1-1-2016

Consistency of supervisor and peer ratings of assessment interviews conducted by psychology trainees

Craig J. Gonsalvez
University of Wollongong, craigg@uow.edu.au

Frank P. Deane
University of Wollongong, fdeane@uow.edu.au

Peter Caputi
University of Wollongong, pcaputi@uow.edu.au

Consistency of supervisor and peer ratings of assessment interviews conducted by psychology trainees

Abstract

Observation of counsellor skills through a one-way mirror, video or audio recording followed by supervisors and peers feedback is common in counsellor training. The nature and extent of agreement between supervisor-peer dyads is unclear. Using a standard scale, supervisors and peers rated 32 interviews by psychology trainees observed through a one-way mirror. Results indicated that peers and supervisors used similar dimensions to cluster the various competencies. Peers rated counsellor performance more positively for general counselling skills but not for specialised techniques. Analyses revealed good supervisor-peer agreement for some items and poor agreement on others, with some differences being unacceptably large. The study has important implications for how feedback involving supervisors and peers might be managed and for peer supervision models.

Keywords

ratings, peer, supervisor, conducted, consistency, interviews, trainees, assessment, psychology

Disciplines

Education | Social and Behavioral Sciences

Publication Details

Gonsalvez, C. J., Deane, F. P. & Caputi, P. (2016). Consistency of supervisor and peer ratings of assessment interviews conducted by psychology trainees. British Journal of Guidance and Counselling, 44 (5), 516-529.

Consistency of supervisor and peer ratings of assessment interviews conducted by psychology trainees

Craig J. Gonsalvez, Frank P. Deane, & Peter Caputi

Craig Gonsalvez
Corresponding author
School of Social Sciences and Psychology &
Clinical and Health Psychology Research Initiatives,
University of Western Sydney, Australia
Tel: 61+2+47360185
Email: c.gonsalvez@uws.edu.au

Frank Deane
School of Psychology & Illawarra Institute for Mental Health,
University of Wollongong,
Northfields Avenue, NSW 2500, Australia
Tel: 61+2+42214523
Email: fdeane@uow.edu.au

Peter Caputi
School of Psychology
University of Wollongong,
Northfields Avenue, NSW 2500, Australia
Tel: 61+2+42213742
Email: pcaputi@uow.edu.au

Abstract

Observation of counsellor skills through a one-way mirror, video or audio recording followed by supervisors and peers feedback is common in counsellor training. The nature and extent of agreement between supervisor-peer dyads is unclear. Using a standard scale, supervisors and peers rated 32 interviews by psychology trainees observed through a one-way mirror. Results indicated that peers and supervisors used similar dimensions to cluster the various competencies. Peers rated counsellor performance more positively for general counselling skills but not for specialised techniques. Analyses revealed good supervisor-peer agreement for some items and poor agreement on others, with some differences being unacceptably large. The study has important implications for how feedback involving supervisors and peers might be managed and for peer supervision models.

Keywords: supervisor ratings, peer ratings, professional supervision, competency assessments, observational methods in supervision

Consistency of Supervisor and Peer Ratings of Assessment Interviews Conducted by Psychology Trainees

Clinical supervision conducted in supervisor-supervisee dyads has been the cornerstone of practitioner training in psychology for decades. Unlike other training components that can effectively be conducted in large or small groups, a significant proportion of clinical supervision is conducted in a one-to-one setting (Norcross, Hedges, & Castle, 2002; Milne, 2009). The dyadic delivery mode makes conventional supervision a resource intensive activity and an expensive component of professional training in psychology (Gonsalvez, Hyde, Lancaster, & Barrington, 2008) and other health disciplines (Spence, Wilson, Kavanagh, Strong, & Worrall, 2001). Several factors have underpinned and maintained such a model of practitioner training for close to a century (Gonsalvez & Milne, 2010). First of all, novice trainees lack both competence and confidence, and have to be supported through phases of misgivings and self-doubt as they deal with high levels of affect and difficult psychological problems (Stoltenberg, Bailey, Cruzan, Hart, & Ukuku, 2014). Secondly, the requirement for intensive supervision is mainly determined by the perceived importance of observation that may be immediate (e.g., through a one-way mirror or co-therapy) or delayed (e.g., through review of video or audio recordings). As a supervisory technique, observation is supported by expert consensus (see Reiser, 2014) and by research (e.g., Townend, Iannetta & Freeston, 2002). Conversely, an over reliance on subjective methods is not recommended, because self-report of case work may be unreliable, may miss important information, and may be vulnerable to bias particularly during early stages of counsellor development, when trainees are less capable of accurate self appraisal (Campbell, 1994; Gonsalvez & Calvert, 2014; Townend, Iannetta & Freeston, 2002). Finally, important knowledge-application (e.g.,

case conceptualization), skills (e.g., generic counseling and other specialized therapy skills) and relationship (e.g., self-awareness and transference reactions) competencies are difficult to assess accurately without recourse to data from some form of observation (Bennett-Levy et al., 2003; Gonsalvez, Oades, & Freestone, 2002; Kaslow et al., 2009). For instance, the way the counsellor communicates affect, the use of body language, and variation of tone, pace and timing of interventions are critical to credible evaluation of the counsellor's empathic skills. In a similar way, observation of behavior is also essential to determine whether and the extent to which the client is actively engaged in or resistant to the counsellor's interventions.

Potential Advantages of Peer Involvement in Supervision

Despite the benefits of close and intensive individual supervision, the expense, availability and accessibility of appropriate supervision is a sufficiently serious problem to prompt discussion of ways to maximize the benefit of such supervision. To maximize learning outcomes from observation, many clinical psychology programs have other trainees observe a supervisor, senior student or peer conducting assessment or therapy. As noted, this observation can occur using one-way mirror, video or audio recordings, with review and feedback occurring in individual or group supervision. Ideally, peer observers actively learn by attempting to understand the client-counsellor dynamics unfolding in the session, generating their own formulations, and discriminating between effective and ineffective intervention strategies. However, there is almost no research about what specific processes or learning occurs for these peer observers. It is unclear whether and to what extent peer appraisals of client experience and counsellor performance are consistent with those of the supervisor and self appraisals by the counsellor. A greater understanding of the domains and the level of agreement among peers, and between supervisor and

therapist/peer, can inform the development of specific strategies to maximize learning for trainees and to enhance the efficiencies and effectiveness of practitioner training and supervision. For example, high levels of anxiety and increased self-doubt during early stages of counsellor development are known to erode self esteem and may negatively bias self evaluations (Stoltenberg, et al., 2014). However, it is unclear whether and in which direction developmental stage affects peer evaluations. If a similar pattern of being overcritical demonstrated towards evaluation of peer performance, it might be advantageous to introduce peer-review after developmental anxieties are largely resolved. There is clearly a need for research to provide data on supervisor and peer trainees' views of observed assessment and therapy sessions.

Given the need to optimise learning opportunities when peers are observing fellow trainees, a series of questions appear relevant: How similar and in what domains are supervisor and peer evaluations comparable? Is there empirical evidence for the efficacy of peer observations? Are there specific competencies that are better accomplished by peer supervision? In general, "few peer or peer-group models have been implemented, and even fewer evaluated for their impact" (Crutchfield & Borders, 1997, p. 221). As notable exceptions, there have been some efforts to examine the outcomes of models by Benshoff and associates (1993, 1996) and by Borders (1991). The models included many traditional supervision activities including goal setting, tape review and case consultation. Evaluation data based on subjective evaluations from a fairly large number of trainees (n = 81) indicated excellent endorsement for peer supervision for each of the models, with participants reporting enhanced counselling and consultation skills, valuable support and valuable learning (Benshoff & Paisley, 1996). However, when a nine-week program of peer supervision was evaluated in a controlled study using objective measures of counselling

effectiveness, results demonstrated small effects in the expected positive direction that were not statistically significant (Cruthfield & Borders, 1997).

Given that observation is an essential component of conventional supervision practice, it is important for peer supervision models to demonstrate good agreement between peer and supervisor evaluation. However, there is a surprising lack of systematic scrutiny of inter-rater agreements and differences when rating counsellor performance and capabilities. This issue has gained renewed vigor following evidence that competency-based ratings (even by supervisors) are likely to be influenced by systematic rating biases (Gonsalvez & Freestone, 2007; Lazar & Mosek, 1993; Robiner, Saltzman, Hoberman, Semrud-Clikeman, & Schirvar, 1997).

Aims of Current Study

The current study aims to systematically examine the level of agreement among supervisors, peers and between supervisor and peer ratings of counsellor performance from behind a one-way mirror. Additionally, the study also explores whether supervisor and peer evaluations are affected by rating biases.

Method

Participants and Setting

The data for the study were 32, first-session, clinical assessment interviews conducted by clinical psychology trainees in an accredited training program in New Zealand. All assessments were conducted at the university psychology clinic which provided general clinical psychology services (assessment and psychotherapy) as well as specialist neuropsychology services. All assessments as part of this study involved referrals for general psychological services and involved adult clients (older than 18 years) usually presenting with mood or anxiety disorders. Most clients were referrals from general practitioners and other health care providers, although self-referrals were

also accepted. The Clinic Director was a senior academic and an experienced clinical psychologist employed by the School of Psychology at the University. The clinical psychologists had an ongoing clinical load through the clinic but also provided teaching support for the clinical psychology program. Cognitive-behaviour therapy (CBT) was the primary orientation of the training program and CBT is the emphasis of most clinical psychology training programs in Australian and New Zealand universities (Kazantzis & Munro, 2011). Of those universities offering specific training in CBT, 87% report assessing trainee competence through some form of observation (Kazantzis & Monro, 2011).

For the current article, the clinical trainee conducting the interview is called the counsellor, peer trainees observing the interview are called peers, and clinical faculty who rated the counsellor's performance are called supervisors. In the context of the present study, supervision practices were restricted to a training clinic and observation of an initial assessment interview. The role of the supervisor(s) was to provide to the counsellor feedback about a wide range of skills and therapy processes that occurred during the interview (e.g., the ability to establish and maintain rapport, microcounselling skills, flow and content of questioning), and to also offer suggestions about case conceptualization, further assessment, treatment planning. Following the counsellor's case formulation and feedback from the supervisor, peers were provided an opportunity to ask questions and to make observations or recommendations (e.g., regarding further assessment or treatment). The role of the supervisor in this context was to provide formative (versus summative) feedback to the counsellor and to manage the question and feedback process from peers. These screened interviews with peer and supervisor(s) as observers were a requirement of all trainees undertaking the clinical psychology training program. At the time these data

were collected interviews were not video or audio-recorded, but video recording is now standard procedure in this program.

Counsellors. Fourteen counsellors (3 males; 11 females) contributed 32 assessment interviews on real clients presenting to the University clinic for treatment, primarily for depression and anxiety. All counsellors were clinical psychology trainees who had completed their Masters degree in Psychology and were enrolled in the Post-Graduate Diploma in Clinical Psychology which was the main qualification for clinical psychology practice in New Zealand at the time of the research. Peers and clinical supervisors observed the interviews through a one-way mirror.

Peers. Ratings were obtained from 19 trainees (5 males; 14 females) who were in their final year of clinical training. Fourteen of these trainees also served as counsellors and conducted one or more of the assessed interviews. Ten interviews were rated by 2 peers, 16 interviews by 3 peers; 4 interviews by 4 peers, and 2 interviews by 5 peers.

Supervisors. Five members (2 males; 3 females) of the clinical faculty within the School of Psychology who were also qualified clinical psychologists and experienced in using the one-way mirror technique, served as supervisors for the study. Seventeen of the 32 interviews were observed and rated by 1 of the supervisors, 14 interviews by 2 supervisors and 1 interview by 3 supervisors. All supervisors were experienced clinical psychologists who had all participated in the screened interview process using one-way mirrors on multiple prior occasions. They also had prior experience conducting these observations in conjunction with other experienced supervisors present. All supervisors would have received some training, typically through workshop attendance. At the time of this research professional bodies had only just commenced formalising training and accreditation of supervisors. Thus,

formal training of the supervisors in supervision methods would have been variable although as noted all had participated with other supervisors in the screened interview process prior to this study commencing. In addition, supervisors periodically met to discuss the supervision processes involved in managing the observation of students conducting these initial assessment interviews.

Measures

A slightly modified version of the Minnesota Therapist Rating Scale (MTRS; DeRubeis, Hollon, Evans, & Bemis, 1982) was used to rate the counsellor's performance during the interview. The scale used by DeRubeis was originally designed to differentiate between elements of CBT and Interpersonal Psychotherapy, and has modest to good psychometric properties including the ability to differentiate reliably between different therapeutic approaches. The MTRS has four subscales, derived from factor analyses: cognitive-behavioural technique (15 items; e.g., Did the therapist work with the client to break problems into their smaller component aspects? To what extent did the counsellor examine the validity of the client's beliefs?), generic therapeutic skills (10 items; e.g., how much rapport was there between therapist and client?), therapist directiveness (4 items; e.g., in general, the person who initiated changes in the flow of the direction of the session was the counsellor/client), and interpersonal psychotherapy skills (IPT**;** 3 items; e.g., to what extent did the content of the session focus on the client's interpersonal relationships). Each item was rated on a Likert-type scale ranging from 1 to 9. Anchors at each end of the scale captured the poles of each item dimension (e.g., *Excellent Rapport – Absence of Rapport, Not at all-Extensively).*

The MTRS was selected for several reasons. First, the CBT components of the scale were consistent with the primary orientation of the clinical psychology training

program. Second, the MTRS also captured features of interpersonal psychotherapy and a wide range of general counselling skills that would be expected in any therapist-client interaction (e.g. rapport). Thus, most items were also applicable to sessions that were primarily assessment focussed. Finally, unlike other rival measures such as the Clinical Skills Assessment Rating Form (Tweed, Graber & Wang, 2010) that yields judgments of "Pass", "Borderline" or "Fail," the MTRS provided ratings that were formative and less evaluative.

The current study used the 32-item scale and response format adopted by DeRubeis, but omitted 5 items (2 from the CBT subscale and 3 from the Generic Therapeutic Skills subscale) that were clearly not applicable to an initial clinical assessment interview (e.g., To what extent do you think the client accepted the nature of therapy?). All four subscales of the original scale were represented in the current measure (CBT, 13 items; General Therapeutic Skills, 7 items; Therapist Directiveness, 4 items; IPT**;** 3 items). Trainees received an introduction to the scale but none of the raters received any standardised training for scoring. We chose not to provide extensive training in the use of these ratings because, in our experience, peer review processes rarely include standardisation in procedures, calibration of judgements, or systematic training in assessing psychotherapeutic skills in others.

Procedure

The clinical interviews were one-to-one intake interviews of about one-hour duration. These interviews were scheduled on a weekly basis and were a routine part of the clinical psychology training at the University. The interviews were allowed to proceed uninterrupted without any feedback or intervention during the session. Clients gave informed consent for the interviews to be observed and, when requested, were given the opportunity to meet the observers behind the one-way mirror. The rationale

and purpose for collecting ratings using a standard scale was explained to all supervisees and supervisors and they agreed to complete the rating forms for research purposes. Additionally, procedures for access to de-identified data was reviewed and approved by the Ethics Committee of the University of Wollongong.

Immediately after the interview, supervisor and peer observers completed the ratings on the modified MTRS. All ratings were completed individually without consultation. The counsellor joined the clinical supervisors and peers to participate in a feedback and discussion session after observers completed their ratings.

Data Analyses and Results

Data Sets

Of the 32 interviews, 2 interviews were excluded from analyses because they were outliers in that they elicited a large number of "not applicable" ratings for most items on the scale. The 30 interviews produced 44 sets of ratings by supervisors (16 interviews were rated by 1 supervisor, 13 interviews were rated by 2 supervisors, and 1 interview by 3 supervisors) and 96 sets of ratings by peers (each interview was rated by 2-5 trainees).

Multi-dimensional Mapping of the Ratings

Data Set A were analysed using a multidimensional scaling approach (using the PROXSCAL algorithm in SPSS) to determine the underlying similarity/dissimilarity of the ratings for the 30 interviews and to compare student and supervisor raters, using the subscale scores of the MTRS as variables. The model indicated good fit to the data (*Stress-1 = 0.056, normalized raw stress = 0.003*).

The results suggested a similar two-dimensional structure for both peers and supervisors (Figure 1). The first dimension (*X*-axis) reflected the extent to which the session was structured, focused, and counsellor driven versus less focused, less

structured and client-driven. Within this dimension, low therapist directiveness and high levels of general counsellor skills (e.g., rapport and alliance) anchored one end, whilst IPT with its specific focus on interpersonal relationships anchored the opposite pole. The collaborative-empiricism of CBT techniques fell in the middle. An examination of the dimension weights suggested a subtle but significant difference between the two categories of raters, with supervisors' ratings dispersed across a wider range of this dimension whereas peer ratings clustered slightly more towards the middle of the dimension. The second dimension (*Y*-axis) related to generic versus specialised therapy techniques, with generic skills anchoring the top end whilst CBT and IPT anchored the opposite pole.

Inter-rater Agreement within Peer and Supervisor Subgroups

To examine between-peer and between-supervisor agreement, we used a subset of the data that comprised all interviews (n = 14) that had ratings by a minimum of two supervisors and two peers. At least two supervisor and two peer ratings were required to calculate inter-rater agreement coefficients. Between supervisor correlations were high on general counseling skills and IPT (intra-class correlations being above .70 in both instances; $p < .01$) and modest on the CBT subscale, ($r = .52$, $p < .10$). Between peer ratings were modest on the CBT, general counseling skills and IPT ($r = .50$ or above in each instance, $p < .10$). Inter-rater agreement for therapist directiveness was poor for both supervisors and peers, with the correlation being negative for peers, suggesting more disagreement than agreement.

Supervisor vs. Peer Ratings

In order to examine the extent to which peer ratings agreed with supervisor ratings, we first established an anchor supervisor for each of the 30 interviews.

Sixteen interviews were rated by a single supervisor who was designated the anchor supervisor. In fourteen interviews where there was more than one supervisor, the anchor supervisor was determined randomly. The 30 interviews generated 92 sets of ratings from peers. Difference scores (compared to anchor supervisor) were computed and scaled on a continuum from complete agreement (identical scores) to levels of disagreements. The data were subjected to χ^2 analyses based on frequency counts over 4 levels of agreement (ranging from ratings that matched, to ratings that deviated by 1, 2, or 3 or more points) for each item. Items that were significant at $p < .05$ level and that were in the direction of better agreement than disagreement are presented in Table 1. Because there was poor between-supervisor agreement for therapist directiveness, this variable was not analysed further.

Overall, there was significant agreement between peer and supervisor ratings for two of the subscale scores. Specifically, about 75% of peer ratings were within good/acceptable agreement limits (within 1 score of the supervisor's rating) for the CBT subscale, and 58% of peer ratings were within acceptable limits for the Generic counseling skills subscale. The frequency distribution for the IPT subscale was in the expected direction, but failed to reach statistical significance.

Peer-to-supervisor agreement for individual items fared much more poorly. Peer ratings showed good agreement with supervisor ratings on only 4 of 27 items. Specifically, 60% or more of peer ratings were within 1 difference point of the supervisor's rating on the following items: rapport, appropriate examination of early relationships in the interview, appropriate amount of client-counsellor verbalizations, and the use of behavioural experiments during the interview. Supervisor-peer agreement was particularly poor on four items: collaborating on an agenda, use of homework, use of open-ended questions, and the adequate use of psychoeducation.

On these items ratings varied by margins that were unacceptable (60% or more peer ratings deviated by 2 or more points, including 40% of ratings that differed by 3 or more points).

Biases Affecting Peer Ratings

Multiple ratings by peers and supervisors for the same interview were averaged to derive mean supervisor and peer ratings for each interview. Ratings were subjected to 2 Group (Supervisor/Peer) X 4 Competencies (4 subscales) ANOVA (Figure 2). The results indicated that counsellors received better ratings from both supervisors and peers for general counsellor skills and therapist directiveness than they did for specialised technical skills (CBT and IPT). Compared with supervisors, peers rated counsellor performance more leniently, giving their peers better scores on general counsellor skills ($p < .005$) and therapist-directiveness ($p < .05$). Peer and supervisor ratings of the counsellor's CBT and IPT skills did not differ.

Discussion

Observation of counsellor performance is a highly recommended method of supervision within professional counselling and psychology (Kaslow et al., 2009; Liese & Beck, 1997; Padesky, 1996). However, its application in supervisory practice is less frequent than desirable (e.g., Townend et al., 2002) and the criteria supervisors use to formulate their evaluations and the reliability of these evaluations have been poorly researched (Gonsalvez & McLeod, 2008; Gonsalvez & Milne, 2010). The study systematically examined supervisor and peer ratings of assessment competencies demonstrated by clinical psychology trainees, and contributes to a better understanding of factors influencing supervisor-peer evaluations. Four issues were examined and will be addressed in order.

Dimensional Structure of Supervisor and Peer Ratings

Overall, an analysis of the dimensional structure of the ratings of the two groups yielded multidimensional solutions that were very similar for the two groups (see Figure. 1). The dimensional structure is a valuable tool to examine in a global sense how raters cluster variables (competencies). Because it is atheoretical, the emergent dimensions are a product of the data and are not confounded by untested assumptions. The overall similarity between the two groups suggest that, at least at a macro-level, both supervisors and peers are using similar dimensions to structure the diverse set of clinical competencies they rated.

Competency-based approaches have dominated recent thinking within clinical supervision (Falender & Shafranske, 2014; Falender, Shafranske & Ofek, 2014; Kaslow et al., 2004), and have spawned the development of competency frameworks that usually include a large number of discrete competencies organised across multiple foundational and functional domains (e.g., Fouad et al., 2009). The proliferation of items may not be supported by empirical approaches that often yield fewer dimensions (e.g., Gonsalvez & Freestone, 2007). Statistical approaches that capture underlying factors, components, or clusters might be helpful to ensure that additional ratings that supervisors are called to make are better grounded and informed by research.

Between Supervisor Ratings

Unfortunately, our results concerning inter-rater reliabilities between supervisors is based on a small sample (n = 14) and generalizations should be made with caution. As may be expected between-supervisor agreement is better than between-peer agreement. The level of agreement reported is modest but is consistent with findings from previous studies where untrained raters are used (Kaslow et al., 2009; Tweed et al., 2010). Researchers have highlighted the need for the development

and more efficient use of structured and psychometrically validated scales within supervision (Ellis, Ladany, Krengel, & Schult, 1996; Gonsalvez & McLeod, 2008). It is possible that better inter-rater reliability (between and within supervisor and peer groups) would have been achieved through training of raters. Future research should examine whether more extensive training, particularly around calibration of scores leads to increased inter-rater reliability.

Agreement Between Supervisor and Peer Ratings

In general, there was better agreement between supervisors and peers on subscale scores than there was on individual items. Of the three subscales examined, 75% of peer ratings were within acceptable agreement limits (within 1 score of the supervisor's rating) for the CBT subscale, and 58% of peer ratings were within acceptable limits for the Generic counseling subscale. The frequency distribution for the IPT subscale was in the expected direction, but failed to reach statistical significance. Peer-to-supervisor agreement for individual items fared much more poorly. Peer ratings showed good agreement with supervisor ratings on only 4 of 27 items. Additionally, the variability between ratings of different peers is sufficiently substantive to be of concern. For instance, the correlation for therapist directiveness within the peer group was negative, suggesting significant disagreement rather than agreement, and 40% of peer ratings were unacceptably deviant (3 or more difference points) on 4 of the 27 items rated. Thus, only the CBT subscale could be recommended for peer ratings and the generic counseling subscale could be used cautiously. However, the low levels of agreement provide valuable data for supervisors who are facilitating feedback sessions involving peer trainees. In those areas where there tends to be low levels of agreement (e.g., therapist directiveness), the supervisor may need to focus on those behaviours that reflect directiveness. Where

sessions have been videotaped, this can be done by reviewing specific examples of the target behaviours. However, when one-way mirrors are used without video-recording, there may be a need to note specific interactions in preparation for discussion during the review and feedback session.

Biases Affecting Peer Ratings

Peer ratings were more lenient (higher) than supervisor ratings on general counselling skills but not on specialised techniques such as CBT and IPT. A leniency bias occurs when ratings of performance are inflated in a positive direction. This becomes apparent particularly when ratings are compared to other performance indicators (e.g., positive subjective appraisal and ratings despite formal fidelity ratings of observed therapy session suggesting only average or poor performance). It is worth noting that rating biases may affect supervisor ratings as well, with research indicating that supervisors' summative assessments are vulnerable to leniency and halo biases (Gonsalvez & Freestone, 2007; Gonsalvez et al., 2013; Robiner et al., 1997), especially when supervisor evaluations occur after a long supervisor-trainee relationship (e.g., at placement end). The current study indicates that leniency biases may be exaggerated in peer evaluations. Consequently, an over emphasis on peer evaluations might lead to inflated self-appraisals and a failure of the trainee counsellor to address inadequacies.. The lack of clearly operationalised criteria and concerns about the subjectivity inherent in evaluation are likely to underlie leniency trends in both supervisors and peers (Robiner et al., 1997; Wahnon, Deane & Gonsalvez, 2014). Additionally, concern over peer disapproval associated with critical appraisal of their performance may also contribute to larger leniency effects observed in peer evaluations..

Limitations

The study's main aim was to determine whether peer ratings of counsellor performance agreed with those of supervisor ratings. The assumption that supervisor ratings are themselves reliable and valid was not robustly tested. Thus, the poor agreement identified for peer ratings may be inflated by variability and inaccuracies of supervisor ratings. Further, because no training was provided to either the supervisors or peers, generalizing these findings to initiatives that include adequate training to raters may be premature and unwarranted. Finally, the study was conducted in a clinical psychology training program and it is unclear how these findings might apply to peer group supervision models outside of this context.

Conclusions

Comment [CG1]: Line 50ff - this sentence is quite cumbersome and needs to be re-phrased for clarity.

Overall the multidimensional scaling results suggested that ratings of both peers and supervisors tend to reflect similar underlying structures or constructs. The exact nature of these underlying dimensions is open to interpretation. Obviously, the content of the scale is likely to influence the number and nature of the dimensions observed. For example, the Cognitive Therapy Rating Scale-Revised (Blackburn, et al., 2001) would provide a more focused analysis of CBT skills. In the current study, the main point was that similar dimensions appeared to be rated. However, agreement about the relative strength of these dimensions did vary.

Agreement between peer and supervisor ratings was acceptable for mean scores derived for the CBT and the Generic counselling skills subscales, but not for the IPT subscale. Given that the training program emphasized a CBT approach, adequate peer-supervisor agreement observed on overall CBT ratings is encouraging. There were low levels of agreement on most individual items. Notably, we observed high levels of disagreement for four items: the use of open-ended questions, negotiation of an agenda, prescribing homework tasks and the use of

psychoeducation. The implication for practice is that these items may require better defined anchors or more clearly defined criteria to enhance rater reliability, especially if the scale is to be used by both supervisors and trainees. Additionally, when providing formative feedback about these specific competencies, supervisors should explain the differences between low, intermediate and high ratings on the scale (for instance, by providing examples) rather than assume that trainees already possess the ability to make these judgments with accuracy. In feedback sessions such clarification could be made in several ways. First, more formal training in the use of particular rating scales would be advisable. This serves the dual purpose of reinforcing key skills which are to be learned as well as making explicit the criteria for competence ratings. Second, peers should be encouraged to clarify the reasons for differences that may occur between self, supervisor and peer ratings.. Video and audio recordings have an advantage in this analytic and reflective process, since they can be replayed and reviewed. In contrast, the effectiveness of techniques that rely on direct observation only (e.g., monitoring through one way mirror or video camera) are compromised by their reliance on the observer's memory or less reliable recording procedures such as note taking. Although experts recommend that supervisors discriminate between and appropriately label criterion-based versus subjective feedback, (Bernard & Goodyear, 2014), survey results indicate that only 52% of supervisors indicated that they regularly and clearly express subjective feedback as personal opinion (Wahnon et al., 2015). Discrepancies between supervisor, peer and self evaluations of counsellor performance highlight the need for supervisors to be reflective and deliberate about their feedback, because good feedback not only identifies how well or poorly the counsellor performed, but clarifies what the counsellor did to merit the concerned evaluation.

Peers tend to provide ratings that are more positive than may be warranted, at least from the supervisor's perspective. These differences were particularly notable for general counsellor skills and therapist directiveness. There is now a need to more systematically examine the factors that might improve consensus (e.g., developmental stage, training, and feedback processes) between evaluations by self, supervisors and peers about counsellor behaviours in assessment and therapy.

References

Bennett-Levy, J., Lee, N., Travers, K., Pohlman, S., & Hamernik, E. (2003). Cognitive therapy from the inside: Enhancing therapist skills through practicing what we preach. *Behavioural and Cognitive Psychotherapy*, *31*, 143–158. doi:10.1017/s1352465803002029

Benshoff, J.M. (1993). Peer supervision in counselor training. *Clinical Supervisor, 11*, 89-102. doi: 10.1300/J001v11n02_08

Benshoff, J. M. & Paisely, P.O. (1996). The structured peer consultation model for school counselors. *Journal of Counseling & Development, 74*, 314-318. doi:10.1002/j.1556-6676.1996.tb01872.x

Bernard, J. M., & Goodyear, R. (2014). *Fundamentals of clinical supervision* (5th ed.). Upper Saddle River, NJ: Pearson Education Inc.

Blackburn, I., James, I. A., Milne, D. L., Baker, C., Standart, S., Garland, A., & Reichelt, F. K. (2001). The Revised Cognitive Therapy Scale: Psychometric properties. *Behavioural and Cognitive Psychotherapy, 29,* 431-446. doi:10.1017/S135246580100404

Borders, L. D. (1991). A systematic approach to peer group supervision. *Journal of Counseling and Development, 69*, 248-252. doi: 10.1002/j.1556-6676.1991.tb01497.x

Campbell, T. W. (1994). Psychotherapy and malpractice exposure. *American Journal of Forensic Psychology, 12*, 5-41. Retrieved from http://www.forensicpsychology.org/journal.htm

Crutchfield, L. B., & Borders, D. L. (1997). Impact of two clinical peer supervision models on practicing school counselors. *Journal of Counseling and Development, 75,* 219-229. doi: 10.1002/j.1556-6676.1997.tb02336.x

DeRubeis, R. J., Hollon, S. D., Evans, M. D., & Bemis, K. (1982). Can psychotherapies for depression be discriminated? A systematic investigation of Cognitive Therapy and Interpersonal Therapy. *Journal of Consulting and Clinical Psychology, 50,* 744-756.doi:10.1037/0022-006X.50.5.744

Ellis, M. V., Ladany, N., Krengel, M., & Schult, D. (1996). Clinical supervision research from 1981 to 1993: A methodological critique. *Journal of Counseling Psychology, 43*, 35-50. doi:10.1037//0022-0167.43.1.35

Falender, C. A., & Shafranske, E. P. (2014). Clinical supervision in the era of competence. In W. B. Johnson & N. Kaslow (Eds.). *Oxford handbook of education and training in professional psychology*. (pp. 291–313). New York, NY: Oxford University Press. doi:10.1093/oxfordhb/9780199874019.013.022

Falender, C.A., Shafranske, E.P. & Ofek, A. (2014). Competent clinical supervision: Emerging effective practices. *Counselling Psychology Quarterly*, 27(4), 393-408. doi:10.1080/09515070.2014.934785

Fouad, N. A., Grus, C. L., Hatcher, R. L., Kaslow, N. J., Hutchings, P. S., Madson, M., et al. (2009). Competency benchmarks: A model for the understanding and measuring of competence in professional psychology across training levels. *Training and Education in Professional Psychology*, *3*(4, Suppl.), S5–S26. doi:10.1037/a0015832

Gonsalvez, C.J., Bushnell, J., Blackman, R., Deane, F., Bliokas, V., Nicholson-Perry, K., ...Knight, R. (2013). Assessment of Psychology Competencies in Field Placements: Standardized Vignettes Reduce Rater Bias. *Training and Education in Professional Psychology, 7*(2) 99-111. doi:10.1037/a0031617

Gonsalvez, C. J., & Calvert, F. L. (2014). Competency-based models of supervision: Principles and Applications, Promises and Challenges. *Australian Psychologist, 49*, 200-208. doi:10.1111/ap.12055

Gonsalvez, C. J., & Freestone, J. (2007). Field supervisors' assessments of trainee performance: Are they reliable and valid? *Australian Psychologist, 42*, 23-32. doi:10.1080/00050060600827615

Gonsalvez, C. J., Hyde, J., Lancaster, S., & Barrington, J. (2008). University psychology clinics in Australia: Their place in professional training. *Australian Psychologist, 43*, 278-285. doi:10.1080/00050060802413529

Gonsalvez, C. J., & McLeod, H., (2008). Toward the science-informed practice of clinical supervision: The Australian context. *Australian Psychologist, 43(2),* 79-87. doi:10.1080/00050060802054869

Gonsalvez, C. J., & Milne, D. L. (2010). Clinical supervisor training in Australia: A review of current problems and possible solutions. *Australian Psychologist, 45,* 233–242. doi:10.1080/00050067.2010.512612

Gonsalvez, C. J., Oades, L. G., & Freestone, J. (2002). The objectives approach to clinical supervision: Towards integration and empirical evaluation. *Australian Psychologist, 37*, 68-77. doi:10.1080/00050060210001706706

Kaslow, N. J., Borden, K. A., Collins, F. L., Forrest, L., Illfelder-Kaye, J., Nelson, P.D.,…Willmuth, M. E. (2004). Competencies conference: Future directions in education and credentialing in professional psychology. *Journal of Clinical Psychology, 60*, 699-712. doi: 10.1002/jclp.20016

Kaslow, N. J., Grus, C. L., Campbell, L. F., Fouad, N. A., Hatcher, R. L., & Rodolfa, E. R. (2009). Competency assessment toolkit for professional psychology. *Training and Education in Professional Psychology, 3* (4, Suppl.), S27-S45. doi:10.1037/a0015833

Kazantzis, N., & Munro, M. (2011). The emphasis on Cognitive-Behavioural Therapy within clinical psychology training at Australian and New Zealand Universities: A

survey of program directors. *Australian Psychologist, 46,* 49-54. doi:10.1111/j.1742-9544.2010.00011.x

Lazar, A., & Mosek, A. (1993). The influence of the field instructor-student relationship on evaluation of students' practice. *The Clinical Supervisor*, *11*, 111-120. doi:10.1300/J001v11n01_08

Milne, D. L. (2009). *Evidence-based clinical supervision: Principles and practice*. Chichester, England: BPS Blackwell. doi:10.1002/9781444308662

Norcross, J. C., Hedges, M., & Castle, P. H. (2002). Psychologists conducting psychotherapy in 2001: A study of the Division 29 membership. *Psychotherapy: Theory, Research, Practice, Training, 39,* 97–102. doi:10.1037//0033-3204.39.1.97

Reiser, R. P. (2014). Supervising cognitive and behavioural therapies. In C. L. Watkins & D.L.Milne (Eds.), *The Wiley International Handbook of Clinical Supervision* (1st ed.) (pp. 493-517). Chicheser, England: Wiley. doi:10.1002/9781118846360.ch24

Robiner, W. N., Saltzman, S. R., Hoberman, H. M., Semrud-Clikeman, M., & Schirvar, J. A. (1997). Psychology supervisors' bias in evaluations and letters of recommendation. *Clinical Supervisor, 16*, 49-72. doi:10.1300/J001v16n02_04

Spence, S. H., Wilson, J., Kavanagh, D., Strong, J., & Worral, L. (2001). Clinical supervision in four mental health professions: A review of the evidence. *Behaviour Change, 18*, 135-155. doi:10.1375/bech.18.3.135

Stoltenberg, C. D., Bailey, K.C., Cruzan, C.B., Hart, J.T., & Ukuku, U. (2014). The integrated developmental model of supervision. In C. L. Watkins & D.L.Milne (Eds.), *The Wiley International Handbook of Clinical Supervision* (1st ed.) (pp. 576-597). Chicheser, England: Wiley. doi:10.1002/9781118846360.ch28

Townend, M., Iannetta, L., & Freeston, M. H. (2002). Clinical supervision in practice: A survey of UK cognitive behavioural psychotherapists accredited by the BABCP.

Behavioural and Cognitive Psychotherapy,30, 485–500. doi:10.1017/S1352465802004095

Tweed, A., Graber, R., & Wang, M. (2010). Assessing trainee clinical psychologists' clinical competence. *Psychology Learning and Teaching, 9*, 50-60. doi:10.2304/plat.2010.9.2.50

Wahnon, T., Deane, F. P., & Gonsalvez, C. (2015). Goal-setting, feedback, and evaluation practices reported by clinical supervisors. Manuscript submitted for publication.

Figure legends

Figure 1. Multi-dimensional scaling structure for supervisor and peer ratings of counsellor competencies.

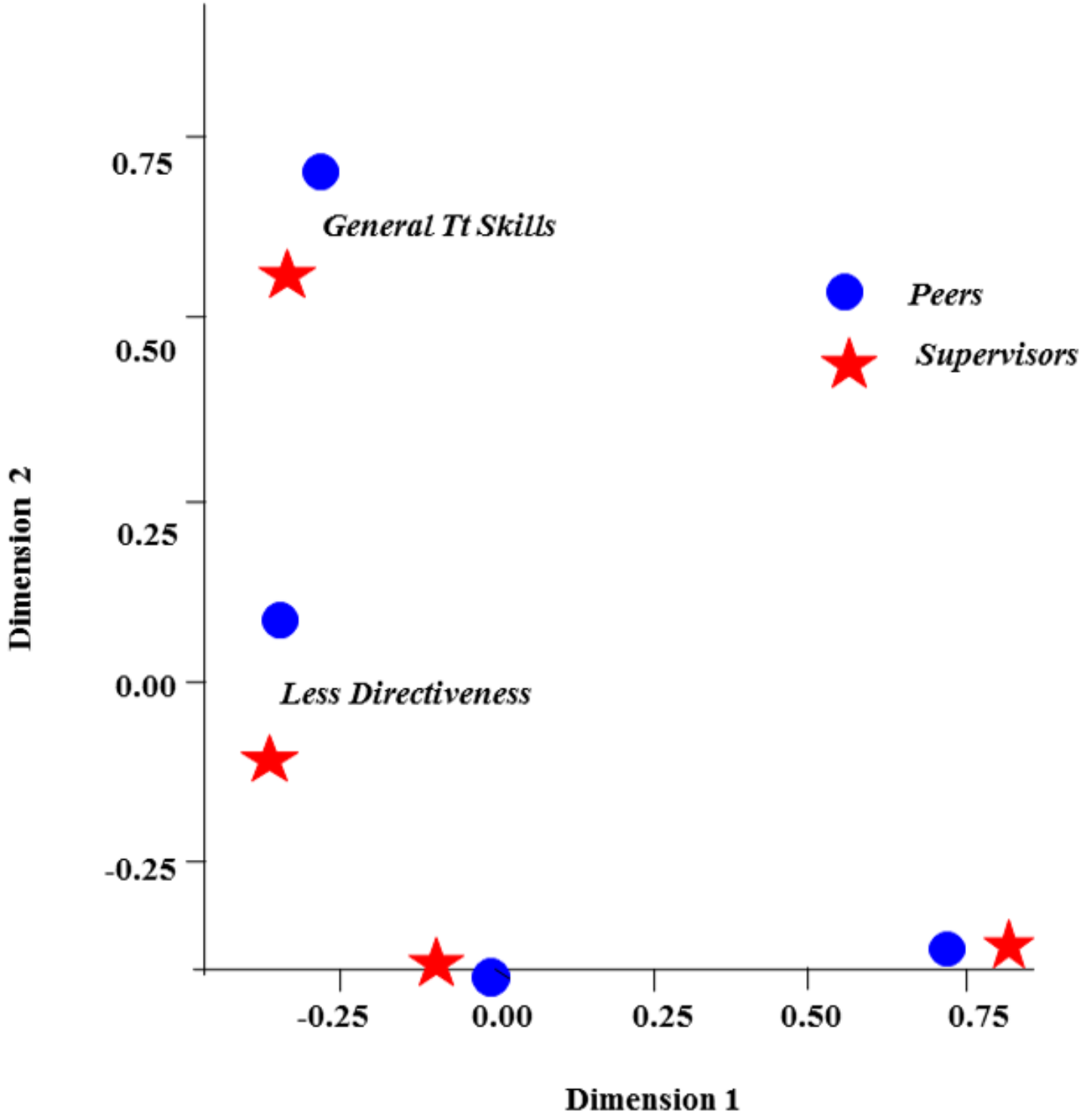
Figure 2. Peer and supervisor ratings of counsellor competencies. (Note: CBT = Cognitive behaviour therapy; Gen Skills = Generic counsellor skills; IPT = Interpersonal psychotherapy; CqDirect = Therapist-directiveness.)

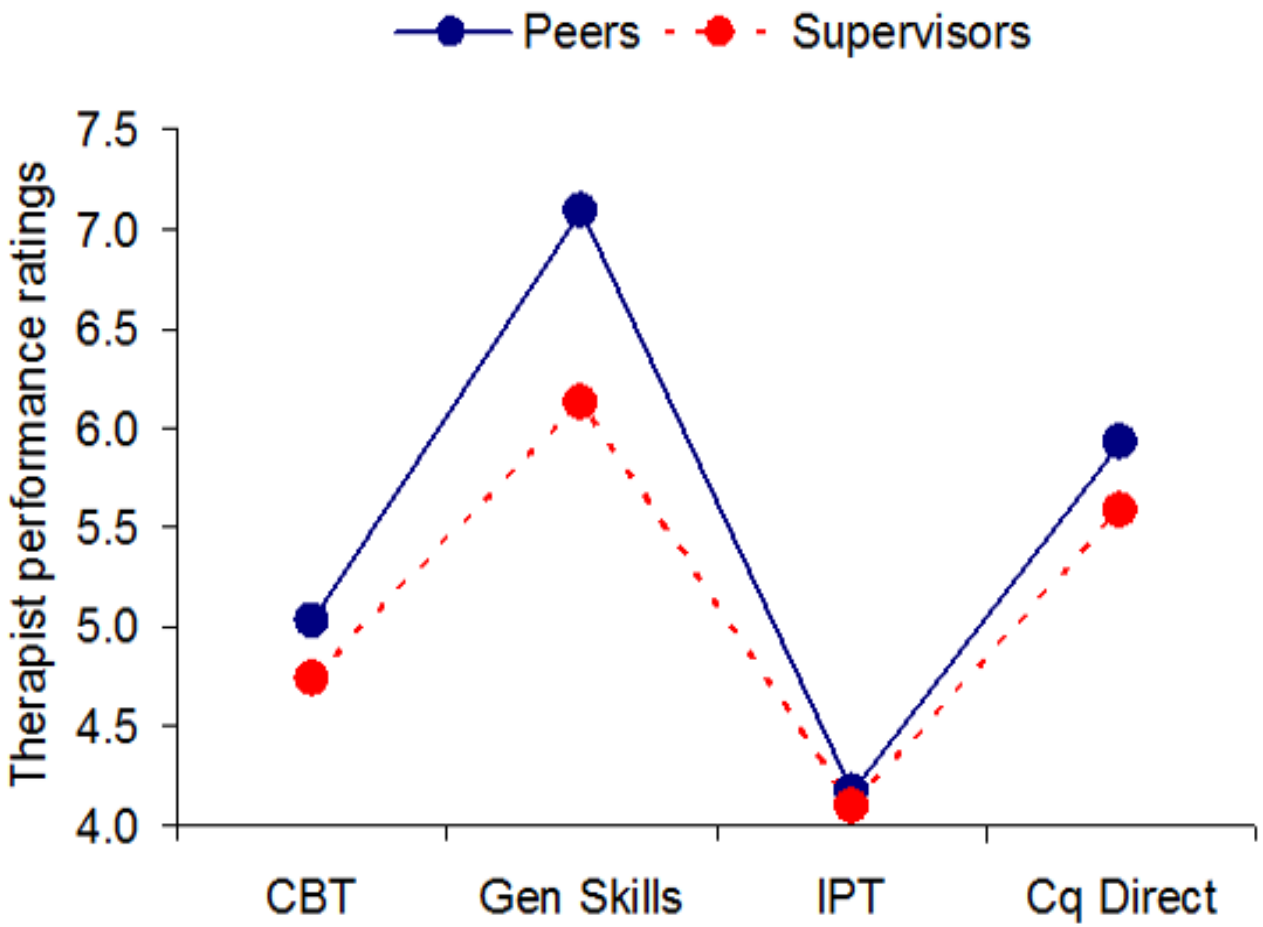
Table 1.

Frequency (in percentages) of Peer-Supervisor Agreement, along a 4-item Scale of Agreement/Disagreement *(δ = Peer-Supervisor Difference Scores).*

Items	χ^2 value	Perfect Agreement	Agree (δ=1)	Disagree (δ=2)	Disagree (δ=3)
		Subscales			
Generic Cq. skills	10.17*	24	34	30	12
CBT	32.09***	28	47	12	12
		Individual Items			
Rapport	28.09***	14	46	29	11
Early relns.	8.04*	32	35	14	19
Cqr:Ct talk ratio	15.56***	17	42	23	17
Beh. Expt.	16.67***	50	21	12	17
Open-ended questions	21.48$^{\#\#\#}$	8	28	23	42$^{\#}$
Negotiated agenda	17.42$^{\#\#\#}$	10	21	27	42$^{\#}$
Homework	15.74$^{\#\#\#}$	17	20	13	50$^{\#}$
Psychoeducation	10.43$^{\#}$	14	23	20	43$^{\#}$

Note: Cq = counselling; Cqr = counsellor; Early relns = extent to which the counsellor related current problems to client's early relationships; Homework = Whether the Cqr prescribed homework tasks. Beh Expt = Whether a behavioural experiment was used and its adequacy. *values are significant in the direction indicating good agreement; $^{\#}$values are significant in the direction suggesting poor agreement with supervisors. *$^{\#}p < .05$. **$^{\#\#}p < .01$. ***$^{\#\#\#}p < .001$.


0.75
0.50
0.25
0.00
-0.25
Dimension 2
General Tt Skills
Peers
Supervisors
Less Directiveness
-0.25
0.00
0.25
0.50
0.75
Dimension 1


Peers
Supervisors
Therapist performance ratings
7.5
7.0
6.5
6.0
5.5
5.0
4.5
4.0
CBT
Gen Skills
IPT
Cq Direct