


ORIGINAL RESEARCH

# Can we agree on the quality of clinical supervision? Inter-rater reliability of the *Short-SAGE* (Supervision: Adherence and Guidance Evaluation) scale

Maria Beckman<sup>1</sup> , Åsa Spännargård<sup>1</sup> and Sven Alfnsson<sup>1,2,\*</sup>

<sup>1</sup>Centre for Psychiatry Research, Department of Clinical Neuroscience, Karolinska Institutet & Stockholm Health Care Services, Stockholm, Sweden and <sup>2</sup>Department of Women's and Children's Health, Uppsala University, Uppsala, Sweden  
\*Corresponding author. Email: [sven.alfonsson@ki.se](mailto:sven.alfonsson@ki.se)

(Received 17 June 2020; revised 31 October 2020; accepted 2 November 2020)

## Abstract

Clinical supervision is a cornerstone in psychotherapist training, but research in this area is hampered by a lack of validated tools for assessing supervision quality. *Short-SAGE* (Supervision: Adherence and Guidance Evaluation) is an observational instrument designed for evaluating supervision in cognitive behavioural therapy. The aim of this study was to evaluate the inter-rater reliability of *Short-SAGE*. Four experienced clinical psychologists participated in three 3-hour *Short-SAGE* coding training sessions, followed by an additional meeting and coding instructions. In a cross-over design, codings of 20 supervision sessions were then assessed with intraclass correlations (ICC), for both the 3- and 7-point scales of the instrument. In the single measure analyses for both scales, only one item showed ICC in the good range, and the rest of the 14 item ICCs were in the poor to fair range. Moreover, on the 3-point scale, five of the 14 inter-rater correlations were non-significant. For research and training purposes, validated tools to assess supervision quality are highly needed. However, instruments for measuring adherence and/or competence are of little value if the coders do not attain inter-rater reliability. Whether quality of supervision is associated with improvements in supervisees' competencies is not yet clear. *Short-SAGE* provides a tool that may enable empirical research in this area. Further studies are needed to assess whether extensive training can improve the inter-rater reliability of *Short-SAGE*.

## Key learning aims

- (1) Readers will be aware of the urgent need for validated tools to assess clinical supervision quality.
- (2) Readers will be familiar with some existing tools for assessing the quality of clinical supervision.
- (3) Readers will be able to identify common problems in the development of instruments for assessing clinical supervision.

**Keywords:** clinical supervision; reliability; SAGE; *Short-SAGE*; supervision integrity

## Introduction

Clinical supervision is a cornerstone in psychotherapist training, but research in this area has been hampered by a lack of validated tools for assessing supervision quality (Watkins, 2012). This is partly due to the many different theoretical models for supervision, together with difficulties in operationalizing supervisor behaviours and supervision features (Watkins, 2011). The absence of assessment tools makes it difficult to draw firm conclusions about even fundamental supervision

components (Freitas, 2002; Milne *et al.*, 2010; Reiser and Milne, 2012), and to date, no supervision model has yet been empirically validated (Alfonsson *et al.*, 2018).

Without empirical guidance, psychotherapy supervision has mostly been modelled by psychotherapy practice. Liese and Beck's model (1997) for supervision in cognitive behaviour therapy (CBT), is one early example. Their model resembles cognitive therapy, and also includes elements used to promote behaviour change in CBT (e.g. goal setting, guided discovery, skills training). Other researchers have instead suggested that supervision may be a specific form of intervention by itself, and therefore should follow principles centred on the supervision context (Falender and Shafranske, 2012; Milne, 2008; Reiser and Milne, 2012). This is clearly expressed in the CORE Competence Framework (Roth and Pilling, 2007), which includes guidelines for both clinical supervision *and* competencies of different psychotherapeutic schools. Supervision in this model is thereby seen as a unique entity that includes specific competencies; to some extent common across different forms of psychotherapy, but also therapy-specific that can be observed and measured. The Supervision Competency Scale (SCS) (Kennerley and Clohessy, 2010; Mueller, 2010) is a tool for assessing CBT supervision built on the CORE Competence Framework. SCS includes 33 items in six areas (i.e. structuring of sessions, enhancing learning, supervisory relationship, other process issues, professional/ethical practice and reflective practice). The instrument is rather extensive, and the validity and reliability are not yet fully known.

Another tool for assessing CBT supervision is SAGE (Supervision: Adherence and Guidance Evaluation) (Milne *et al.*, 2011). Originally developed to assess the competencies of evidence-based supervision (EBCS) (Milne, 2009; Milne and Dunkerley, 2010), SAGE mirrors existing observational coding instruments for evaluation of psychotherapy, such as the Cognitive Therapy Scale (CTS-R) (Blackburn *et al.*, 2001). The original 23-item version of SAGE was empirically derived to assess the competence of supervisors based upon direct observation of practice samples. The 23 items included different aspects of relationship variables, supervisor's competencies, and supervisee's engagement in experiential learning (Milne *et al.*, 2011). As clinical supervision is difficult to reliably assess, the authors suggest a minimum of three to four supervision sessions in order to receive a reasonable assessment of competence. Recently, SAGE has been refined into a shorter 14-item version (*Short-SAGE*), which has undergone some preliminary psychometric evaluation supporting its usability (Reiser *et al.*, 2018). However, rating therapist or supervisor competence is challenging, and poor inter-rater reliability is a common problem in observational coding instruments. Additionally, measures of reliability of ratings often vary widely both across and within studies (Roth *et al.*, 2019), and many studies use incorrect statistical procedures, or fail to fully report information necessary to interpret the inter-rater reliability results (Hallgren, 2012). *Short-SAGE* has not yet been evaluated in this regard, and the aim of this study was therefore to assess the inter-rater reliability of *Short-SAGE*.

## Method

### Design

Data in this study were retrieved from an ongoing study of basic and advanced level psychotherapy students' courses in CBT. In the study, via audio recordings of therapy sessions, supervisors monitored and assessed students' clinical performance when they provided psychotherapy in diverse clinical settings. The supervision sessions included performance feedback, and were conducted in groups with three to four supervisees (i.e. each supervisee received approximately 45 minutes of supervision per session). All sessions were audio recorded. The supervisors were experienced psychotherapists (i.e. 14–32 years of professional psychotherapy experience, and 4–24 years of experience as CBT supervisors), with advanced training in CBT and additional training in CBT

supervision. They were not trained or instructed to follow the specific EBCS model, but to comply with the broader CBT supervision framework (e.g. review and provide feedback on students' therapy sessions, foster reflective thinking using Socratic methods, encourage students' responsibility for further learning, and, when necessary, teach/model therapeutic skills). All supervisors were independently contracted to the training centre and not part of the permanent staff. The research team approached all supervisors currently providing supervision at the training centre, and all chose to participate in the study.

This study used audio recordings from 48 supervision sessions, provided over 2 months by 12 different supervisors, collected from both basic and advanced CBT training courses. From these, a random sample of 20 recordings of approximately 50 minutes each were extracted for analysis. Prior to analysis, in order to delete identifiable patient information from the recordings, each recording was scrutinized by an independent staff member. The audio recordings were then randomly assigned to three of four coders in a cross-over design, and then independently coded with *Short-SAGE*. All four coders were clinical psychologists with advanced training in CBT. They were working as teachers in clinical and/or supervision courses, and had practical experience of supervision in diverse settings. The coders also had experience in assessing CBT and clinical supervision, and were familiar with the EBCS model of supervision. Additionally, they all had various previous training in other coding procedures, such as the Cognitive Therapy Scale, Teacher's PETS and The Motivational Interviewing Treatment Integrity Code. Prior to coding, the coders received the *Short-SAGE* manual and were asked to familiarize themselves with the coding instructions. They then participated in three 3-hour *Short-SAGE* workshops, in which three randomly selected supervision sessions were analysed and discussed in order to promote a common understanding of the instrument and to reach scoring consensus. The training outline was based on the *Short-SAGE* manual, and the description of Loades and Armstrong (2016). In each workshop, the coders listened to a recorded supervision session and then discussed the ratings of each item until the rationale was clarified and consensus was reached. The coders had an additional meeting regarding two items they perceived as most difficult to interpret (i.e. Prompting and Conceptualizing). To further the coders' understanding of these items, contact was made with the founders of SAGE, who provided more elaborate instructions. No coder in the group served as the gold standard or master coder. The goal was instead to reach consensus between all coders, for all scores, in each session. After the four workshops with joint coding, based on the level of agreement at the fourth workshop, the inter-rater agreement was deemed high enough for the independent coding. For the subsequent codings, each coder was provided with the *Short-SAGE* manual. All recordings were coded independently and submitted anonymously. The coding sheets were then compiled by an independent staff member, and provided with a code number.

### **Instrument**

*Short-SAGE* contains 14 items (i.e. Managing, Agenda-setting, Formulating, Questioning, Prompting, Demonstrating, Teaching, Training/Experimenting, Evaluating, Feedback, Reflecting, Conceptualizing, Planning, and Experiencing), each scored on a 7- or 3-point Likert scale (Reiser *et al.*, 2018). The factor structure is consistent with the underlying conceptual framework that guided the original 23-item SAGE version, including two of the major components: (1) *The Supervision Cycle* (i.e. specific supervisor behaviours which are believed to facilitate optimal experiential learning); and (2) *The Supervisee Cycle* (i.e. specific, observable supervisee learning competencies) (Reiser *et al.*, 2018). In the main, 7-point scale of *Short-SAGE*, 0 to 2 indicates incompetent/not demonstrated, 3 to 4 indicates competent, and 5 to 6 indicates expert competence. The 3-point scale, or the RAG (red-amber-green) scale, provides a coarser (i.e. incompetent/competent/expert) rating for each item, and is mainly used for training purposes.

**Table 1.** Mean *Short-SAGE* scores for each item for each coder

| Item                      | Coder A<br>Mean (SD) | Coder B<br>Mean (SD) | Coder C<br>Mean (SD) | Coder D<br>Mean (SD) |
|---------------------------|----------------------|----------------------|----------------------|----------------------|
| 1. Managing               | 3.1 (1.2)            | 3.0 (0.9)            | 3.2 (1.0)            | 3.2 (1.0)            |
| 2. Agenda-setting         | 1.4 (1.4)            | 1.7 (1.3)            | 2.2 (0.9)            | 2.0 (1.5)            |
| 3. Formulating            | 3.6 (0.7)            | 3.1 (0.7)            | 3.0 (0.5)            | 3.2 (0.9)            |
| 4. Questioning            | 3.2 (1.0)            | 2.7 (0.6)            | 3.3 (0.9)            | 3.5 (1.3)            |
| 5. Prompting              | 2.8 (0.7)            | 2.5 (0.6)            | 2.9 (0.9)            | 2.9 (1.2)            |
| 6. Demonstrating          | 1.2 (1.7)            | 1.5 (1.4)            | 1.7 (1.6)            | 1.4 (1.5)            |
| 7. Teaching               | 3.8 (0.8)            | 3.8 (0.8)            | 4.0 (0.8)            | 3.9 (0.6)            |
| 8. Training/experimenting | 0.8 (1.6)            | 0.8 (1.4)            | 1.5 (1.4)            | 0.7 (0.9)            |
| 9. Evaluating             | 3.3 (1.0)            | 3.5 (0.6)            | 3.6 (0.7)            | 3.4 (0.7)            |
| 10. Feedback              | 1.2 (1.4)            | 0.8 (0.8)            | 1.6 (1.0)            | 1.0 (0.7)            |
| 11. Reflecting            | 4.2 (0.5)            | 3.6 (0.6)            | 3.2 (1.1)            | 3.5 (0.7)            |
| 12. Conceptualizing       | 1.5 (1.5)            | 1.5 (1.3)            | 1.6 (1.1)            | 1.0 (0.5)            |
| 13. Planning              | 3.8 (0.6)            | 3.0 (0.5)            | 3.4 (1.1)            | 3.8 (0.8)            |
| 14. Experiencing          | 2.6 (1.4)            | 2.3 (1.3)            | 2.4 (1.5)            | 2.6 (1.7)            |

### Data analyses

The inter-rater reliability was assessed with a two-way, random effects, absolute agreement, intraclass correlation (ICC) (Hallgren, 2012; Koo and Li, 2016; McGraw and Wong, 1996). Results for both single (i.e. the reliability of the ratings based on ratings provided by a single coder) and average (i.e. the reliability of the ratings, based on the mean value of ratings provided by several coders) measures are presented, for both the 7- and 3-point scale of *Short-SAGE*. Following published guidelines, the chosen design should result in a 90% ability to detect modest (0.4) correlations between raters' scores (Bujang and Baharum, 2017; Walter *et al.*, 1998). The ICCs were interpreted according to the recommendations of Cicchetti and Sparrow (1981): <0.39 as poor; 0.40–0.59 as fair; 0.60–0.74 as good; and 0.75–1.00 as excellent.

### Results

Table 1 shows the mean *Short-SAGE* 7-point scale scores for each item, for each coder. In all coded supervision sessions, for the *Short-SAGE* 7-point scale, no item reached the instrument's highest value (i.e. 6), so the range was somewhat restricted. Additionally, four of the 14 items showed floor effects (i.e. Formulating, Teaching, Evaluating and Reflecting) with no scores lower than 2 (Table 2). This indicates that all supervision sessions included elements of supervisors actively encouraging the supervisees to analyse/synthesise and generate clinical presentations, didactic information from the supervisors, monitoring activities from the supervisors, and that the supervisees in all 20 sessions, in the light of their own understanding, summarised relevant events from their therapy sessions. Moreover, four of the 14 items showed ceiling effects: three of them (i.e. Prompting, Demonstrating, and Training/Experimenting) with no scores higher than 4, and one of them (Feedback) with no scores higher than 3 (Table 2). This indicates that none of the supervisors, in an expert way, prompted/cued the supervisee about relevant material, modelled/illustrated skills or engaged the supervisees in experiential learning, and that none of them, in a proficient way, let the supervisees summarise the supervision session.

All inter-rater correlations for the 7-point scale were statistically significant. For the single measures, six of the 14 items were in the poor range (i.e. <.40), seven in the fair range (i.e. 0.40–0.59), and one was in the good range (i.e. 0.60–0.74). For the average measures, three of the items were in the fair range, six in the good range, and five were in the excellent range (i.e. 0.75–1.00) (Table 2).

For the 3-point scale, five of the 14 inter-rater correlations turned out non-significant (i.e. Prompting, Teaching, Evaluating, Feedback and Reflecting). For the remaining nine items, the single measure analyses resulted in six items in the poor range (i.e. <.40), two in the fair

**Table 2.** Range and intra-class correlation coefficients (ICC) for the *Short-SAGE* 7-point scale

| Item                      | Range | ICC <sup>a</sup><br>(95% CI) | ICC <sup>b</sup><br>(95% CI) | <i>p</i> |
|---------------------------|-------|------------------------------|------------------------------|----------|
| 1. Managing               | 1–5   | .40 (.13 to .67)             | .67 (.30 to .86)             | .002     |
| 2. Agenda-setting         | 0–5   | .53 (.22 to .76)             | .77 (.46 to .91)             | .001     |
| 3. Formulating            | 2–5   | .29 (.04 to .57)             | .55 (.12 to .80)             | .008     |
| 4. Questioning            | 0–5   | .35 (.09 to .62)             | .62 (.24 to .83)             | .002     |
| 5. Prompting              | 0–4   | .43 (.16 to .68)             | .69 (.37 to .87)             | .001     |
| 6. Demonstrating          | 0–4   | .39 (.13 to .66)             | .66 (.30 to .85)             | .001     |
| 7. Teaching               | 2–5   | .52 (.26 to .75)             | .77 (.51 to .90)             | .001     |
| 8. Training/experimenting | 0–4   | .54 (.28 to .76)             | .78 (.54 to .91)             | .001     |
| 9. Evaluating             | 2–5   | .35 (.08 to .63)             | .62 (.20 to .84)             | .006     |
| 10. Feedback              | 0–3   | .49 (.23 to .73)             | .74 (.47 to .89)             | .001     |
| 11. Reflecting            | 2–5   | .18 (–.02 to .46)            | .40 (–.10 to .72)            | .037     |
| 12. Conceptualizing       | 1–5   | .53 (.26 to .76)             | .77 (.51 to .90)             | .001     |
| 13. Planning              | 1–5   | .23 (.01 to .51)             | .47 (.01 to .76)             | .015     |
| 14. Experiencing          | 0–5   | .71 (.51 to .86)             | .88 (.75 to .95)             | .001     |

<sup>a</sup>single measure.<sup>b</sup>average measure.**Table 3.** Range and intra-class correlation coefficients (ICC) for the *Short-SAGE* 3-point scale

| Item                      | Range | ICC <sup>a</sup><br>(95% CI)    | ICC <sup>b</sup><br>(95% CI)     | <i>p</i> |
|---------------------------|-------|---------------------------------|----------------------------------|----------|
| 1. Managing               | 0–2   | .32 (.04 to .61)                | .58 (.10 to .82)                 | .013     |
| 2. Agenda-setting         | 0–2   | .39 (.13 to .66)                | .66 (.31 to .85)                 | .001     |
| 3. Formulating            | 0–2   | .25 (.01 to .54)                | .50 (.01 to .78)                 | .028     |
| 4. Questioning            | 0–2   | .48 (.22 to .72)                | .74 (.46 to .87)                 | .001     |
| 5. Prompting              | 0–1   | .19 (–.04 to .44) <sup>c</sup>  | .42 (–.15 to .74) <sup>c</sup>   | .062     |
| 6. Demonstrating          | 0–2   | .45 (.18 to .70)                | .71 (.40 to .88)                 | .001     |
| 7. Teaching               | 0–2   | .20 (–.07 to .51)               | .42 (–.23 to .75)                | .077     |
| 8. Training/experimenting | 0–2   | .68 (.45 to .84)                | .86 (.71 to .94)                 | .001     |
| 9. Evaluating             | 0–2   | .05 (–.19 to .37)               | .14 (–.88 to .64)                | .344     |
| 10. Feedback              | 0–1   | –.04 (–.23 to .35) <sup>c</sup> | –.13 (–1.29 to .50) <sup>c</sup> | .611     |
| 11. Reflecting            | 0–2   | .22 (–.05 to .53)               | .46 (–.15 to .77)                | .055     |
| 12. Conceptualizing       | 0–2   | .37 (.09 to .64)                | .63 (.22 to .84)                 | .005     |
| 13. Planning              | 0–2   | .28 (.01 to .58)                | .54 (.01 to .80)                 | .025     |
| 14. Experiencing          | 0–2   | .28 (.04 to .56)                | .54 (.10 to .79)                 | .005     |

<sup>a</sup>single measure.<sup>b</sup>average measure.<sup>c</sup>range too restricted for adequate analysis.

range (i.e. 0.40–0.59), and one in the good range (i.e. 0.60–0.74). The average measure analyses for the remaining nine items resulted in four items in the fair range, four in the good, and one item in the excellent range (i.e. 0.75–1.00) (Table 3).

For both scales, there were rather large discrepancies between the ICC for single and average measures, indicating low levels of percentage agreement across items (Table 2 and 3). In a *post-hoc* analysis, all coders were compared pairwise in order to detect any outlier with consistently lower inter-rater reliability, but no single coder stood out in this regard.

## Discussion

The aim of this study was to assess the inter-rater reliability of *Short-SAGE*. For the 7-point scale, the range was somewhat restricted, and the analyses revealed floor effects for four of the 14 items, and ceiling effects for another four. For both scales, for the single ICC measures, only one item was in the good range, and the rest of the items were in the fair to poor range. Moreover, on the 3-point

scale, five of the 14 inter-rater correlations turned out to be non-significant. This 3-point RAG scale (i.e. red-amber-green) has been proposed by the authors as an educational tool; it is useful as a basis for supervisor-supervisee discussions. Poor inter-rater reliability may be less problematic in non-evaluative contexts. However, results of the current study indicate that the RAG scale is not reliable for evaluating purposes where the exact level (e.g. 'fail' and 'pass') is important, at least not without extensive coder training. Moreover, the average measurements showed, not surprisingly, better results for both scales. However, as the *Short-SAGE* assessment is normally done by a single rater, the single-measures results are most relevant for the assessment of the scale's inter-rater reliability.

Unfortunately, this study's results did not provide additional information regarding whether larger samples of sessions and/or more extensive coder training would have generated higher ICC scores. As Syed and Nelson (2015) state in their article on Guidelines for Establishing Reliability when Coding Narrative Data: reliability is not a product, but a process that involves multiple time-intensive steps. However, as the literature rarely describes the process of training of coders who measure treatment fidelity (Kramer Schmidt *et al.*, 2019), it is difficult to know exactly how that training should be conducted. The *Short-SAGE* manual proposes a 1-day training workshops for raters, guided by the full SAGE manual. Other researchers have instead proposed a considerable amount of training, conducted in a stepped training approach, with a level of inter-rater reliability specified *a priori* (Hallgren, 2012; Syed and Nelson, 2015).

Interestingly, for some of the supervision sessions, many of the items scored 0 (Table 2), indicating absence of features, or highly inappropriate performance in that specific domain. This is especially surprising when it comes to Agenda-setting, Questioning, Demonstrating, and Feedback; supervisor behaviours that could arguably be expected in most, if not all, CBT supervision sessions. However, while CBT supervision text books quite unanimously promote supervisor behaviours, such as the use of agenda, Socratic questioning and modelling (e.g. Watkins and Milne, 2014), our experience tells us that CBT supervision content varies to a large degree. Clinical supervision has not been monitored or scrutinized as closely as some psychotherapy methods, and the content has not been studied more objectively until recently (Alfonsson *et al.*, 2018). Hopefully, clinical supervision can develop in a similar way as psychotherapy, including more transparency and a closer adherence to published guidelines. That being said, few of the supervision techniques, including those described in *Short-SAGE*, have been experimentally explored, and to a large extent, we still do not know exactly which supervision behaviours are effective in the training of psychotherapists. In other words, even if it is possible to improve the reliability of *Short-SAGE*, both the validity of the instrument, and the underlying model that *Short-SAGE* is supposed to measure, are still unclear. Taken together, it is difficult to know if the low ICC levels in this study are related to the training of coders, the sample (i.e. both ceiling and floor effects) and/or the instrument itself.

### Limitations

This study has important limitations: Our sample of both supervision sessions (i.e., 20) and coders (i.e., 4) were small. However, neither the *Short-SAGE* manual, nor the article describing the instrument's psychometric properties (Reiser *et al.*, 2018) contain any information on the recommended number of sessions or coders for assessing the inter-rater reliability of *Short-SAGE*. To our knowledge, inter-rater reliability has not previously been investigated for a supervision coding instrument. However, instruments for assessing CBT competence, such as CTS-R, have been able to prove adequate inter-rater reliability (Blackburn *et al.*, 2001). In their study, Blackburn and colleagues had a total of 102 sessions coded by two out of four coders. The present study had a similar approach using fewer sessions (i.e., at least three coders coded each recorded session in a cross-over design). However, the restricted ranges, with both floor and ceiling effects, and the use of absolute agreement in this study's analyses, resulted in lower



statistical power than expected. The restricted ranges also limits the conclusions that can be drawn regarding a context that includes a fuller range of supervisor behaviours, which thus limits the generalizability of the results. A larger sample of sessions and coders may be needed to further analyse the inter-rater reliability of *Short-SAGE*, but coding of sessions consume large amounts of resources, and are often associated with practical difficulties. Moreover, in this study, the coders did not have any formal training in using *Short-SAGE*. Similar to the proposed *Short-SAGE* manual's one-day training workshops for raters, the coders in this study had participated in three inhouse three-hour *Short-SAGE* workshops, provided by two clinical psychologists with advanced training in CBT, and experience in using SAGE in their work as supervisors. In the workshops, three recorded supervision sessions were analysed. They also had an additional meeting regarding two difficult items, and received more elaborate instructions from the founders of SAGE. The results of the study indicate that a more extensive training proposed by researchers like Hallgren (2012) and Syed and Nelson (2015), which requires considerable efforts, may be essential for reaching an adequate inter-rater reliability. Taken together, the results of this study do not provide information regarding whether the low ICC levels were related to the training of coders, the coded sample (i.e. both ceiling and floor effects) and/or the instrument itself (i.e. the *Short-SAGE*).

### Conclusions

For both research and training purposes, validated tools for assessing supervision quality are highly needed. However, instruments for measuring adherence and/or competence are of little value if the coders do not attain inter-rater reliability. In this study, only one of the 14 items of *Short-SAGE* was in the good range, and the rest of the items were in the fair to poor range. Unfortunately, the results did not provide additional information regarding whether more extensive training and/or larger samples of sessions and coders would have generated a higher degree of correlation and agreement between items. Due to limited research in this area, it is unclear whether expert codings of global scores of supervisor behaviours is a valid method for measuring supervision quality. More behaviour-oriented approaches, like those developed for motivational interviewing (Moyers *et al.*, 2005), may be more accurate. Codings that provide more detailed information clearly provide richer data, and might also more easily generate agreement between items. *Short-SAGE* may be used to assess supervision quality in both research and training settings. Whether quality of supervision is associated with improvements in supervisees' competencies is as yet unclear. *Short-SAGE* provides a tool that may enable empirical research in this area. Further studies are needed to assess whether extensive training can improve the instrument's inter-rater reliability.

**Acknowledgements.** None.

**Financial support.** This research received no specific grant from any funding agency, commercial or not-for-profit sectors.

**Conflict of interest.** None.

**Ethics statements.** All participants were informed about the study procedure and provided informed consent. The procedure was approved by the Regional Ethics Committee Board (No. 2018/1735-31/3), and all authors have abided by the Ethical Principles of Psychologists and Code of Conduct as set out by the BABCP and BPS.

### Key practice points

- (1) Clinical supervision is a cornerstone in psychotherapist training, but there are few available tools to assess clinical supervision quality.
- (2) For both research and training purposes, tools for assessing supervision are highly needed. However, instruments for measuring adherence and/or competence are of little value if the coders do not attain inter-rater reliability.
- (3) Whether quality of supervision is associated with improvements in supervisees' competencies is as yet unclear.

## Further reading

- Alfonsson, S., Parling, T., Spännargård, Å., Andersson, G., & Lundgren, T. (2018). The effects of clinical supervision on supervisees and patients in cognitive behavioral therapy: a systematic review. *Cognitive Behaviour Therapy*, *47*, 206–228.
- Reiser, R. P., Cliffe, T., & Milne, D. L. (2018). An improved competence rating scale for CBT Supervision: Short-SAGE. *The Cognitive Behaviour Therapist*, *11*.

## References

- Alfonsson, S., Parling, T., Spännargård, Å., Andersson, G., & Lundgren, T. (2018). The effects of clinical supervision on supervisees and patients in cognitive behavioral therapy: a systematic review. *Cognitive Behaviour Therapy*, *47*, 206–228.
- Blackburn, I.-M., James, I. A., Milne, D. L., Baker, C., Standart, S., Garland, A., & Reichelt, F. K. (2001). The revised cognitive therapy scale (CTS-R): psychometric properties. *Behavioural and Cognitive Psychotherapy*, *29*, 431–446.
- Bujang, M. A., & Baharum, N. (2017). A simplified guide to determination of sample size requirements for estimating the value of intraclass correlation coefficient: a review. *Archives of Orofacial Science*, *12*.
- Cicchetti, D. V., & Sparrow, S. A. (1981). Developing criteria for establishing interrater reliability of specific items: applications to assessment of adaptive behavior. *American Journal of Mental Deficiency*, *86*, 127–137.
- Falender, C. A., & Shafranske, E. P. (2012). The importance of competency-based clinical supervision and training in the twenty-first century: why bother? *Journal of Contemporary Psychotherapy*, *42*, 129–137.
- Freitas, G. J. (2002). The impact of psychotherapy supervision on client outcome: a critical examination of 2 decades of research. *Psychotherapy: Theory, Research, Practice, Training*, *39*, 354.
- Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: an overview and tutorial. *Tutorials in Quantitative Methods for Psychology*, *8*, 23–34.
- Kennerley, H., & Clohessy, S. (2010). Becoming a supervisor. *Oxford Guide to Surviving as a CBT Therapist*, 323.
- Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, *15*, 155–163.
- Kramer Schmidt, L., Andersen, K., Nielsen, A. S., & Moyers, T. B. (2019). Lessons learned from measuring fidelity with the Motivational Interviewing Treatment Integrity code (MITI 4). *Journal of Substance Abuse Treatment*, *97*, 59–67.
- Liese, B. S., & Beck, J. S. (1997). Cognitive therapy supervision. In C. E. Watkins (ed), *Handbook of Psychotherapy Supervision*. Hoboken, NJ, USA: John Wiley & Sons Inc.
- Loades, M. E., & Armstrong, P. (2016). The challenge of training supervisors to use direct assessments of clinical competence in CBT consistently: a systematic review and exploratory training study. *The Cognitive Behaviour Therapist*, *9*.
- McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, *1*, 30.
- Milne, D. (2008). CBT supervision: from reflexivity to specialization. *Behavioural and Cognitive Psychotherapy*, *36*, 779. doi: [10.1017/s1352465808004773](https://doi.org/10.1017/s1352465808004773)
- Milne, D., & Dunkerley, C. (2010). Towards evidence-based clinical supervision: the development and evaluation of four CBT guidelines. *The Cognitive Behaviour Therapist*, *3*, 43–57. doi: [10.1017/s1754470x10000048](https://doi.org/10.1017/s1754470x10000048)
- Milne, D., 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. doi: [10.1521/ijct.2010.3.3.278](https://doi.org/10.1521/ijct.2010.3.3.278)
- Milne, D. L. (2009). *Evidence-Based Clinical Supervision: Principles and Practice*. John Wiley & Sons.
- Milne, D. L., Reiser, R. P., Cliffe, T., & Raine, R. (2011). SAGE: preliminary evaluation of an instrument for observing competence in CBT supervision. *The Cognitive Behaviour Therapist*, *4*, 123–138.
- Moyers, T. B., Martin, T., Manuel, J. K., Hendrickson, S. M., & Miller, W. R. (2005). Assessing competence in the use of motivational interviewing. *Journal of Substance Abuse Treatment*, *28*, 19–26.
- Mueller, M. (2010). *Oxford Guide to Surviving as a CBT Therapist*. Oxford University Press.
- Reiser, R. P., Cliffe, T., & Milne, D. L. (2018). An improved competence rating scale for CBT Supervision: Short-SAGE. *The Cognitive Behaviour Therapist*, *11*.
- Reiser, R. P., & Milne, D. (2012). Supervising cognitive-behavioral psychotherapy: pressing needs, impressing possibilities. *Journal of Contemporary Psychotherapy*, *42*, 161–171.
- Roth, A. D., Myles-Hooton, P., & Branson, A. (2019). Judging clinical competence using structured observation tools: a cautionary tale. *Behavioural and Cognitive Psychotherapy*, *47*, 736–744.
- Roth, A., & Pilling, S. (2007). A competence framework for the supervision of psychological therapies. Retrieved from: [https://www.researchgate.net/publication/265872800\\_A\\_competence\\_framework\\_for\\_the\\_supervision\\_of\\_psychological\\_therapies](https://www.researchgate.net/publication/265872800_A_competence_framework_for_the_supervision_of_psychological_therapies)
- Syed, M., & Nelson, S. C. (2015). Guidelines for establishing reliability when coding narrative data. *Emerging Adulthood*, *3*, 375–387.



- Walter, S., Eliasziw, M., & Donner, A. (1998). Sample size and optimal designs for reliability studies. *Statistics in Medicine*, 17, 101–110.
- Watkins, C. E. (2011). Psychotherapy supervision since 1909: some friendly observations about its first century. *Journal of Contemporary Psychotherapy*, 41, 57–67.
- Watkins, C. E. (2012). Psychotherapy supervision in the new millennium: Competency-based, evidence-based, particularized, and energized. *Journal of Contemporary Psychotherapy*, 42, 193–203.
- Watkins Jr, C. E., & Milne, D. L. (2014). *The Wiley International Handbook of Clinical Supervision*. John Wiley & Sons.

---

**Cite this article:** Beckman M, Spännargård Å, and Alfnsson S. Can we agree on the quality of clinical supervision? Inter-rater reliability of the *Short-SAGE* (Supervision: Adherence and Guidance Evaluation) scale. *The Cognitive Behaviour Therapist*. <https://doi.org/10.1017/S1754470X20000562>