Australian Psychologist

APS Australian Psychological Society

SPECIAL ISSUE

Brief Online Training with Standardised Vignettes Reduces Inflated Supervisor Ratings of Trainee Practitioner Competencies

Josephine Terry,[1]* Craig Gonsalvez[1], and Frank Patrick Deane[2]

[1]School of Social Sciences and Psychology, Western Sydney University, and [2]School of Psychology, University of Wollongong

Objective: Supervisor assessments of trainee competence are integral to ensuring that clinical psychology trainees reach competency benchmarks. The commonly used Clinical Psychology Practicum Competencies Rating Scale (CΨPRS) has been shown to elicit inflated ratings of competency. Hence, the aim of this study is to examine whether brief supervisor training reduces ratings by providing objective criteria with which supervisors can assess trainee competency.

Method: The ratings included were of 124 psychology trainees from nine Australian university clinical programmes. Of 170 supervisors, 32 completed the online training immediately prior to commencing the CΨPRS. Training required supervisors to rate the competency level described in five standardised vignettes (Beginner through to Competent). Vignette ratings, as determined by a panel of expert supervisors, were provided as feedback. A sixth *calibration vignette* was also rated (no feedback provided). Firstly, CΨPRS ratings from the trained and untrained supervisors were compared. Secondly, the difference between supervisor and expert ratings of the calibration vignettes were compared across trained and untrained groups.

Results: Trained supervisors provided lower CΨPRS ratings than untrained supervisors. In addition, trained supervisors (vs untrained supervisors) provided ratings of the calibration vignette that more accurately matched the ratings provided by the expert panel.

Conclusions: Brief online training using standardised vignettes was associated with lower CΨPRS ratings. The standardised vignettes helped calibrate supervisors' ratings and likely attuned supervisors to the skills and competency levels that are expected at particular developmental stages. As a consequence, training appeared to reduce ratings, arguably resulting in more accurate assessments of trainee performance.

Key words: competency assessment; field placement evaluations; online training; psychology practicum assessment; rater biases; supervisor evaluations.

What is already known on this topic

1 Likert-type ratings of competencies in psychology such as the Clinical Psychology Practicum Competencies Rating Scale (CΨPRS) are prone to leniency and halo biases.
2 Among other factors, a lack of clear criteria for assessing competence are likely to contribute to the effects of leniency and halo biases
3 Brief training can improve the accuracy and reliability of ratings of performance and functioning in occupational and clinical settings, respectively.

What this paper adds

1 This study highlights the potential value of an innovative method that uses standardised vignettes to help calibrate supervisors' ratings.
2 The article demonstrates that brief online training for supervisors using standardised vignettes improves the accuracy of CΨPRS ratings.
3 The vignette methodology has wide applications for competency assessments in psychology and other health disciplines.

Trainee psychologists are required to reach performance benchmarks set by training institutions and accrediting bodies, such as the Australian Psychology Accreditation Council (APAC), in order to obtain registration and to practice in the field. These benchmarks indicate the minimal level of practitioner knowledge, skills, and conduct required for a trainee to be considered an entry-level graduate psychologist. Accurate competence assessment is critical as it ensures that trainees have attained the standards expected by their training institutions, their employers, and relevant regulatory bodies, and as

Correspondence: Josephine Terry, School of Social Sciences and Psychology, Western Sydney University, Locked Bag 1797, Penrith, NSW 2751, Australia.
Email: j.terry@westernsydney.edu.au

Accepted for publication 30 August 2016

doi:10.1111/ap.12250

expected by clients and the public more broadly (Nelson, 2007; Tweed, Graber, & Wang, 2010).

There are significant and varied implications for stakeholders if performance is not at the required level. Failure to identify deficits in competence risks causing harm to consumers of psychological services and, in extreme cases, can result in negligence and legal consequences. In a broader context, behaviour that violates the standards of professional practice has a detrimental impact on the discipline of psychology (Overholser & Fine, 1990). Identification of underperforming trainees is also critical for directing remediation and possible dismissal (Forrest, Elman, Gizara, & Vacha-Haase, 1999). Therefore, it is important that assessment of competency accurately reflects actual performance. To this end, there has been significant effort directed towards developing valid and reliable competency evaluation rating forms (CERFs). Typically, these assessment forms use Likert scales to rate trainee competence across a range of foundational (e.g., professional skills, ethical attitude and behaviour) and functional (e.g., case conceptualisation, intervention) domains (Fouad et al., 2009; Gonsalvez et al., 2015; Rodolfa, Bent, Eisman, Nelson, & Ritchie, 2005).

Field supervisors have a central role in ensuring the accuracy of assessment, but they have reported that the assessment process is a significant source of stress within their supervisory role (Bogo, Regehr, Roxanne, & Regehr, 2007; Pease, 1988). Therefore, it is important that supervisors have access to, and engage with, resources that are designed to support this role. Given the high degree of responsibility and time constraints placed on field supervisors, these resources need to be easily accessible, time efficient, effective, and applicable to a wide variety of placement types and settings (Chafouleas, Riley-Tillman, Jaffery, Miller, & Harrison, 2015; Olsen, Donaldson, & Hudson, 2010). Given that field supervisors come from diverse settings, any resources designed to assist in the accurate assessment of trainees should be standardised to ensure that there is consistency in supervisors' expectations of trainees and in the assessment process. Indeed, supervisors should be familiar with the assessment tools and of the implications of their assessments. For instance, providing an overly favourable assessment of a trainee might obscure true performance, and consequently, important opportunities for remediation might be lost.

The tendency for supervisors to provide ratings of trainee performance that may not reflect actual performance has been raised in the literature. Recent research has highlighted issues of reliability and validity of trainee assessments, suggesting that Likert-scale assessments in particular are vulnerable to leniency and halo biases (Bogo, Regehr, Hughes, Power, & Globerman, 2002; Gonsalvez & Crowe, 2014; Gonsalvez & Freestone, 2007; Robiner, Saltzman, Hoberman, Semrud-Clikeman, & Schirvar, 1998). For instance, Gonsalvez et al. (2015) examined competency ratings provided by supervisors using the *Clinical Psychology Practicum Competencies Rating Scale* (CΨPRS). The CΨPRS measured competence across nine domains (e.g., *relational skills*, *clinical assessment*), and ratings were based on a four-stage developmental model that placed trainees along a continuum from Beginner (Stage 1) to Competent (Stage 4). Ratings were given at the end of placement and were for trainees at various stages of training. The researchers found that ratings reached the ceiling early in training, with only 1.6% of ratings falling in the lower half of the scale. In fact, at the end of their second placement, trainees were rated as performing at a level consistent with a recent graduate in their first job (*Stage 4: Competent*). This is surprising as in Australia, the second placement is typically the trainee's first field experience outside the university training clinic, and in most cases, they are only halfway through their 2-year training. These findings suggest that practicum assessments are vulnerable to biases that elicit overly favourable ratings of trainees.

Leniency bias is driven by a supervisor's disinclination to give low ratings of performance, and instead, they report trainees as having displayed levels of competence higher than they deserved (Robiner, Fuhrman, & Ristvedt, 1993; Wolf, 2015). This may be motivated by a desire to avoid conflict and to be viewed favourably by the trainee (Gonsalvez, Wahnon, & Deane, 2016). It may also be influenced by the extent to which a supervisor defines their role as supportive or evaluative (Vinton & Wilke, 2011). Mediating factors might also include the supervisor's concerns about the consequences low ratings have for the credentialing of the trainee and a desire to be perceived positively by the relevant stakeholders at the trainee's institution. A lack of familiarity with the assessment tool and the anchor points along the rating scale may also play a role (Dudek, Marks, & Regehr, 2005). In a survey of 113 supervisors, 58% indicated they believed their ratings of supervisees were subject to leniency bias. When asked to rate the likely source of this bias from a list of 11 options, most (52%) indicated that a "lack of objective measures for competence and incompetence" and "lack of clear criteria for competence and incompetence" (43%) contributed strongly or very strongly to biases in these assessments (Gonsalvez et al., 2016).

A second bias known to affect supervisor ratings in psychology (Gonsalvez & Freestone, 2007) and in other health disciplines is the halo effect (Bogo et al., 2002; Pease, 1988; Wolf, 2015). The halo bias occurs when an overall impression of an individual systematically biases ratings on a range of different traits in the direction of the valence. Thus, halo effects could be positive and lead to inflated scores or be negative, leading to lower ratings than are warranted. Remediation of the effects of leniency and halo biases on practicum ratings can take two courses. The first involves increasing the reliability of the assessment tool and by making changes that facilitate accurate ratings (e.g., randomising item order, providing clear anchor points along the rating scale, non-Likert scale forms of assessment; see Bogo et al., 2002; Gonsalvez et al., 2013, 2015; Gonsalvez & Freestone, 2007). The second course of action assumes that leniency and halo biases are a rater issue, and remediation involves changing rater behaviour through education and feedback (e.g., Stamoulis & Hauenstein, 1993; Støre-Valen et al., 2015). Of course, an either/or approach is unlikely to yield the desired results because these two factors most likely interact.

As a first step in minimising the impacts of rater bias, our research investigates the effects of a brief online training on CΨPRS ratings. Prior research in other disciplines and settings (e.g., medicine, clinical assessment, behavioural assessment of children, occupational skills assessment) has suggested that behavioural assessment training (both online and in vivo) is a means for reducing bias and improving rating accuracy (Chafouleas et al., 2015; Jelley & Goffin, 2001; Schanche,

Høstmark Nielsen, McCullough, Valen, & Mykletun, 2010; Stamoulis & Hauenstein, 1993; Støre-Valen et al., 2015; Thornton & Zorich, 1980). For instance, online training has been shown to improve the reliability of Global Assessment of Functioning (GAF)[1] scores (Støre-Valen et al., 2015). The training required mental health clinicians to rate vignettes that described fictitious cases that reflected a range of functioning. After each vignette, the clinician's rating was compared to the mean rating provided by a panel of experts, and the deviation score was fed back to the clinicians. The authors found that there was an increase in the concordance between clinician and expert ratings as a function of training. Our study takes a similar vignette-based training approach, but in our case, supervisors rate the level of trainee competency described in the vignettes, and expert ratings are provided as feedback. The central question is whether training attenuates the high scores given using the CΨPRS, hypothesised to be driven by rater bias (Gonsalvez et al., 2015).

The CΨPRS adopts a stage-based approach to skills and competency development. In other words, it assumes that trainees improve incrementally over multiple placements and that the developmental trajectory is roughly similar across all domains (Gonsalvez et al., 2015). Multiple items present descriptions of skills integral to each of the 10 competency domains (see Appendix A). Supervisors provide a rating that reflects the developmental stage of the trainee's current performance in reference to a standard of professional practice (*Stage 4: "comprising capabilities and skills on par with a clinical psychologist working in their first job following completion of their Master's degree"*). The rating scale provides two end-point anchors, Stage 1 (*Beginner*) and Stage 4 (*Competent*), with Stages 2 and 3 at equidistant points between Stages 1 and 4. Supervisors are provided with descriptions of the four stages, and these descriptions are used to guide the rating of the trainee along the developmental trajectory (see Appendix B).

For ratings to be accurate, supervisors must become familiar with the descriptions of the stages. They must also develop an accurate mapping of the descriptors to the *in situ* performance of the trainee. In other words, the supervisor must have a good idea of what, for example, an ability to use *active and responsive listening skills* looks like in practice at each of the stages. One possible means of facilitating accurate ratings is to familiarise supervisors to the level of skill expected at the various stages by presenting brief but detailed vignettes that describe realistic performances of a hypothetical trainee at various stages of development. The aim is to minimise the effects of rater bias by providing supervisors an opportunity to develop strong mappings between stage descriptors and *in situ* performance.

In the current research, our aim is to reduce the ceiling effects that are commonly seen in trainee assessments by presenting a brief online training immediately prior to the commencement of trainee assessment. Apart from the empirical goals of the current research, the pragmatic considerations in devising the training were its accessibility and duration (5–10 min). The objective was to integrate the training into the assessment process without dramatically increasing the burden on supervisors. In addition, by integrating the training into the CΨPRS, we could ensure that any impacts of the training were immediately measurable. The training consisted of the presentation of five descriptive vignettes (see Appendix C) to supervisors, to which they assigned a particular stage of development. Importantly, supervisors were provided with feedback as to their accuracy (see Method section for more detail). To measure the effect of the training, a group of supervisors completed the training task, while another group did not (*Trained vs Untrained groups*). Our primary dependent measure was the CΨPRS ratings of actual trainees across nine domains, yielding a 2 × 9 mixed repeated-measures design. It was hypothesised that CΨPRS ratings would be lower in the Trained compared to the Untrained group. In accordance with previous literature (Gonsalvez et al., 2015), we expected differences in ratings across domains, but we did not expect an interaction between training and domain.

To further test the effects of training, all supervisors were presented with a test vignette (*calibration vignette*) immediately prior to commencing the assessment (and immediately after training in the Trained group). Supervisors indicated the developmental stage described in the vignette (as per the training task but with no feedback). We compared stage ratings across the Trained and Untrained groups. It was hypothesised that the stage ratings provided by the Trained group (compared to the Untrained group) would be closer to the intended stage described in the vignette.

Method

Participants

CΨPRS assessments were provided by 170 university clinic and field supervisors with Masters or Doctoral qualifications in clinical psychology from an accredited training institution. In addition, they had relevant post-qualification clinical psychology experience to become eligible for full membership of the Australian Psychological Society (APS) College of Clinical Psychologists and had current supervisor accreditation with the Psychologists Board of Australia. Of the cohort of supervisors participating in this study, 32 opted to complete the voluntary training task prior to assessing their trainees. Supervisors in the Trained group had an average of 8.20 years (standard deviation [*SD*] = 2.32) of experience in clinical practice and 5.69 years (*SD* = 3.77) of supervisory experience. Those in the Untrained group had greater levels of experience in both clinical practice (*M* = 9.41 years, *SD* = 1.39) and in supervising trainees (*M* = 7.14 years, *SD* = 3.34). The difference across groups was significant for both clinical practice experience ($t(168) = 3.95$, $p = .01$) and supervisory experience ($t(168) = 2.24$, $p = .04$). Therefore, correlations between these demographic variables and our dependent measures were conducted. There were no significant correlations between years of (a) clinical practice and CΨPRS scores (mean of nine domains), $r(170) = -.02$, $p = .85$ or calibration vignette measures, $r(170) = .11$, $p = .15$; (b) supervisory experience and CΨPRS scores, $r(170) = -.12$, $p = .13$; or calibration vignette measures, $r(170) = -.02$, $p = .82$. As there were no statistical relationships between the demographic and dependent variables, supervisor experience was not considered a mediating factor in the comparison of the Trained and Untrained groups.

CΨPRS assessments were of 124 psychology trainees enrolled in a Clinical Psychology Masters or Doctoral programme at one of the nine participating Australian universities. Programmes were accredited by the APAC and the Clinical College of the APS. Prior to commencing their professional programme,

trainees had completed a 4-year psychology degree at the undergraduate level.

Over the course of a Clinical Master's programme, trainees typically complete four (six in the case of doctoral programs) field placements. The first placement is in the university's psychology clinic, and subsequent placements are with external agencies. Each placement requires 200–300 placement hours, with a minimum of 80–100 h of face-to-face client contact. The placement setting, client population, disorder, and the severity of conditions being addressed varies widely. The data presented in this study were assessments completed at the end of one or more of the trainee's placements.

Materials and Procedure

The CΨPRS used in this research is an online trainee assessment and is a revision of earlier versions of the CΨPRS (Gonsalvez et al., 2015). It is a 60-item rating scale comprising 10 overall domain items and 50 subdomain items measuring specific skills within each of the 10 domains (*Counselling, Clinical Assessment, Case Conceptualisation, Intervention, Ethical Attitude and Behaviour, Scientist-Practitioner Approach, Professionalism, Psychological Testing, Reflective Practice, and Response to Supervision.* See Appendix A for all items). Each item is rated on a 1.0- to 4.9-point visual analogue scale, ranging from Beginner (*Stage 1*) to Competent (*Stage 4*), with intermediate, equidistant anchors being *Stage 2* and *Stage 3* (see Appendix B for Stage descriptions). Mid-points of the four stages are respectively 1.5, 2.5, 3.5, and 4.5. Stages are in reference to a defined standard of competent professional practice, comprising capabilities and skills on par with a newly graduated clinical psychologist working in their first job. Supervisors first completed the 10 overall domain items. This was followed by a random presentation of the 50 subdomain items.

In addition to the trainee assessment items, an optional brief training was included in the CΨPRS. Supervisors who opted to complete the training were presented with five vignettes that described the performance of a hypothetical trainee (See Appendix C). Each vignette was designed to represent performance at a particular developmental stage (Beginner through to Competent) and covered one of five domains (*Counselling, Clinical Assessment, Intervention, Ethical Attitude and Behaviour, and Professionalism*). Supervisors used a visual analogue scale to indicate the developmental stage described in each vignette. After completing the five training items, supervisors were provided with mean ratings as determined by a panel of expert supervisors from Australian universities ($N = 25$). The calibrators (12 females, 13 males) were mostly university Clinic Directors or Practicum Coordinators and had considerable clinical experience as registered psychologists ($M = 18.68$ years in practice, $SD = 5.37$), as supervisors ($M = 14.76$ years; $SD = 6.83$), and in rating the competencies of psychology trainees ($M = 10.45$ years; $SD = 6.55$). Their ratings acted as a form of feedback, allowing supervisors to compare their own ratings with those of a group of experts.

All supervisors, irrespective of whether they opted to complete the training, rated a single vignette item (*calibration vignette*) immediately prior to providing the trainee assessment. The vignette described the performance of a hypothetical trainee in one of five competency domains (*Case Conceptualisation, Intervention, Psychological Testing, Scientist-Practitioner Approach, and Professionalism*. See Appendix D). One of the five vignettes was randomly selected for each participant. While no feedback was provided, these calibration vignettes had been rated by the same expert calibrators who rated the training vignettes.

Prior to completing the CΨPRS, all supervisors whose data are presented in this study, endorsed an option to provide consent for their de-identified data to be included in the research.

Results

CΨPRS Ratings

A mean was calculated for each of the domains by averaging the overall score and the corresponding subdomain items. A mixed repeated-measures analysis of variance (ANOVA) was conducted on these means with Domain as a within-subject factor (*nine domains*) and Training (*Trained, Untrained*) as a between-subjects factor. Psychological testing was excluded from the analysis as only 75 trainees were assessed on this domain. Analyses revealed a main effect of Domain, $F(8, 1168) = 31.78$, $p = .00$, partial $\eta^2 = .18$. To investigate this main effect, pair-wise comparisons were conducted with an adjusted alpha of $p = .0014$. Means, *SD*s and 95% confidence intervals (CIs) are reported in Table 1, and superscripts indicate the domains that were not significantly different from one another (see Table 1 note for further explanation).

Importantly, there was a main effect of Training with lower ratings in the group that participated in the brief online training task compared to the group that did not, $F(1, 146) = 4.37$, $p = .038$, partial $\eta^2 = .03$ (see Table 1 for means, *SD*s, and 95% CIs). There was no Domain by Training interaction ($p = .43$).

Calibration Difference Score

For each supervisor, their calibration vignette rating was compared to the mean rating for the same item given by the expert supervisors. Specifically, a difference score was calculated by subtracting the mean expert supervisor rating from the supervisor's rating. A positive difference score suggests an inflated rating, whereas a negative difference score indicates stringency.

An independent sample *t*-test compared the mean difference scores across the Trained and Untrained groups (see Figure 1). The results indicated that the Trained group had a lower difference score ($M = -1.0$, $SD = 0.59$) than the Untrained group ($M = .19$, $SD = 0.71$), $t(168) = 2.18$, $p = .03$. The two difference score means (Trained, Untrained) were also compared to zero with one-sample *t*-tests. These revealed that the mean difference between supervisor and expert calibrator scores was significantly greater than zero in the Untrained group ($t(134) = 3.04$, $p = .00$). However, the difference score in the Trained group was not significantly less than zero ($t(34) = .98$, $p = .33$).

Discussion

Our findings suggest that the completion of a brief online training task (5–10 min) lowers supervisor ratings of actual trainee performance (relative to no training). Specifically, supervisors who completed the training task gave lower competency ratings compared to supervisors who did not complete the

Table 1 CΨPRS Rating Means (*M*), Standard Deviations (*SD*), Confidence Intervals (95% CI) for the Total Sample and the Trained and Untrained Groups

Domains	Total sample (*N* = 148)			Trained (*N* = 32)			Untrained (*N* = 116)		
	M	*SD*	95% CI	*M*	*SD*	95% CI	*M*	*SD*	95% CI
Counselling	4.28[a]	0.65	4.09, 4.34	4.10	0.60	3.88, 4.33	4.33	0.66	4.22, 4.45
Clinical assessment	4.20[b]	0.66	3.99, 4.25	4.00	0.60	3.75, 4.21	4.26	0.66	4.14, 4.38
Case conceptualisation	4.16[b]	0.66	3.95, 4.20	3.93	0.60	3.70, 4.16	4.22	0.67	4.10, 4.34
Intervention	4.17[b]	0.66	3.96, 4.22	3.94	0.60	3.72, 4.17	4.23	0.67	4.12, 4.35
Ethical attitude/behaviour	4.37[c]	0.60	4.17, 4.41	4.15	0.56	3.95, 4.36	4.42	0.60	4.32, 4.53
Scientist-practitioner	4.21[ab]	0.66	4.00, 4.26	4.00	0.60	3.77, 4.23	4.26	0.67	4.14, 4.39
Professionalism	4.38[c]	0.61	4.21, 4.45	4.24	0.56	4.03, 4.45	4.42	0.62	4.31, 4.53
Reflective practice	4.28[a]	0.62	4.09, 4.33	4.10	0.55	3.88, 4.31	4.33	0.63	4.22, 4.44
Response to supervision	4.38[c]	0.62	4.19, 4.43	4.20	0.59	4.00, 4.40	4.43	0.61	4.32, 4.54
Total mean	4.20	0.61	4.08, 4.32	4.07	0.61	3.86, 4.28	4.32	0.61	4.21, 4.44

Notes. Rating scale ranges from 1 = Beginner through to 4.9 = Competent. [abc]Domains that share a superscript are not significantly different from each other ($p > .0014$). For example, the counselling, scientist-practitioner, and reflective practice domains share the superscript "a," indicating that their mean ratings are not significantly different from one another. Alternatively, counselling and clinical assessment do not share a common superscript, indicating that their mean ratings are significantly different. CΨPRS = Clinical Psychology Practicum Competencies Rating Scale.

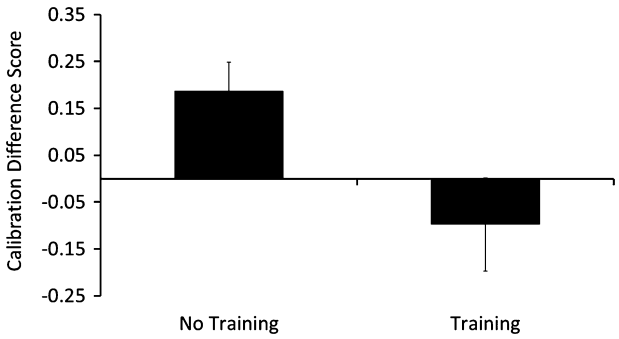

Figure 1 Calibration Vignette: Mean Difference Between the Supervisor Ratings and the Expert Calibrators as a Function of Training (Untrained vs Trained). *Note*. Error bars represent standard error of the mean (SEM).

training. These lowered ratings most likely represent increased accuracy when assessing trainee performance. Further evidence of this arises from our analyses of differential ratings of a calibration vignette between supervisors and experts. We found that the difference in ratings between supervisors and experts was lower if supervisors completed training. The trained supervisor ratings matched the expert calibrator ratings more closely, whereas the untrained supervisors provided significantly higher scores than the experts. Consequently, ratings of the calibration vignettes suggest that training realigns supervisors' mapping of the stage descriptors in CΨPRS to *in situ* performance (as described in the vignettes).

These findings provide evidence for the potential value of using standardised vignettes to calibrate supervisors' ratings. In the context of a brief online training task, these vignettes familiarise supervisors with the performance and competency standards expected at particular points along the developmental continuum. This is consistent with other research demonstrating the effectiveness of training in improving rater reliability and accuracy across varied disciplines (Chafouleas et al., 2015; Jelley & Goffin, 2001; Schanche et al., 2010; Stamoulis & Hauenstein, 1993; Støre-Valen et al., 2015; Thornton & Zorich, 1980). It may be that providing clear frames of reference indirectly attenuates vulnerability to rater biases that is commonly observed in CERFs (Dudek et al., 2005; Gonsalvez et al., 2015; Stamoulis & Hauenstein, 1993). In other words, high ratings potentially driven by rater bias can be reduced to more accurate levels by providing objective and clear criteria (via behaviourally descriptive vignettes) that are matched to a trainee's expected stage of development. Feedback on rating performance during training is also crucial as it allows supervisors to compare their ratings with expert supervisors' ratings. This has the benefit of highlighting potential rater biases and calibrates supervisors' mapping of the CΨPRS stages to actual performance.

Although it is clear that vignette-based training led supervisors to give lower CΨPRS ratings, it is unclear whether post-training ratings represent an attenuation of the leniency and/or halo effects. It would be valuable for future research to tease out the interplay between these two biases. If training reduces leniency effects, post-training CΨPRS ratings would be moderated across low, average, and high ratings. Conversely, if vignette-based training primarily attenuated halo effects (both positive and negative), post-training CΨPRS ratings would be lower in the case of positive halo bias and higher in the case of negative halo bias. Unfortunately, this study was not designed to differentiate between the two biases, and baseline scores obtained by the sample were uniformly high ($M = 4.20$), precluding an examination of training effects across a range of low and high scores.

While our preliminary findings are promising, further research is needed to determine if the effects of training are sustained over time and if training has additional benefits for less experienced supervisors compared to more experienced supervisors (e.g., Støre-Valen et al., 2015). We also note that training and calibration vignettes used in this study were drawn from a subset of competency domains. It is possible that training effects may be strengthened if supervisors are trained on all domains and are required to rate calibration vignettes across all developmental stages and domains. Nonetheless, we found no interaction between Domain and Training, and in fact, training

reduced CΨPRS ratings across all domains, suggesting that the benefits of training carried over to other *untrained* domains. However, this requires further investigation.

Clinical supervisors from a number of professions (e.g., psychology, social work, medicine) have reported that their role as gatekeeper and evaluator is a source of concern, particularly when this conflicts with their goal to maintain a positive and supportive relationship with the trainee (Vinton & Wilke, 2011). Prior research has found that almost half of supervisors are concerned about the effects of feedback on trainees' self-esteem. A total of 35% "strongly" endorsed guilt or fear of feeling responsible for lengthening or terminating their supervisees' education/internship as a source of bias in their ratings (Gonsalvez et al., 2016). Indeed, completing trainee assessments can be difficult and stressful (Bogo et al., 2007; Pease, 1988). These factors have been cited as possible drivers of rater bias (Gonsalvez et al., 2016; Vinton & Wilke, 2011; Wolf, 2015). While we do not claim that vignettes presented in a brief training directly reduce leniency or halo biases, we do suggest that the objective anchor points attenuate the effects of bias on trainee ratings. Specifically, the training vignettes provide an objective, behaviourally rich, and accurate description of expected trainee performance across the developmental continuum. Together, the goals of training are intended to facilitate an accurate mapping of the stage-based descriptors provided within the CΨPRS to *in situ* trainee performance. In summary, our findings indicate that the typically seen inflated ratings of trainee performance can be attenuated with a brief online training that is accessible, feasible, time efficient, and cost effective.

Acknowledgements

Financial support for this publication has been provided by the Australian Government Office for Learning and Teaching. The views in this publication do not necessarily reflect the views of the Australian Government Office for Learning and Teaching. The authors thank the members of the research team, including site coordinators and project consultants.

Note

[1] The GAF is used by mental health clinicians and physicians to measure psychological, social, and occupation functioning.

References

Bogo, M., Regehr, C., Hughes, J., Power, R., & Globerman, J. (2002). Evaluating a measure of student field performance in direct service: Testing reliability and validity of explicit criteria. *Journal of Social Work Education*, *38*(3), 385–401.

Bogo, M., Regehr, C., Roxanne, P., & Regehr, G. (2007). When values collide. *The Clinical Supervisor*, *26*, 99–117. doi:10.1300/j001v26n01_08

Chafouleas, S. M., Riley-Tillman, T. C., Jaffery, R., Miller, F. G., & Harrison, S. E. (2015). Preliminary investigation of the impact of a web-based module on direct behavior rating accuracy. *School Mental Health*, *7*, 92–104. doi:10.1007/s12310-014-9130-z

Dudek, N. L., Marks, M. B., & Regehr, G. (2005). Failure to fail: The perspectives of clinical supervisors. *Academic Medicine*, *80*(10), S84–S87. doi:10.1097/00001888-200510001-00023

Forrest, L., Elman, N., Gizara, S., & Vacha-Haase, T. (1999). Trainee impairment: A review of identification, remediation, dismissal, and legal issues. *Counseling Psychologist*, *27*, 627–686. doi:10.1177/0011000099275001

Fouad, N. A., Grus, C. L., Hatcher, R. L., Kaslow, N. J., Hutchings, P. S., Madson, M. B., … Crossman, R. E. (2009). Competency benchmarks: A model for understanding and measuring competence in professional psychology across training levels. *Training and Education in Professional Psychology*, *3*(4 SUPPL. 1), 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*, 99–111. doi:10.1037/a0031617

Gonsalvez, C., & Crowe, T. (2014). Evaluation of psychology practitioner competence in clinical supervision. *American Journal of Psychotherapy*, *68*(2), 177–193.

Gonsalvez, C., Deane, F., Knight, R., Nasstasia, Y., Shires, A., Nicholson Perry, K., … Bliokas, V. (2015). The hierarchical clustering of clinical psychology practicum competencies: A multisite study of supervisor ratings. *Clinical Psychology*, *22*(4), 390–403. doi:10.1111/cpsp.12123

Gonsalvez, C. J., & Freestone, J. (2007). Field supervisors' assessments of trainee performance: Are they reliable and valid? *Australian Psychologist*, *42*(1), 23–32. doi:10.1080/00050060600827615

Gonsalvez, C., Wahnon, T., & Deane, F. P. (2016). Goal-setting, feedback, and evaluation practices reported by clinical supervisors. *Australian Psychologist*. doi:10.1111/ap.12175

Jelley, R. B., & Goffin, R. D. (2001). Can performance-feedback accuracy be improved? Effects of rater priming and rating-scale format on rating accuracy. *Journal of Applied Psychology*, *86*(1), 134–144. doi:10.1037//0021-9010.86.1.134

Nelson, P. D. (2007). Striving for competence in the assessment of competence: Psychology's professional education and credentialing journey of public accountability. *Training and Education in Professional Psychology*, *1*, 3–12. doi:10.1037/1931-3918.1.1.3

Olsen, H., Donaldson, A. J., & Hudson, S. D. (2010). Online professional development: Choices for early childhood educators. *Dimensions of Early Childhood*, *38*, 12–17.

Overholser, J. C., & Fine, M. A. (1990). Defining the boundaries of professional competence: Managing subtle cases of clinical incompetence. *Professional Psychology: Research and Practice*, *21*(6), 462–469. doi:10.1037/0735-7028.21.6.462

Pease, B. B. (1988). The ABCs of social work student evaluation. *Journal of Teaching in Social Work*, *2*(2), 35–50. doi:10.1300/j067v02n02_04

Robiner, W. N., Fuhrman, M. J., & Ristvedt, S. (1993). Evaluation difficulties in supervising psychology interns. *Clinical Psychologist*, *46*, 3–13. doi:10.1037/e554872011-003

Robiner, W. N., Saltzman, S. R., Hoberman, H. M., Semrud-Clikeman, M., & Schirvar, J. A. (1998). Psychology supervisors' bias in evaluations and letters of recommendation. *Clinical Supervisor*, *16*(2), 49–72. doi:10.1300/j001v16n02_04

Rodolfa, E. R., Bent, R. J., Eisman, E., Nelson, P. D., & Ritchie, P. (2005). A cube model for competency development: Implications for psychology educators and regulators. *Professional Psychology: Research and Practice*, *36*, 347–354. doi:10.1037/0735-7028.36.4.347

Schanche, E., Høstmark Nielsen, G., McCullough, L., Valen, J., & Mykletun, A. (2010). Training graduate students as raters in psychotherapy process research: Reliability of ratings with the Achievement of Therapeutic Objectives Scale (ATOS). *Nordic Psychology*, *62*(3), 4–20. doi:10.1027/1901-2276/a000013

Stamoulis, D. T., & Hauenstein, N. M. A. (1993). Rater training and rating accuracy: Training for dimensional accuracy versus training for ratee differentiation. *Journal of Applied Psychology*, *78*(6), 994–1003. doi:10.1037/0021-9010.78.6.994

Støre-Valen, J., Ryum, T., Pedersen, G. A., Pripp, A. H., Jose, P. E., & Karterud, S. (2015). Does a web-based feedback training program result in improved reliability in clinicians' ratings of the Global Assessment of Functioning (GAF) scale? *Psychological Assessment*, *27*(3), 865–873. doi:10.1037/pas0000086

Thornton, G. C., & Zorich, S. (1980). Training to improve observer accuracy. *Journal of Applied Psychology*, *65*, 351–354. doi:10.1037/0021-9010.65.3.351

Tweed, A., Graber, R., & Wang, M. (2010). Assessing trainee clinical psychologists' clinical competence. *Psychology Learning and Teaching*, *9*(2), 50–60. doi:10.2304/plat.2010.9.2.50

Vinton, L., & Wilke, D. J. (2011). Leniency bias in evaluating clinical social work student interns. *Clinical Social Work Journal*, *39*(3), 288–295. doi:10.1007/s10615-009-0221-5

Wolf, K. (2015). Leniency and halo bias in industry-based assessments of student competencies: A critical, sector-based analysis. *Higher Education Research and Development*, *34*(5), 1045–1059. doi:10.1080/07294360.2015.1011096

Appendix A. Clinical Psychology Practicum Competencies Rating Scale (CΨPRS) Items

1. Counselling Competencies

Overall Rating Item

Demonstrates empathic understanding, application of basic counselling techniques, and collaborative goal formulation with clients.

Sub-Domain Items

- Applies basic counselling techniques appropriately including clarification, paraphrase and summarising responses.
- Forms and communicates an empathic understanding to clients, carers, and significant others.
- Formulates client goals in a collaborative manner.
- Demonstrates accurate empathy in complex situations where affect is covert, controlled or denied.

2. Clinical Assessment Competencies

Overall Rating Item

Performs adequate assessments in a time efficient and in a personally/socio-culturally sensitive manner, appropriately prioritises issues, and assesses risk.

Sub-Domain Items

- Demonstrates knowledge of psychopathology and diagnostic criteria for clients seen at the placement.
- Demonstrates a systematic and logical sequence of questioning during the clinical assessment interview.
- Skilful and efficient in conducting a clinical assessment, including a mental state examination.
- Undertakes clinical assessments in an interpersonally engaging and in a socio-culturally sensitive manner.

3. Case Conceptualisation Competencies

Overall Rating Item

Appropriately integrates information from multiple sources to inform appropriate case conceptualisations, diagnoses, and treatment plans.

Sub-Domain Items

- Makes appropriate use of diagnostic frameworks (e.g., *DSM5*) to arrive at correct diagnoses and differential diagnoses.
- Draws upon different psychological theories and approaches to derive a meaningful case conceptualisation.
- Integrates cultural knowledge into case conceptualisation.
- Integrates assessment and other information into realistic treatment plans.

4. Intervention Competencies

Overall Rating Item

Skilfully implements appropriate, empirically supported treatment interventions; monitors treatment progress and outcomes.

Sub-Domain Items

- Demonstrates knowledge of principles and procedures of relevant interventions
- Demonstrates effective application of theoretical knowledge of evidence-based treatment methods (e.g., Cognitive Behavioural Therapy [CBT], Interpersonal Psychotherapy [IPT], Motivational Interviewing [MI]).
- Implements interventions relevant to the needs of the client.
- Demonstrates flexibility and responsiveness in the application of treatments and/or in the implementation of manualised programs.
- Efficiently conducts evidence-based treatment approaches (e.g. CBT, IPT, MI). Fluently transitions between elements/techniques.
- Overcomes common difficulties in therapy through skilful interviewing to maintain therapy direction and progress.
- Uses appropriate measures to regularly monitor treatment progress and outcomes.

5. Ethical Attitude and Behaviour

Overall Rating Item

Demonstrates knowledge of ethical/professional codes, standards and guidelines, and commitment to their application. Maintains appropriate and respectful boundaries and seeks consultation on ethical issues.

Sub-Domain Items

- Demonstrates knowledge of ethical/professional codes, standards and guidelines.

- Recognises ethical and legal issues that arise across the range of professional activities, and demonstrates good discernment and judgment in these situations.
- Acknowledges the limits of one's competence and makes appropriate referrals when required.
- Demonstrates commitment to ethical practice across a range of clinical situations.

6. Scientist Practitioner Competencies

Overall Rating Item

Demonstrates knowledge of theoretical and research evidence related to diagnosis, assessment and intervention. Shows respect for scientific methods and empirical evidence and commitment to their application to clinical practice.

Sub-Domain Items

- Demonstrates knowledge of theoretical and research evidence related to assessment, diagnosis, case conceptualisation and treatment, and to intervention monitoring and evaluation of interventions.
- Demonstrates the ability to critically analyse and evaluate the empirical literature.
- Demonstrates respect for, and use of, the scientific method in clinical practice.
- Demonstrates systematic and habitual application of scientific principles (e.g., hypothesis test into assessment, diagnosis, case conceptualisation and treatment, and to intervention monitoring and evaluation of interventions.

7. Professionalism

Overall Rating Item

Demonstrates effective organisation and time management. Clear and professional expressive skills, professional dress and demeanour. Good interactional skills with colleagues and other professionals.

Sub-Domain Items

- Demonstrates responsibility and accountability, reliably and punctually attending client appointments and work-related activities.
- Demonstrates an organised, disciplined, and timely approach to maintaining case notes and records.
- Effectively prioritises competing tasks
- Demonstrates concern for the welfare of others including the profession, organisation and community, and shows respect for cultural values and diversity.
- Clearly and effectively communicates in verbal, non-verbal and written forms for a range of purposes.
- Conducts self professionally in dress and demeanour.
- Works collaboratively with colleagues across a range of disciplines.
- Copes professionally with disapproval and criticism, and works constructively towards resolution of interpersonal conflicts at work.
- Demonstrates progress in developing an integrated sense of self as a professional psychologist.

8. Psychological Testing Competencies

Overall Rating Item

Applies knowledge to correctly select, administer, score and interpret common psychometric tests, and to generate psychometric reports. Demonstrates knowledge of psychometric issues and testing theory.

Sub-Domain Items

- Correctly administers and score common/core psychological tests.
- Demonstrates knowledge of psychometric issues, testing theory, and bases of assessment methods.
- Interprets and integrates information in accordance with psychometric principles.
- Demonstrates ability to write psychological test reports that are clear, accurate, and tailored appropriately to the user.

9. Reflective Practice

Overall Rating Item

Demonstrates self-care, self-awareness and reflectivity reflection on own emotions, beliefs, values and behaviour and their effect on others. Appropriately self corrects.

Sub-Domain Items

- Demonstrates problem-solving ability, organised reasoning, intellectual curiosity and flexibility.
- Demonstrates affect tolerance, understanding of interpersonal conflict, tolerance of ambiguity and uncertainty.
- Demonstrates consideration of the way in which personal issues and concerns impact on one's professional practice.
- Effectively uses observation and feedback including supervision to hone reflection skills.
- Actively reflects on ways in which others' cross-cultural values and perspectives influence one's own responses and vice versa.
- Accurately assesses own strengths and weaknesses and level of competence and plans necessary learning to address gap.
- Demonstrates appropriate and timely care of personal health and wellbeing to ensure effective professional functioning.

10. Response to Supervision

Overall Rating Item

Demonstrates good preparation and collaboration within supervision, openness to and effective use of feedback.

Sub-Domain Items

- Demonstrates adequate preparation for supervision.
- Seeks and accepts supervisory input, including direction. Appropriately balances autonomy and dependency needs.

Appendix B. Clinical Psychology Practicum Competencies Rating Scale (CΨPRS) Stage Descriptors

Stages	Description of stages
Stage 1. Beginner	Knowledge, skills, attitude value and relationship competencies are yet to be developed or at an early stage of development, and are on par with trainees commencing training without any practicum experience. Frequent minor or major inadequacies may be apparent, including difficulty applying knowledge to practice, difficulty managing sessions or conducting specific tasks, or little awareness of process issues. In later placements, a Stage 1 rating indicates failure to demonstrate adequate competency, with more frequent or intensive supervision required than would be expected.
Stage 2.	Knowledge, skills, attitude value, and relationship competencies are developing, and while more basic competencies are demonstrated under some circumstances, they may be inconsistent or not generalised. More complex competencies may be absent. Minor inadequacies occur frequently, and major problems may occur occasionally, although insufficient to cause serious harm. In later placements, a Stage 2 rating may indicate a failure to demonstrate adequate competency in the domain or a requirement for additional supervision to ensure adequate performance.
Stage 3.	The trainee demonstrates a moderate repertoire of basic knowledge, skills, attitude value, and relationship competencies, which are generalised to a wide range of common contexts, with more complex competencies emerging. There is a growing independence and responsibility for their own practice, with only minor inadequacies occurring.
Stage 4. Competent	The trainee demonstrates a wide repertoire of basic to advanced knowledge, skills, attitude value, and relationship competencies applied across a wide range of contexts. Performance is consistent with competencies of a graduate who has just completed all requirements of their professional Master's degree. There is an appropriate level of independence and development of adequate professional identity.

Appendix C. Training Vignettes

Domain	Vignette	Calibration score (Max 5)
Counselling competencies	Trainee TA relates effectively with clients in commonly encountered situations, and this capability is developing in more complex cases. She/he maintains a comfortable, warm, and respectful demeanour with most client situations. She/he frequently demonstrates good reflective listening skills and makes appropriate emotional and meaningful responses that help validate client experiences and clarify client issues. She/he appropriately directs and guides client focus in most client situations but tends to become less effective when dealing with complex presentations, including client resistance.	3.71 (0.38)
Clinical assessment competencies	Trainee TB collects sensitive information and uses session time effectively in most cases. She/he integrates collected information into hypothesis, diagnosis, and case formulations for commonly encountered cases and is developing this skill with more unusual or difficult cases. She/he displays an awareness of incorporating socio-cultural factors into clinical assessments but is sometimes inconsistent in integrating this information. She/he is capable of conducting risk assessments and/or formulating risk management plans for standard cases but needs some assistance for complex cases (e.g., multiple diagnoses).	3.40 (0.54)
Intervention competencies	Trainee TC demonstrates the ability to conduct a few structured behavioural and cognitive techniques relatively well but has a limited repertoire of CBT skills. The trainee appears unable to move fluently from one technique to the other, making the session feel disjointed and significantly reducing the effectiveness of the strategies employed. She/he is often able to identify negative cognitions and makes attempts to pose Socratic questions, but these attempts are typically restricted to a variant of "what's the evidence for that?" Slow but modest progress is made during typical sessions with cooperative clients presenting with low levels of severity. With more difficult cases, progress is less obvious and may stall.	2.30 (0.39)
Ethical attitude and behaviour	Trainee TD generally follows most aspects of the relevant legal, professional, and cultural ethical guidelines. She/he recognises the relevant ethical issues in simple cases but occasionally has difficulties with more complex cases. She/he displays a developing awareness of one's own values and biases, including cultural biases. She/he displays the capacity to apply an appropriate problem-solving approach to ethical issues encountered, but these may be simplistic. She/he does not always recognise when it might be helpful to seek appropriate consultation and supervision in order to guide her/his ethical practice.	2.82 (0.62)

Professionalism	Trainee TE requires close supervision in order to ensure that workload responsibilities are being adequately met in a timely manner. She/he is able to communicate with other team members and respond to direct instructions. Some difficulties present in prioritising competing demands and being appropriately assertive within the team when needed. Minor instances of poor record keeping, poor case preparation, or unprofessional demeanour have occurred. Self-reflection and self-awareness are limited, leading to overly negative or positive self-evaluations. There are also some concerns about punctuality and the occasional insensitive comment when interacting with peers and professionals.	1.90 (0.42)

Appendix D. Expert Calibration Vignettes

Domain	Vignette	Calibration score (Max 5)
Case formulation	Trainee YA's attempts at case formulation are fairly simplistic and mostly derive from a menu-driven approach linking intervention strategies to symptoms/problems rather than from an approach based on an understanding of underlying principles and/or key processes. Consequently, K assesses and formulates appropriate simple interventions but demonstrates difficulty applying formulated interventions to the client's individual context or circumstances. She/he requires assistance to modify intervention plans as new information emerges. She/he requires assistance in the translation of formulations into a language the client will understand and is tentative in their client communication.	2.14 (0.33)
Intervention	Trainee YB has a modest repertoire of interactional/intervention skills that allow fair progress with clients presenting with low-to-moderate levels of severity/complexity. The trainee's performance during a typical session is patchy, being interspersed with competent performance of simple interventions and other segments evidencing limited or laboured progress. The therapist evidences difficulty to move efficiently from one therapy episode/technique to the other, reducing to some extent the effectiveness of the interventions. She/he displays awareness of process issues, including client resistance, and makes initial attempts to address underlying dynamics. However, the interventions are only of limited value as these efforts lack the incisiveness, sophistication, and fluency characteristic of more advanced trainees.	2.68 (0.49)
Psychological testing	Trainee YC is able to generate some hypotheses leading to appropriate test selection for straightforward cases, but needs direction for more complex presentations. She/he generally balances the need to follow standardised test administration procedures while maintaining rapport and managing the client. The trainee shows adequate knowledge of psychometrics and test theory and is able to interpret test scores and discrepancies with some assistance. Her/his ability to derive appropriate recommendations from test data are slightly limited, leading to the occasional neglect of some central issues. With some supervisory input, the trainee is capable of producing written reports that show sufficient structure, accuracy, and clarity.	2.92 (0.40)
Scientist-practitioner approach	Trainee YD demonstrates a commitment to bringing the scientific method to their work. The trainee generally consults the scientific literature and other relevant materials, such as tests or school reports, to assess and treat their clients. She/he regularly uses new information from clients to formulate and test hypotheses. The trainee usually makes attempts to systematically assess client progress and consider alternate hypotheses when treatment is not progressing. When the literature or research evidence is less clear, they have difficulty formulating a theory-informed strategy to devise an appropriate way forward, instead seeking direction from their supervisor.	3.22 (0.48)
Professionalism	Trainee D has minor problems with consistently discharging workload responsibilities in an effective and timely manner. Inexperience or limited skills in prioritising demands within or across different professional roles contribute to variable outcomes and/or work-related stress. Excessive time may be devoted to less important aspects of the job. Lack of confidence often leads to an alternating pattern of self-directed learning and requests for guidance and support. She/he works fairly well within a team, but demeanour and communication styles lack the authority and autonomy of a mature professional.	2.44 (0.61)