framework would need to be developed by integrating empirical findings with professional consensus, in this way articulating the sets of activities usually assumed to be associated with better learning outcomes (p. 6).

https://doi.org/10.1017/S1754470X11000079 Published online by Cambridge University Press

One of the key elements in their competencies map for CBT includes 'An ability to use recordings/direct observation to monitor the supervisee's ability to implement CBT techniques ... using appropriate instruments'. In summary, clinical supervision appears to promote effective therapy, although our understanding of this link is hampered by the generally poor measurement of supervision. However, there is a growing acceptance of a competency-based approach (Newman, 2010).

The shortage of valid and reliable instruments for evaluating supervision and for establishing competence in supervisory practices is a particularly serious deficiency, as the concept of competence lies at the heart of modern professional training and registration (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 explicitly defining the technical, cognitive and emotional attributes that constitute safe and effective mental health practice (Falender & Shafranske, 2004; Epstein & Hundert, 2006; Roth & Pilling, 2008). Epstein & Hundert (2006) 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).

Although the above reviews suggest that almost no suitable instruments exist with which to measure competence in supervision, a recent systematic review of observational instruments actually located ten tools, some with good psychometric data (Milne & Reiser, 2011). Rigour in an instrument is reflected in its psychometric properties (Ellis & Ladany, 1997; Kazdin, 1998; Madson & Campbell, 2006), and the relevant criteria include several forms of validity (hypothesis, face, content, predictive, concurrent), and of 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 is explicitly derived from a theoretically consistent construct, so that appropriate inferences can be drawn and tested, leading potentially to theory development. Logically, it is as necessary as any of the other psychometric criteria, a 'strong rope' linking theory to hypotheses, elegant statistical testing and research findings.

In terms of rigour, the best two instruments identified in the Milne & Reiser (2011) review were reported by Milne *et al.* (2002), and by Parsons & Reid (1995). A number of previous studies of supervision have used the Milne *et al.* (2002) direct observation instrument 'Teachers PETS' (Process Evaluation of Training and Supervision) to evaluate the effectiveness of clinical supervision. For example, Milne & Westerman (2001) studied the effects of bi-weekly consultation on the clinical supervision of three supervisees over an 8-month period, using a multiple baseline $N = 1$ design. Video recordings of selected sessions were coded by independent observers, blind to the consultancy phases. Supervisory behaviours and supervisee reactions were coded at 15-s intervals (i.e. the momentary time sampling method; MTS), to determine the relative frequency of each supervisory behaviour over the three phases of the study. Results of the study suggested that the targeted supervisor behaviours increased in frequency over time, across supervisees and phases (especially guiding experiential learning). In a second $N = 1$ study, Milne & James (2002) analysed the impact of consultancy on the effectiveness of routine clinical supervision for one supervisor and six supervisees. Video recordings were coded by three observers and the results indicated

https://doi.org/10.1017/S1754470X11000079 Published online by Cambridge University Press

that the profile of supervisory behaviours also improved, in terms of more experiential learning, a greater range of learning methods, more experimenting, and less time spent passively reflecting.

However, the MTS method is time-consuming and requires extensive training. Moreover, the results produced by 'Teachers PETS' do not explicitly quantify competence. Similarly, although highly reliable, the competence checklist used by Parsons & Reid (1995) is limited to one supervisory behaviour, that of providing feedback (although eight elements of feedback were observed, such as 'praise' and 'error identification'). This restricted specification of supervisory competence limits the utility of their instrument for general evaluations of the wider spectrum of supervisory behaviours. For example, in PETS the item 'feeding back' is one of 18 supervisory behaviours. Therefore, the best tools in the Milne & Reiser (2011) review have significant shortcomings, justifying the development of a new instrument. This tool should be quicker to complete, enable easier training of raters, and provide an explicit rating of competence in supervision through using an established competence scale, such as that of Dreyfus & Dreyfus (1986).

These pragmatic criteria raise the question of what constitutes a 'good' instrument. The traditional reply has been 'good reliability and validity', but these criteria are insufficient. That is, although an instrument like 'Teachers PETS' has good reliability and validity, it is less impressive in terms of its 'implementation' and 'yield'. Implementation concerns pragmatic issues, such as whether an instrument is available, whether it can be applied readily, and whether or not extensive training is required (Barkham *et al.* 1998). In this sense, an instrument may satisfy psychometric criteria but be impractical to apply. Yield refers to whether the information that an instrument produces has some utility (e.g. outcome benchmarking). The emphasis on yield is shared by other authors, albeit using different terms, such as 'impact' and 'consequential validity' (e.g. Vleuten & Schuwirth, 2005). 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 impact. To explain these functions, one might use observation to 'profile' supervision, in terms of highlighting the strengths and weaknesses of that supervision, in relation to a given 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 approach (e.g. see Henggeler *et al.* 2002). A 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). 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. For example, Bambling *et al.* (2006) linked an observational measure of the supervisors' adherence (to one of two approaches, including CBT) to self-report measures of therapeutic alliance and symptom change.

In summary, the present paper addresses the need for an instrument with which to measure competence in supervision by presenting a new tool, one that is suitable for evaluating competence in CBT supervision. The primary method of administering SAGE was through direct observation, although we also created a supervisor self-rating version and a version that allowed the supervisee to rate the supervisor. Our final two objectives are to apply the criteria

https://doi.org/10.1017/S1754470X11000079 Published online by Cambridge University Press

relating to an instrument's implementation and yield, so providing a rounded but preliminary 'DIY' assessment of the new tool (Barkham *et al*. 1998).

Method and Results

For the sake of clarity, details of the various methods that were used to assess SAGE are now provided alongside the respective results (there are a series of six related psychometric analyses). These six analyses are concerned with the 'design' aspects of SAGE, namely: content validity, hypothesis validity, construct validity, criterion validity, discriminant validity and reliability (inter-rater and internal consistency). The 'implementation' and 'yield' dimensions are considered in the Discussion section.

Content validity

The individual items within SAGE reflect the assumption that the supervisory relationship plays a vital moderating role, as stressed by many experts (e.g. Watkins, 1997; Bernard & Goodyear, 2004; Falender & Shafranske, 2004). Continuing this appeal to the applied psychology literature (see Milne, 2007, for the reasoning-by-analogy argument that underpins this work), we combined the findings from the supervision literature on the supervisory relationship or 'alliance' (e.g. Efstation *et al.* 1990) with the related clinical psychology literature (e.g. Norcross, 2001, 2002), as others have done previously (e.g. Bordin, 1983). In addition, we drew on the most closely related existing tools: the 23 items of SAGE were derived from two existing instruments, 'Teachers' PETS' (Milne *et al.* 2002) and 'CBT STARS' (James *et al.* 2005). Teachers' PETS was originally designed as a way to observe and record a range of 'leadership' behaviours, and derived its items from a search of the applied psychology literature, including instruments designed to assess different forms of leadership (e.g. teachers, athletic coaches, clinical supervisors). PETS consistently demonstrated good reliability and validity across several studies (Milne, 2009). Further evidence that this strategy yielded items that duly mapped onto clinical supervision was found within the content and divergent validity assessments (see below). Like PETS, SAGE only attempts to measure one aspect of supervision (i.e. the 'formative' function of facilitating learning; Proctor, 1988). However, as set out in the Introduction, these instruments differ in several ways, particularly that SAGE is explicitly a competence-rating tool.

Second, having compiled these items, a form was developed to assess the face and content validity of SAGE, based on the definition by Anastasi & Urbina (1997). It consisted of four questions: Was the SAGE manual easy to read? Does the manual aid your understanding of supervision? Would it serve to raise standards? How would you judge the validity of the manual, overall? Each question was rated on a 3-point Likert scale, with 1 indicating 'not yet acceptable', 2 indicating 'acceptable', and 3 indicating 'good'. Six CBT experts in the UK, known to the first author (i.e. a convenience sample) were contacted and asked to read through the SAGE manual, before completing the evaluation form. Four of these experts duly contributed ratings.

Overall, the four experts rated SAGE as having 'good' face and content validity, being judged easy to read, aiding the reader's understanding of supervision, serving to raise standards in the field, and in terms of its overall validity (mean = 2.67, S.D. = 0.49).

https://doi.org/10.1017/S1754470X11000079 Published online by Cambridge University Press

Table 1. *A summary of the items contained within SAGE*

SAGE items	Brief definition
Common factors	
1. Relating	Core conditions, ‘restorative’
2. Collaborating	Alliance
3. Managing	Scaffolded, optimal challenge, ‘normative’
4. Facilitating	Improving grasp (including perplexity)
Supervision cycle	
5. Agenda setting	Needs-led/developmental objectives
6. Demonstrating	Modelling
7. Discussing	Review, disagree, problem solving
8. Evaluating	Closely monitor (e.g. clinical data)
9. Experiencing	Expressing and processing affective aspects
10. Feeding back (giving)	Offer praise, strengths/weaknesses
11. Feeding back (receiving)	Elicit (e.g. helpful events/transfer)
12. Formulating	Analysis, synthesis, explanation
13. Listening	Attending and summarizing
14. Observing	Live/tape material
15. Prompting	Reminders and cues
16. Questioning	Gather information, raise awareness
17. Teaching	Informing/educating (symbolic)
18. Training	Experiential learning (e.g. role play)
Supervisees’ cycle	
19. Experiencing	Awareness, identification and processing of affect (assimilation)
20. Reflecting	Summarizing and integrating subjective material
21. Conceptualizing	Integrating objective material (e.g. theories/findings)
22. Planning	Decision-making about actions
23. Experimenting	Enacting plans (in and out of supervision, e.g. trial-and-error learning through role play/reality checking

The forms also contained several comments. These included recommendations for future development of the tool (e.g. ‘Useful to have some video examples of a low, good and high ratings, so that the rater has a standard to compare with’; ‘Provides a clear conceptual model of supervision competence’; ‘It looks as if it would be quite easy to use with good face validity, which may lead to raising of standards’).

Hypothesis validity

We sought to develop a tool with which to observe the competence of supervision in relation to CBT and the closely related approach of evidence-based clinical supervision (EBCS; Milne, 2009), one that was explicitly grounded in relevant theory (Kolb, 1984). The resulting instrument, SAGE, was a 23-item instrument, as summarized in Table 1.

The theory underpinning SAGE assumes that a competent supervisor will utilize a range of supervisory behaviours (e.g. goal-setting and corrective feedback) to enable the supervisee to engage in experiential learning (i.e. a combination of action, reflection, conceptualization and experiencing, as set out in Kolb, 1984, and elaborated in Milne, 2009), within the context of

https://doi.org/10.1017/S1754470X11000079 Published online by Cambridge University Press

an effective supervisory relationship. Based on this reasoning, we hypothesized a three-factor structure, which we labelled 'common factors', 'supervision cycle' and 'supervisees' cycle' (see Table 1).

Each SAGE item is defined in a coding manual and scored on a competence rating scale that ranges from 'incompetent' to 'expert', based on the Dreyfus model (Dreyfus & Dreyfus, 1986). SAGE is an 'event recording' or 'global rating' tool (i.e. ratings are based on observing a sample of behaviours, thus providing an overall judgement about the quality of the sample). SAGE also has a qualitative feedback section to record the details of the observed supervision, and to make suggestions on ways to improve the observed sample of supervision.

Construct validity

We conducted a factor analysis of SAGE, based on self-ratings by 176 qualified mental health professionals who participated in a one-day workshop on clinical supervision. These ratings were made at the start of the workshop, based on their supervision experiences over the past year. They were not trained in the use of the scale, only receiving a brief outline plus an opportunity to ask questions. All were employed within one NHS Trust; most were mental health nurses, with the majority of participants being female, aged between 30 and 50 years. They were experienced supervisors and supervisees, and represented a range of theoretical orientations, although most adopted a CBT approach to their practice.

We predicted that we would obtain the three-factor solution outlined above (i.e. 'common factors', 'supervision cycle' and 'supervisees' cycle'). Data were analysed using SPSS Statistics, version 17.0 (SPSS Inc., USA). The factor analysis used principal axis factoring; rotation, had it proved necessary, would have been direct oblimin. The Kaiser–Meyer–Olkin measure of sampling adequacy was 0.96. Examination of the eigenvalues (19.06, 1.15, 0.71, 0.62 for the first four, respectively) and the Scree plot indicated a single factor, supervisory competence, accounting for 76.6% of the variance. Internal consistency was 0.98. Our prediction was therefore rejected.

Criterion (predictive) validity

For this part of the instrument evaluation we reasoned that EBCS should have a significantly greater effect on the supervisees' learning than CBT supervision (i.e. supervision-as-usual). This was because it was explicitly grounded in theories of how experiential learning is facilitated, and drew on supervision methods with evidence to support their effectiveness (Milne, 2009). Although criterion validity is typically assessed in terms of the correlation between a new instrument and some criterion using large survey samples, it is also sometimes applied in relation to performance or status (Kazdin, 1998). Therefore, we utilized supervisor performance data to assess the criterion validity of SAGE, comparing CBT supervision against EBCS within a longitudinal, $N = 1$ multiple phase (A-B-A-B) design. The CBT condition was 'supervision as usual', whereas the EBCS condition placed a greater emphasis on providing an optimally challenging environment (see items 3 and 4 in Table 1), intended to facilitate experiential learning within a developmentally informed approach (e.g. greater attention to the supervisee's affective experience and to enactive learning than in CBT supervision; see Milne 2009). This represents an experimental evaluation of the relative effect of the two approaches

https://doi.org/10.1017/S1754470X11000079 Published online by Cambridge University Press

on SAGE scores. If SAGE contains valid items, it should detect differences between these two methods of supervision, assuming fidelity. We attempted to maximize fidelity by taping all sessions, which were listened to by a consultant (i.e. 'supervision-of-supervision') leading to fortnightly feedback and guidance to the supervisor. Amongst other things, this was intended to ensure that the supervisor adhered to the relevant approach.

An $N = 1$ design is recommended in studies of this kind, due to its high internal validity (Oliver & Fleming, 2002). The phases within the design were alternating baseline phases (phase A: supervision as usual, i.e. CBT supervision), and intervention phases (phase B: EBCS). Six participants took part in the criterion validity assessment. The consultant (i.e. the supervisor's supervisor) was a male chartered clinical psychologist aged 58 years and based at a UK University. Specializing in staff development and a Fellow of the British Psychological Society, he had 31 years of relevant experience and was considered proficient in both providing and evaluating clinical supervision (e.g. he had published extensively in the supervision literature, including four prior $N = 1$ studies). The supervisor was a male licensed psychologist aged 56 years working as the director of a training clinic (mental health) in North America. There was also one supervisee, a 35-year-old female, who was a post-doctoral student at the supervisor's clinic. There were three client participants, male and female adults presenting with complex mental health problems to this training clinic, all were clients of the supervisee.

The study came about following a workshop led by the first author that was attended by the supervisor, who then invited the first author to collaborate on the present study. Therefore, the selection of the consultant, supervisor, supervisee and client was based on convenience sampling, namely their availability and willingness to be involved in a study of supervision, as part of an existing research programme. Ethical clearance was given by the relevant NHS Research and Development Department (first author), the Human Subjects Internal Review Board at Palo Alto University (second author), and by a UK university's Ethics Committee (Psychology Department: third and fourth authors).

The procedure was for the consultant to hold hour-long telephone consultancy sessions with the supervisor throughout the study (i.e. during all phases), typically after every second supervision session. These consultancy sessions consisted of the consultant giving a detailed review of the previous supervision session, which he had already listened to on audio tape, and then subsequently rated using the SAGE instrument. The SAGE results were forwarded as feedback to the supervisor, prior to consultancy. Using both the SAGE manual as a guide and his skills as a supervisor, the consultant attempted to facilitate the use of EBCS and CBT supervision by the supervisor. For a more detailed description of the consultancy procedure used, see Milne & Westerman (2001, p. 447).

We obtained trends that consistently indicated that EBCS resulted in greater learning by the supervisee (i.e. items 19–23; see Table 1). Expressed in terms of the observed differences in the SAGE learning scores between phases, EBCS was associated with enhanced learning on four of these five SAGE learning items (mean 32% improvement). Statistical analysis of covariation across the overall A-B-A-B sequence supported this descriptive summary and subjective interpretation: the supervisee's learning significantly covaried with EBCS ($Z = 2.45$, $p = 0.01$). By contrast, non-significant findings were obtained for all CBT sequences. The conventional $N = 1$ longitudinal (i.e. session-by-session) data are presented within a separate paper describing a comparative evaluation of the CBT and EBCS approaches (Milne *et al.*, unpublished data).

https://doi.org/10.1017/S1754470X11000079 Published online by Cambridge University Press

Discriminant validity

Naturalistic recordings of three supervision sessions provided by three volunteer supervisors and their current supervisees were selected for this analysis, as they were believed to represent three distinct approaches, i.e. CBT, psychodynamic, and systemic supervision. The supervisors were all qualified mental health practitioners working within the same NHS Trust. Two were male and one female (the psychodynamic supervisor). All were in the age range 30–50 years.

The recordings were produced to a professional standard by a university TV studio, and were part of an EBCS manual for training novice supervisors (Milne, 2010). These three sessions were considered to be distinct, because of the espoused theoretical orientations of the three supervisors, an assumption which was supported by the assessment of the first author and by the comments of 20 experienced supervisors viewing these tapes (as a learning exercise within a supervision workshop). Each supervision session lasted between 45 min and 1 hour. The systemic and psychodynamic tapes were purposively sampled, as there was only one clear-cut example of each approach in our sample (most volunteers drew on a CBT approach). The CBT tape was selected as the supervisor concerned was an approved supervisor within the local CBT Diploma training course. However, the one rater involved was blind to these three approaches. He had established good inter-rater reliability in a prior study.

We found that each of the three supervisory approaches contained theoretically congruent behaviours, ones that equated in a suitably differentiated way to the SAGE items. We made the comparisons by calculating the mean percent frequency of the observed SAGE items for each tape (i.e. we profiled the three forms of supervision). This was because a comparison based on the supervisors' competence (the conventional way to rate SAGE items) would not have revealed the relative profiles of these three supervision approaches, and therefore would not have allowed us to examine the capacity of SAGE to discriminate these approaches. Specifically, the CBT supervision session contained behaviours that corresponded to all but one of the SAGE items (i.e. item 18, 'experiencing', was not observed). SAGE ratings indicated that the CBT session contained a considerable amount of supervisor teaching (12.8% of total observed behaviours) and supervisee reflecting (11.6%). This was followed in frequency by questioning the supervisee (9.9%), and by the supervisor observing (9.5%) and listening (7.3%) to the supervisee. By contrast, the psychodynamic supervision session contained mostly supervisee reflecting (32.3%), perhaps related to the supervisor's listening (18.6%), questioning (9.1%), and teaching (10.5%). Notably, only a small number of SAGE items could be detected in the content of the psychodynamic session. For example, this tape did not include any agenda-setting, demonstrating, receiving feedback, observing, training or experimenting. By contrast, within the systemic supervision session only one SAGE item was not observed: 'demonstrating'. Similar to the psychodynamic session, the systemic session was predominantly a combination of reflecting (32.7%), listening (24.9%), questioning (11.7%) and discussing (5%). This study is described more fully in a forthcoming paper by Cliffe & Milne.

Overall, the sample of CBT supervision session corresponded most closely to SAGE, as it was observed more frequently across the observed 21 items than the other two approaches. By contrast, the final SAGE category of 'other' supervisory/supervisee behaviours was never observed during any of the three recordings. This suggests that SAGE may be sufficiently broad to be able to assess these different types of clinical supervision,

https://doi.org/10.1017/S1754470X11000079 Published online by Cambridge University Press

but that it is most attuned to CBT supervision, corroborating the content validity of SAGE.

Reliability

In terms of inter-rater reliability, the third and fourth authors were the raters who completed SAGE within the present study. They were a male and a female, aged 21 years, and both were undergraduate students. They were initially trained in using the SAGE manual by discussing and rating an audio tape with the first author, over a 90-min period. They then coded a succession of supervision tapes jointly, stopping the tapes where necessary to discuss and clarify individual ratings. These ratings used the Dreyfus competence scale, as described above (as opposed to the frequency calculation used within the discriminant validity section). Following this training phase (approximately 6 hours), a randomly selected baseline audio tape was used for an inter-rater reliability assessment, based on independent observation. The independent ratings of this session were compared and exact percent agreement was calculated, together with Pearson's r analysis and a more robust reliability measure, Cohen's kappa coefficient (Landis & Koch, 1977). The exact percent agreement between the observers' ratings was calculated using the formula:

$$\frac{\text{number of agreements} - \text{number of disagreements}}{\text{total number of ratings (observations)}} \times 100.$$

At the end of the observer training period, exact inter-rater percent agreement was 73%, Pearson's correlation was $r = 0.815$ ($p = 0.001$), and the kappa coefficient was $\kappa = 0.54$, a moderate strength of agreement (Landis & Koch, 1977). In summary, inter-rater reliability was found to be acceptable (Lombard *et al.* 2006). As noted under the construct validity analysis above, the internal consistency of the self-rated version of SAGE was 0.98.

Discussion

There are few instruments available with which to measure competence in clinical supervision (Ellis & Ladany, 1997), and the lack of valid and reliable supervision instruments is a key impediment to research on supervision (Watkins, 1998). In order to begin to address this deficiency, we have outlined a new observational measure, SAGE, for which we presented some generally promising but initial psychometric data (e.g. acceptable inter-rater reliability: $r = 0.815$, $p = 0.001$, $\kappa = 0.54$). SAGE also appears to have face validity, based on the ratings of supervision experts. Less satisfactory, in relation to our model of supervision, was the construct validation of SAGE, which indicated that there was just one factor, termed 'supervisory competence'. This suggests that the hypothesized three-factor structure is invalid (see Table 1), at least for the sample and task assessed, although there is at least one strong factor. Predictive validity was assessed in terms of SAGE's ability to detect changes in supervisory competence over time, and we found evidence to support the anticipated relative effectiveness of the EBCS approach in terms of the supervisee's learning (i.e. SAGE values were 32% higher during EBCS supervision, a significant improvement over CBT supervision). SAGE also appeared to be able to distinguish between three theoretically distinct approaches:

https://doi.org/10.1017/S1754470X11000079 Published online by Cambridge University Press

CBT, systemic, and psychodynamic supervision. This was treated as a discriminant validity check, and indicated that SAGE was best attuned to CBT supervision (e.g. all SAGE items were observed within the sample session, and the ratings corresponded to the CBT approach as described in the literature).

In addition to these psychometric or 'design' criteria, we should also comment on the remaining dimensions of the DIY approach. In terms of 'implementation', SAGE is available on request to the first author, free of charge, and (based on the present findings) can potentially be used reliably by observers with minimal training, being based on MTS, PETS tends to require approximately 3 hours to code one 60-min session, whereas SAGE can be completed in about 5 min after viewing the session. Another part of the rationale for developing SAGE was that it has good face validity and can also be completed in relatively brief periods of time, yielding highly descriptive data. As regards 'yield', SAGE can be used to profile a supervisor's competence, highlighting strengths and weaknesses. The obtained profile can provide a rough indication of whether or not the supervisor is implementing a particular approach, as per the above comparison between CBT, systemic, and psychodynamic supervision. The inclusion of an established competence scale helps to benchmark the proficiency level of the observed supervision. In particular, the final items within SAGE allow assessment of the effectiveness of the sampled supervision, in terms of whether the supervisee is being encouraged to reflect, conceptualize, plan, etc. (mini-impacts). In summary, judged in terms of the DIY criteria, SAGE has shown some promise as an instrument that can help to address the measurement problems within the supervision field.

Limitations of the study and recommendations for future research

It could be argued that our attention to validity is disproportionate, 'validity . . . is not a large concern' 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 (Barlow *et al.* 2009, p. 130), unlike traditional self-report instruments that treat the data as a sign of something unobservable and hence inferred to be underlying the items (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). There is also scope to develop better inter-rater reliability for SAGE. It may be that a more sophisticated training procedure would have achieved better results, or it may be that it requires experienced supervisors to achieve really high levels of reliability.

In general, and consistent with the above point, the present study presents preliminary data, and there is a clear case for both deepening and broadening this psychometric analysis. For example, at the time of writing we are undertaking a generalizability analysis of SAGE, to include a re-evaluation of its factor structure. This should clarify whether our current one-factor solution is valid, and also contribute to item refinement.

A third issue concerns the distinctiveness of our two experimental conditions, CBT supervision and EBCS, in that it could be argued that EBCS is simply CBT supervision done thoroughly. Furthermore, in both cases the supervision was of CBT-based clinical work, which could further confound the intended comparison (i.e. at a clinical level, both

https://doi.org/10.1017/S1754470X11000079 Published online by Cambridge University Press

are indeed supervision of CBT). Although narrative accounts of CBT supervision suggest a strong similarity to EBCS (e.g. Padesky, 1996; Liese & Beck, 1997), there do appear to be conceptually distinct aspects of EBCS, due to it drawing on ideas about human development and learning from beyond the CBT supervision literature, alongside a different emphasis on some shared variables (e.g. EBCS places greater stress on the behavioural and affective aspects of supervision; Milne, 2008). This perspective is supported in the present study, as we obtained statistically different findings for the two approaches (i.e. we met the 'non-equivalence' criterion that is used within NICE). Therefore, we believe that, while EBCS overlaps significantly with CBT supervision, it is both conceptually and operationally distinct (just as the defining nature of motivational interviewing has been distinguished from the concept of CBT; Miller & Rollnick, 2009; and CBT distinguished from other therapies; Ablon & Jones, 2002).

One weakness that we do recognize as inherent in SAGE is that it only measures one aspect of supervision (i.e. the 'formative' function of facilitating learning; Proctor, 1988), and in a rather general way. Theoretically, there are many valid criteria by which to evaluate supervision. For example, in their review, Wheeler & Richards (2007) noted that researchers had measured a number of 'mini-impacts' of supervision (including supervisee self-awareness, skills development and self-efficacy). This echoes the long-standing argument that various interactional (e.g. the supervisory alliance) and mini-outcome variables merit attention (e.g. case re-conceptualization; Lambert, 1980; Holloway, 1984), for methodological and conceptual reasons.

Finally, in relation to the self-report versions of SAGE (i.e. supervisor and/or supervisee completed) versus the observer-rated version, we note that such multiple methods of measurement tend to have a low level of concordance. In one respect this is precisely the justification for their complementary use; to gather contrasting information, so as to better 'triangulate' the phenomenon under study. To illustrate, when asked to rate the same sessions, the correlation between supervisor and supervisee evaluations have been found to be non-significant (Efstation *et al.* 1990; Zarbock *et al.* 2009). This replicates the data regarding therapist and patient ratings of therapy (e.g. Hill & Lambert, 2004).

Conclusions

In their review of the effectiveness of supervision, Lambert & Ogles (1997) noted that 'advances in knowledge can be expected to increase with advances in criterion measurement' (p. 441). We have introduced a new instrument for advancing the measurement of supervision, i.e. SAGE, applying the rounded DIY criteria to it. It has been noted that developing instruments to measure professional competennce 'remains a challenge' (Kaslow *et al.* 2009, p. S42) and this is certainly true of supervision. In the case of SAGE, although it is strong in relation to the 'implementation' and 'yield' criteria, our psychometric ('design') assessments of SAGE were preliminary in nature, so we plan to improve this analysis within ongoing research (e.g. a generalizability study). In the meantime, we view SAGE as a promising way to observe CBT supervision, one that should be supplemented by better-developed instruments that are used to obtain self-report data on complementary variables (e.g. assessing the supervisory alliance, as perceived by supervisors and supervisees).

https://doi.org/10.1017/S1754470X11000079 Published online by Cambridge University Press

Acknowledgements

We are grateful to Thomas Meyer and Dave Lee for their helpful comments on the psychometric analysis; and to Mark Freeston for conducting and interpreting the factor analysis, plus offering comments on the comparison between CBT supervision and EBCS.

Declaration of Interest

None.

Recommended follow-up reading

Falender CA, Cornish JAE, Goodyear R, Hatcher R, Kaslow NJ, Leventhal G, Shafranske E, Sigmon ST, Stoltenburg C, Grous C (2004). Defining competencies in psychology supervision: a consensus statement. *Journal of Clinical Psychology* 60, 771–785.

References

Ablon JS, Jones EE (2002). Validity of controlled clinical trials of psychotherapy: findings from the NIMH treatment of depression collaborative research program. *American Journal of Psychiatry* **159**, 775–793.

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

BABCP (2009). *Standards of Conduct, Performance and Ethics in the Practice of Behavioural and Cognitive Psychotherapy*. Bury: British Association for Behavioural and Cognitive Psychotherapy.

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 Behaviour 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 (2004). *Fundamentals of Clinical Supervision*, 3rd edn. London: Pearson.

Bordin ES (1983). Supervision in counselling: contemporary models of supervision: a working alliance-based model of supervision. *The Counselling Psychologist* **11**, 35–42.

BPS (2009). *Directory for Applied Psychology Practice Supervisors*. Leicester: British Psychological Society.

DoH (2009). *New Horizons: Towards a Shared Vision for Mental Health*. 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.

Efstation JF, Patton MJ, Kardish CM (1990). Measuring the working alliance in counsellor supervision. *Journal of Counselling Psychology* **37**, 322–329.

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

https://doi.org/10.1017/S1754470X11000079 Published online by Cambridge University Press

Epstein RM, Hundert EM (2006). Defining and assessing professional competence. *Journal of American Medical Association* **287**, 226–235.

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

Falender CA, Shafranske EP (2008). *Casebook for Clinical Supervision*. Washington DC: American Psychiatric Association.

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

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.

Gregorie TK, Propp J, Poertner J (1998). Supervisor's 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.

Hill CE, Lambert MJ (2004). Methodological issues in studying psychotherapy process and outcome. In: *Handbook of Psychotherapy and Behaviour Change* (ed. M. J. Lambert), pp. 84–135. New York: Wiley.

Holloway EL (1984). Outcome evaluation in supervision research. *The Counselling Psychologist*, **12**, 167–174.

HPC (2010). *Draft Standards for Doctoral Programmes in Clinical Psychology*. London: Health Professions Council.

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, Grus CL, Campbell LF, Fouad NA, Hatcher RL, Rodolfa ER (2009). Competency assessment toolkit for professional psychology. *Training and Education in Professional Psychology* **3**, S27–S45.

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

Kenkel MB, Peterson RL (2010). *Competency-Based Education for Professional Psychology*. Washington, DC: American Psychiatric Association.

Kolb DA (1984). *Experiential Learning: Experience as the Source of Learning and Development*. New Jersey: Prentice-Hall.

Lambert MJ (1980). Research and the supervisory process. In: *Psychotherapy Supervision: Theory, Research and Practice* (ed. A. K. Hess), pp. 423–450. New York: Wiley.

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.

Landis JR, Koch GG (1977). The measurement of observer agreement for categorical data. *Biometrics* **33**, 159–174.

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

Lombard M, Snyder-Duch J, Bracken CC (2006). Content analysis in mass communication: assessment and reporting of inter-coder reliability. *Human Communication Research* **2**, 587–604.

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.

https://doi.org/10.1017/S1754470X11000079 Published online by Cambridge University Press

Madson MB, Campbell TC (2006). Measure of fidelity in motivational enhancement: a systematic review. *Journal of Substance Abuse Treatment* **31**, 67–73.

Miller WR, Rollnick S (2009). Ten things that motivational interviewing is not. *Behavioural and Cognitive Psychotherapy* **37**, 129–140.

Milne DL (2007). Developing clinical supervision through reasoned analogies with therapy. *Clinical Psychology and Psychotherapy* **13**, 215–222.

Milne DL (2008). CBT supervision: from reflexivity to specialisation. *Behavioural and Cognitive Psychotherapy* **36**, 779–786.

Milne DL (2009). *Evidence-Based Clinical Supervision: Principles and Practice*. Chichester: BPS Blackwell.

Milne DL (2010). Can we enhance the training of clinical supervisors? A national pilot study of an evidence-based approach. *Clinical Psychology and Psychotherapy* **17**, 321–328.

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 (2011). Observing competence in CBT supervision: a systematic review of the available instruments. *The Cognitive Behaviour Therapist*. doi:10.1017/S1754470×11000067.

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 of Cognitive Therapy* **3**, 278–294.

Milne DL, Westerman C (2001). Evidence-based clinical supervision: rationale and illustration. *Clinical Psychology and Psychotherapy* **8**, 444–445.

Newman CF (2010). Competency in conducting CBT: foundational, functional, and supervisory aspects. *Psychotherapy Theory, Research, Practice, Training* **47**, 12–19.

Norcross JC (2001). Purposes, processes and products of the task force on empirically supported therapy relationships. *Psychotherapy: Theory/Research/Practice/Training* **38**, 345–356.

Norcross JC (2002) *Psychotherapy Relationships That Work: Therapist Contributions and Responsiveness to Patients*. Oxford: Oxford University Press.

Oliver JR, Fleming RK (2002). Applying within-subject methodology to transfer training research. *International Journal of Training and Development* **4**, 173–180.

Padesky CA (1996). Developing cognitive therapist competency: teaching and supervision models. In: *Frontiers of Cognitive Therapy* (ed. P. Salkovskis), pp. 266–292. New York: Guilford Press.

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 Behavior Analysis* **28**, 317–322.

Proctor B (1988). A cooperative exercise in accountability. In: *Enabling and ensuring* (ed. M. Marken and M. Payne), pp.21–34. Leicester: Leicester National Youth Bureau and Council for Education and Training in Youth and Community Work.

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.

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

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

Watkins CE (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.

https://doi.org/10.1017/S1754470X11000079 Published online by Cambridge University Press

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. *Counseling and Psychotherapy Research* **7**, 54–65.

Zarbock G, Drews M, Bodansky A, Dahme B (2009). The evaluation of supervision: construction of brief questionnaires for the supervisor and the supervisee. *Psychotherapy Research* **19**, 194–204.

Learning objectives

By studying this paper carefully, readers will be able to:

(1) Summarize the argument for developing a new instrument to measure supervisory competence.

(2) Discuss the criteria for a 'good' instrument.

(3) Evaluate the extent to which SAGE is a 'good' instrument.

https://doi.org/10.1017/S1754470X11000079 Published online by Cambridge University Press