2013

Assessment of psychology competencies in field placements: Standardized vignettes reduce rater bias

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

John Bushnell
University of Wollongong, bushnell@uow.edu.au

Russell Blackman
University of Wollongong, russellb@uow.edu.au

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

Vida Bliokas
University of Wollongong, vida@uow.edu.au

See next page for additional authors

Publication Details

Gonsalvez, C. J., Bushnell, J., Blackman, R., Deane, F., Bliokas, V., Nicholson-Perry, K., Shires, A., Nasstasia, Y., Allan, C. & 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.

Assessment of psychology competencies in field placements: Standardized vignettes reduce rater bias

Abstract
Supervisors' ratings of psychology trainees' competence in field settings are a critical component of training assessment. There has been little systematic research regarding the validity of these assessments, but the available evidence suggests we have a problem! Supervisors' judgments may be affected by systemic biases that pose a serious threat to assessment credibility. The current study is part of a research collaboration among six universities that endeavors to develop and evaluate a new method the use of vignettes against outcomes derived from a conventional rating scale. Individual vignettes were designed and subjected to a rigorous process of peer-review and revisions, before final vignettes were assigned calibration scores by a group of experts. A catalogue of vignettes (n = 41) that represent various domains of competence across several developmental stages was compiled. University and field supervisors used the conventional rating scale and the vignette-matching procedure (VMP) to evaluate competencies at end-placement. Data from a pilot (n = 20) and a follow-up study (n = 57) suggest that compared with a conventional rating scale, the VMP reduced leniency and halo biases. The VMP has the potential to improve outcomes of competency assessments in field placements and merits further research and development.

Keywords
psychology, assessment, field, competencies, placements, bias, standardized, vignettes, reduce, rater

Disciplines
Education | Social and Behavioral Sciences

Publication Details
Gonsalvez, C. J., Bushnell, J., Blackman, R., Deane, F., Bliokas, V., Nicholson-Perry, K., Shires, A., Nasstasia, Y., Allan, C. & 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.

Authors
Craig J. Gonsalvez, John Bushnell, Russell Blackman, Frank Deane, Vida Bliokas, Kathryn Nicholson-Perry, Alice Shires, Yasmina Nasstasia, Christopher Allan, and Roslyn Knight

Assessment of Psychology Competencies in Field Placements:

Standardized Vignettes Reduce Rater Bias[1]

Authors: Craig J. Gonsalvez*[1], John Bushnell[2], Russell Blackman[1], Frank Deane[1], Vida Bliokas[3], Kathryn Nicholson-Perry[4], Alice Shires[5], Yasmina Nasstasia[6], Christopher Allan[1], and Roslyn Knight[7]

*Author for correspondence

Craig J. Gonsalvez

University of Wollongong

Northfields Avenue

Wollongong, NSW 2522

Australia

Email: craigg@uow.edu.au

Tel: 61+2+42213674

Fax: 61+2+42214163

Author Affiliations

1: School of Psychology and Illawarra Institute for Mental Health, University of Wollongong

2: Faculty of Medicine, University of Wollongong

3: Port Kembla Hospital, Illawarra Shoalhaven Local Health District,

4: School of Psychology, University of Western Sydney

5: School of Psychology, University of New South Wales

6: School of Psychology, University of Newcastle

7: School of Psychology, Macquarie University

Abstract

Supervisors' ratings of psychology trainees' competence in field settings are a critical component of training assessment. There has been little systematic research regarding the validity of these assessments, but the available evidence suggests we have a problem! Supervisors' judgments may be affected by systemic biases that pose a serious threat to assessment credibility. The current study is part of a research collaboration among six universities that endeavours to develop and evaluate a new method – the use of vignettes - against outcomes derived from a conventional rating scale. Individual vignettes were designed and subjected to a rigorous process of peer-review and revisions, before final vignettes were assigned calibration scores by a group of experts. A catalogue of vignettes (*N* = 41) that represent various domains of competence across several developmental stages was compiled. University and field supervisors used the conventional rating scale and the vignette-matching procedure (VMP) to evaluate competencies at end-placement. Data from a pilot (N = 20) and a follow-up study (N=57) suggest that compared to a conventional rating scale, the VMP reduced leniency and halo biases. The VMP has the potential to improve outcomes of competency assessments in field placements and merits further research and development.

Keywords: competency assessments, practicum assessment, field placement evaluations, internship assessment, assessment by vignettes, vignette-matching procedure, leniency bias, halo bias.

Assessment of Psychology Competencies in Field Placements:

Standardized Vignettes Reduce Rater Bias

Over the past decade, psychology, as a discipline, has been at the forefront of describing and assessing competencies (Roth & Pilling, 2008; Rubin, Bebeau, Leigh, Lichtenberg, Smith, et al., 2007). Competencies and benchmarks for professional psychology for each of the developmental stages have been defined and systematically organised (Fouad, Grus, Hatcher, Kaslow, Hutchings, et al., 2009). Competency assessments across several health professions have been reviewed leading to the formulation of guidelines for competency assessment (Kaslow, 2004; Kaslow, Rubin, Bebeau, Leigh, Lichtenberg, et al., 2007), and key challenges and solutions to competency measurement at both macro and micro levels have been described (Lichtenberg, Portnoy, Bebeau, Leigh, Nelson, et al., 2007). Finally, a professional tool-kit has been assembled that catalogues commonly used assessment instruments and discusses their uses, merits and demerits (Kaslow, Grus, Campbell, Fouad, Hatcher, et al., 2009). By any measure, these developments represent significant landmarks of progress. The competency paradigm has probably gained the most momentum in the area of clinical supervision (Falender & Shafranske, 2011). From its earliest stages, competency-based approaches have espoused a developmental framework positing that growth from beginner to competent practitioner would occur through several intermediate stages (Watkins, 1995). Developmental models have dominated supervision training and practice for decades and have recently been integrated within competency-based models (Falender, Cornish, Goodyear, Hatcher, Kaslow, et al., 2004; Gonsalvez, Oades, & Freestone, 2002). The commitment to a competency-based philosophy and pedagogy has implications for curriculum design, supervision methods and techniques, and should drive assessment tasks and procedures. It is also assumed that such an approach will yield

accountability, transparency, and demonstrated evidence of competency attainment during training and through the life-span of the professional. (Kaslow et al., 2007).

Measurement of Competence and Competencies

Reliable and accurate measurement of competence is fundamental to the effective implementation of the competency paradigm to psychology. "Quite obviously, the gold standard is to demonstrate competency, but to do so requires having an assessment process for competency in place." (Leigh et al., 2007, p. 463). In operational terms, this would entail the careful crafting for both individuals and training programs, developmentally appropriate assessment tasks that are consistent with competency requirements prescribed by professional societies and regulatory authorities such as accrediting and registration bodies. The rationale for effective measurement is obvious. Accurate measurement of a trainee's profile of strengths and needs across the diverse domains of competence and over time is of major importance for trainees and training institutions. Reliable measurement informs formative feedback for all trainees and helps protect the public by identification, remediation, or dismissal of trainees who lack competence. Thus, the reliable and valid assessment of competence and competencies is critical to the effectiveness of the competency approach, and serious inadequacies would constitute a significant barrier to Psychology's pursuit of a competency-based paradigm. Unfortunately, Psychology lags behind other disciplines such as medicine in terms of the breadth and frequency with which summative assessments are used, both during training and during post-qualification professional development (Leigh, Smith, Bebeau, Lichtenberg, Nelson, et al., 2007; Townsend, McIlvenny, Miller, & Dunn, 2001). Popular assessment tasks such as multiple-choice, short-answer, and essay assignments may be appropriate tasks to measure knowledge competence but be incapable of capturing progress in skills, relationship, and attitude-value competencies (Lichtenberg et al., 2007;

Pachana et al., 2011). This deficiency is probably most pronounced in the assessment of competencies in field placements.

Assessment of Competencies in Field Placement and Internships.

In terms of time and resource investments per trainee, clinical supervision is the most expensive component of professional training within several sub-disciplines in psychology (Gonsalvez & Milne, 2010). Most of this supervision occurs in a one-to-one context in university clinics, external field placements/externships/rotations, or/and in year-long internships (terms differ across countries and disciplines; the broader term *'placement'* will be used in the current manuscript). Although a variety of formative and summative assessment methods are used within psychology, competency evaluation rating forms (CERF) are used widely by training institutions supervisors in many countries to rate competencies at end-placement (Baird, 2005; Gonsalvez & Freestone, 2007; Kaslow et al., 2009; Tweed, Graber & Wang, 2010). CERFs are easy to use, inexpensive, and are sufficiently versatile to measure a range of global and specific competencies. Evaluations by field supervisors are credible because of their professional qualifications and practice-expertise. Moreover, because field supervisors have access to direct observation of trainee performance across a wide variety of real-life situations over a fairly extended period of time (Gonsalvez & Freestone, 2007), their judgments have high ecological validity and merit serious consideration by training institutions. A survey indicated that Directors of training ranked internship supervisors' evaluations of trainees as first among 36 other quality assurance measures of professional training (Norcross, Stevenson & Nash, 1986). What is inconsistent and of concern, is that this large and fairly long-term investment in practicum-based training is not supported by evidence that assessments of competence are credible.

On the contrary, there is a significant body of evidence indicating that a number of biases operate. The largest study we are aware of reported results from 291 end-placement

reports attained by 131 clinical psychology trainees evaluated by 130 field and university supervisors over a 12-year period (Gonsalvez & Freestone, 2007). Field supervisors consistently avoided assigning average and below average grades. The same cohort of students obtained higher grades for their practicum from field supervisors than they did for their coursework. Moreover, between-domain ratings for the same supervisor produced very high correlations, but supervisors' ratings at completion of a placement poorly predicted ratings trainees attained during a subsequent placement. These results were interpreted as suggesting systematic halo and leniency biases (Gonsalvez & Freestone, 2007). The concern about leniency and halo rating biases affecting field supervisor judgements is supported by other research within psychology (Borders & Fong, 1991; Dienst & Armstrong, 1988; Robiner, Saltzman, Hoberman, Semrud-Clikeman & Schirvar, 1997), and from other health disciplines including social work (Bogo, Regehr, Hughes, Power & Globerman, 2002; Bogo, Regehr, Hughes, Power, Woodford, et al., 2004; Lazar & Mosek, 1993), medicine (Williams, Klamen & McGaghie, 2003) and nursing (Chambers, 1998; Dolan, 2003). Further, a survey of internship supervisors found most supervisors (58%) believed that competence ratings made by themselves and by their peer supervisors were biased, compared to smaller numbers who indicated they were unsure (22%) or believed that ratings were not biased (10%). Leniency and central tendency biases were identified as being most prevalent (Robiner et al., 1997).

Inaccurate ratings, particularly those which are too lenient, reduce opportunities for trainees to develop their skills and ultimately, may erode public confidence if practitioners are being credentialed without appropriate attainment of competence. The leniency bias could foster inflated self perceptions and prevent necessary and appropriate remediation strategies. Robiner et al., (1997) concluded, "It may not be an exaggeration to consider the existence and extent of supervisory bias to be the most critical quality assurance issue confronting clinical

psychology…." (p. 62). Regretfully, fifteen years later, there is little by way of progress to report. Several factors are likely to have contributed to this lack of progress: the intuitive appeal and face validity of supervisor's judgments, the absence of a theoretical framework to conceptualise the multitude of practitioner skills and their assessment, and the lack of validated instruments to provide more credible and pragmatic assessment alternatives (Lichtenberg et al., 2007).

Biased Rater or Inappropriate Instrument?

It is possible that the observed inadequacies of CERF-type ratings are a consequence of a poorly developed instrument rather than the effect of rater bias. Although instruments to assess the fidelity of specific therapeutic approaches are available (e.g., the Cognitive Therapy Rating Scale; Blackburn, James, Milne, Baker, Standart, et al., 2001), the use of such scales is time consuming, requires training, and is often insufficient to cover the range of relevant competencies. There have been no measures with established psychometric properties that assess broader trainee competencies (Baird, 2005; Gonsalvez & Freestone, 2007). However, several attempts to overcome scale inadequacies by improvements in rating technologies have proved unsuccessful. For example, an attempt to persuade supervisors to assign lower scores by changing a 5-point scale to a 6-point scale failed to remedy the leniency bias (Gonsalvez & Freestone, 2007). A further source of bias could stem from demands on field supervisors to rate a trainee's competence "in reference to performance of trainee peers". This is problematic when the rater has no normative reference-point against which to anchor their rating. An endeavour to remedy this was undertaken in a collaborative study across five universities in Australia. The requirement to rate a trainee's performance based on "peer performance" (a relative anchor) was substituted by having supervisors' rate competence against a notional absolute anchor – readiness to practice. This and other modifications (e.g., providing better operational definitions of anchor points) had little effect

and supervisor ratings continued to manifest marked halo and leniency biases (Bushnell, Nicholson, Blackman, Allan, Nasstasia et al, 2011). Finally, after several attempts to improve field supervisor ratings, researchers in social work concluded, "trying to improve field evaluation scales may be the academic equivalent of rearranging the deck chairs on the Titanic." (Regehr, Bogo, Regehr, & Power, 2007, p. 338). In effect, it is possible that the use of a Likert-type rating is itself the source of at least some of the observed problems, with the rating continuum providing a framework that fosters rather than mitigates against rater biases. The effectiveness of alternative instruments to assess competencies needs to be examined.

Use of Vignettes

Concerns about field supervisor rating biases in social work have led an influential group of researchers to discard rating scale-based instruments and trial the use of vignettes (Bogo et al., 2002, 2004). They designed a catalogue of 20 vignettes that offered descriptors of a trainee functioning at different levels. The vignettes were intended to capture typical competency profiles attained by students across diverse competency levels. At end-placement, field supervisors were required to read all 20 vignettes and to pick out all vignettes that matched their trainee's performance. The chosen vignettes were then scrutinised before supervisors chose one or two vignettes that best matched the trainee's competencies. This method resulted in a broader distribution across performance levels. Supervisors who were unwilling to assign low ratings to trainees on the conventional rating scale were willing to match the same trainees to vignettes that represented poor performance levels.

The work of Bogo and associates (2002, 2004) is pioneering but the "prototype" model has potential limitations, especially if such an approach is expected to yield reliable and valid ratings across the spectrum of competency domains. First, the prototype approach is built on the assumption that the variability among trainee performance can be effectively

captured by a relatively limited number (20) of exemplars. A more serious concern is that this approach is theoretically inconsistent with most competency-based approaches that assumes a relative independence for the domains posited (see Fouad et al., 2009). This allows for the possibility that a person who scores high on one domain (e.g., clinical assessment skills) may obtain a range of scores (from low to high) on another (e.g., ethical practice). These assumptions are yet to be empirically tested nevertheless, the adoption of a model that preserves the independence of the domains appears warranted at this early stage of competency assessment.

The Present Study

The current study has two main aims: the description of the development and standardization procedures for a catalogue of vignettes designed to assess clinical psychology competencies at the end of field placements, and the pilot testing of these vignettes through the comparison of outcomes derived from the new, vignette procedure and a conventional rating scale.

Method

Participants

Field supervisors were recruited from six clinical psychology training programmes offered by six universities that were fully accredited by the Australian Psychologists Registration Board and the Australian Psychological Society. Field supervisors satisfied University requirements for clinical qualifications (had a Clinical Masters degree, a Clinical Doctorate degree or a PhD in Clinical Psychology), had experience as clinical psychologists, and were or were eligible to be full members of the APS College of Clinical Psychologists. In accordance with university and APS requirements, summative evaluations were completed by field supervisors at mid- and at end-placement. The format of the mid-placement evaluation

varied across the participating universities, but a uniform rating scale, described below was used for end-placement ratings.

The trainees were students who gained entry into a Masters or Doctoral clinical program of one of the participating universities. They had completed a minimum of four years of full-time psychology training at the undergraduate level and had diverse levels of experience ranging from no professional experience to several years of experience as a generalist psychologist. As part of their clinical training, trainees completed intensive coursework at their respective universities and concurrently enrolled in 3 or more field placements during a two-year period. The initial placement was usually in the university's psychology clinic, and subsequent placements occurred in external agencies. Each placement included between 200-300 placement hours, including a minimum of 80-100 hours of face-to-face client contact. Placement experiences varied with most occurring 2 or 3 days per week.

*Clinical Psychology Practicum Competencies Rating Scale (*CΨPRS*)*

The CΨPRS, a CERF-type rating scale, was used for field supervisor ratings of students at end-placement. The CΨPRS consists of 69-items (60 individual items + 9 items assessing overall domain performance) that covers 8 broad domains of clinical competence (see Table 1 for domains) and one additional domain that evaluates 'Rate of Progress and Response to Supervision During Placement'. The CΨPRS was developed from earlier versions of similar scales used by the participating universities, and the list of practicum competencies identified by Hatcher and Lassiter (2007). The CΨPRS is based on a developmental model incorporating 4 stages of development from Stage 1 (Beginner) through to Stage 4 (Competent; Stages 2 and 3 were left unlabelled). Supervisors were required to rate students in reference to a notional absolute anchor (readiness for competent practice), using a visual analogue scale that ranged from zero (Stage 1/Beginner) to ten (Stage 4/Competent),

with intermediate anchors being Stage 2 and Stage 3 (See Table 1). Supervisors completed the CΨPRS online at the end of each placement, separately for each item for each student, with the domains presented in a fixed sequence, one domain at a time. The Rate of Progress and Response to Supervision During Placement domain was always completed last.

Insert Table 1 about here

Development of Vignettes

The domains and developmental stages adopted by the CΨPRS (see Table 1) were used as a template to generate vignettes for the vignette matching instrument, yielding a matrix of 32 cells for the clinical domains (8 clinical domains x 4 stages). In addition, for the Intervention domain, parallel vignettes for CBT and psychodynamic therapies were generated. For the progress domain, five types of progress during placement were demarcated to depict unsatisfactory, slow, inconsistent, developing well, and excellent progress. Hence, the overall task entailed designing a catalogue of 41 finalised vignettes across the clinical and progress domains. The catalogue of vignettes was developed in two phases. The first phase involved the development of 25 vignettes (Domains 1 to 3, 4b, 5 and 9), and the second phase involved the development of the remaining 16 vignettes (Domains 4a, 6, 7, 8). Procedural details for Steps 1 and 4 remained identical for the two phases. Minor procedural variations were implemented for Steps 2 and 3 in Phase 2. The procedures employed in Phase 2 proved more efficient and are therefore reported below and recommended for any attempt at replication (See Figure 1).

Insert Figure 1 about here

Step 1: Development of Vignettes, Version 1(V1). In the initial step, each of the domains (see Table 1) was assigned to two of the six Clinical Research Investigators. All six of the Clinical Research Investigators were employed as university Psychology Clinic

directors. Together, the Clinical Research Investigators drafted 82 vignettes, 2 vignettes per cell, to ensure parallel sets of vignettes for all domains. They were instructed: to use the items within the CΨPRS domains as a generic guide to identify aspects of a competency, to ensure that the vignette captured key aspects of the competency rather than attempting to include all aspects, to attempt to titrate the competencies depicted in each vignette in terms of the four developmental stages of competency attainment (see Table 1 for descriptions of the 4 developmental stages), and to restrict word length of each vignette to about 100 words (see Appendix for a sample vignette). Hence, the set of 4 vignettes for each clinical domain was required to capture the step-wise progression towards competence within the specific domain. For each domain, the vignette authors were asked to use the 0-to-10 continuum adopted by the CΨPRS to anchor the 4 vignettes in an attempt to ensure that the four vignettes fell within the following bands of the visual analogue scale (Vignette 1: scores 1-2; Vignette 2: scores 3-5; Vignette 3: 6-8; Vignette 4: 9-10).

Steps 2 & 3: Development of Vignettes, Versions 2 (V2) and 3 (V3). Steps 2 and 3 involved the recruitment of two external experts with domain-specific expertise (e.g., psychometrics, professional ethics) to help develop the revised Version 2 vignettes (V2-vignettes, see Figure 1). A vignette development subcommittee was formed and comprised of the two external experts and two other content experts on the research team. One subcommittee member took primary responsibility for authorship of a domain. This primary subcommittee member examined the two V1-vignettes for each cell (8 vignettes for each clinical competency domain) and generated a single revised vignette for each cell (referred to as V2 vignettes in Figure 1). Instructions for forming the V2-vignettes were similar to those provided in the development of the V1 vignettes and included the specification that the set of revised vignettes for each domain should fall within the four bands of competence.

In Step 3, V2-vignettes were subjected to peer review and commentary. Two members of the vignette development subcommittee independently and anonymously provided revisions to the V2-vignettes using track-changes in a word-processing program (see Figure 1)..

Step 4. Finalisation of V3-Vignettes. In Step 4, the principal investigator of the project and the subcommittee member who took primary responsibility for vignette development in Step 2 jointly finalised the text for each vignette. This involved incorporating suggested peer revisions, as deemed necessary.

Vignette Matching Procedure (VMP): Pilot Study 1

Because the development of vignettes happened in two phases, only 25 vignettes (Domains 1,2,3, 4a, 5, and 9) were available for this study. Twenty field supervisors who had completed end-placement CΨPRS ratings for 20 trainees during a previous month participated in Study 1. For each domain, supervisors were presented the set of four vignettes (five for the Progress domain) concurrently, before being required to pick one of the vignettes that best matched the trainee's performance. Trainees obtained the developmental scores (Stage 1 to Stage 4) associated with the specific vignette. No attempt was made to match CΨPRS and vignette ratings, and supervisors were instructed that vignette ratings would have no bearing on trainee assessments. The results from the Study 1 are presented in Table 2. Following encouraging results from Study 1, the research progressed to Phase 2 that included the development of the remaining vignettes, vignette calibration and an additional field study.

Calibration of V3-vignettes

The initial expectation was that there would be good consensus among the clinical research investigators with regard to where on the 0-10 point visual analogue scale, each of the V3-vignettes should be anchored. However, pilot testing suggested that there was less than optimal agreement. For example, there were several instances when between-

investigator ratings of the same vignette varied by 2 scale points (on the 10-point scale). Two options were considered: the vignettes could be assigned anchor scores by the vignette committee or they could be 'normed' by a larger group. The latter option was favoured and consequently, the vignettes were subjected to a calibration test. A group of University Psychology Clinic Directors from Australia or New Zealand was used as the criterion group to calibrate the vignettes. University Psychology Clinic Directors are qualified clinical psychologists holding full membership in the Clinical College of the Australian Psychological Society. They design and coordinate clinical services at the university psychology clinics, provide supervision to trainees undertaking placements within the clinic, coordinate clinical placements and supervision of trainees during their external placements (externships), and have oversight of the assessment of trainee practicum competencies both within the university clinic and at externships. The group's expertise in clinical psychology practice, supervision, and their familiarity with placement activities and placement evaluation qualified the Clinic Directors to serve within the criterion group. The Clinic Directors were contacted by email and 15 agreed to participate. A computerised online platform was used to present the V3-vignettes. Vignettes were presented individually in random sequence, and the criterion experts were asked to identify the domain (from a list of 9 domains) represented by the presented vignette, and to indicate the point on the visual analogue scale ranging from Beginner (0) to Competent (10) where the vignette best fitted. Criterion experts completed their task independently and anonymously and received a $30 entertainment/book voucher as part compensation for their time. To examine reliability among the criterion experts, intraclass correlation coefficients were determined for each of the domains. A two-way random approach was adopted using the 'absolute agreement' strategy.

Vignette Matching Procedure (VMP): Study 2.

Concurrent with the calibration exercise, the V3-vignettes were subjected to a field trial. All field supervisors who submitted end-placement evaluations for trainees were invited to participate in this study by submitting the mandatory CΨPRS ratings and by completing the optional VMP immediately after the CΨPRS. Data were obtained for 57 trainees (response rate of 36%) across the group of participating universities during a 5-month period. Of the 57 trainees, 30 had ratings following completion of their first placement (labelled Novices in the current study), 16 had ratings following completion of their 2nd or 3rd placement (labelled Advanced Beginners in the current study), 10 had ratings following completion of their 4th placement, and 1 person had missing data. A computer-based program presented the vignettes in sequence from the first through to the last domain. Within each domain, vignettes were presented in either ascending or descending order, one vignette at a time. A random program determined whether the series commenced with an ascending or descending order, with the two orders alternating between domains. Supervisors reviewed each vignette and made a judgment as to whether the profile of competencies demonstrated by the trainee was at a level higher than, equal to, or lower than the developmental profile captured by the vignette. The series within the domain terminated when the trainee's competence level was identified (e.g., when the trainee was identified as possessing competencies higher than vignette 2 but lower than vignette 3). Thus, not all vignettes within a domain were necessarily presented for each trainee.

Supervisors were instructed that the VMP was in the experimental stage so their scores based on the vignettes would have no bearing on the summative assessments of the trainees. Following the use of the VMP, supervisors completed a 4-item evaluation about the face-validity and utility of the VMP. Following completion of the task, supervisors had the option of claiming a $30 book or movie voucher as token compensation for their research participation. Completion of the vignette procedure by supervisors took about 35 minutes.

CΨPRS and VMP Competency Scores. For the analyses, mean competency scores (across domain items) for each student were computed for each of the domains of the CΨPRS. Because each vignette was calibrated by the criterion group, a profile of competency scores across domains could be computed for each trainee based on the calibration score of the vignette to which the trainee was matched. For instance, a trainee matched to the Stage 2 vignette of Domain 1 (Relational skills), received a competence score of 3.27, whereas a trainee matched to Stage 3 received a competence score of 6.29. In instances where competence was rated higher than Stage 2 but lower than Stage 3, trainees were assigned a score mid-way between the two calibration scores (e.g., 4.78 for Domain 1).

All stages of the research were approved by the University of Wollongong's Human Research Ethics Committee and ratified by Ethics Committees of all other participating universities.

Results

The results of Pilot Study 1 are presented in Table 2.

Insert Table 2 about here

Vignette Calibration

Intraclass correlation coefficients (***r***) for the domains and the calibration scores assigned to the vignettes by the group of criterion experts are presented in Table 3. High intraclass correlations ($p < .001$) were found for each of the domains. The vignettes were also required to meet four criteria. All vignettes met Criterion 1, (Accurate identification of the domain represented by the vignette by 95% or more of experts); 38 of 41 vignettes met Criterion 2 (mean calibration scores fell within designated bands); 34 of 41 vignettes met Criterion 3 (calibration score standard deviations did not exceed 1.5); and 37 of 41 vignettes met Criterion 4 (difference between mean scores of adjacent vignettes within a domain did

not exceed 4.0 units). Eleven of the 41 vignettes (27%) violated one or more criteria and are currently undergoing revision.

Insert Table 3 about here

Vignette Matching Procedure: Study 2

The results from Study 2 are presented in Table 4. To compare whether the two instruments (CΨPRS and VMP) yielded different distributions, the data (frequencies) were subjected to a log linear model analyses for the 2 Instruments X 9 domains X 4 Stages. The results indicate significant interactions for Instrument X Stage, χ^2 (6) = 187.65, p<.001, for Instrument X Domain, χ^2 (16) = 74.29, p<.001, and for Domain X Stage, χ^2 (24) = 152.88, p<.001. In effect, the results indicate that the VMP yielded a wider distribution and lower scores (higher frequencies in Stage 2 and lower frequencies in Stage 4) than did the CΨPRS, with VMP-CΨPRS differences being more pronounced on some domains than on others. Across domains, almost all (99.8%) supervisor ratings on the CΨPRS fell within Stage 3 (around 25%) and Stage 4 (75%) performance bands with less than 1% of ratings falling within Stage 2 (0.20%) or Stage 1 (0%). In contrast, on the VMP, across all domains, 7.6% of trainees were judged to have skills within Stage 2, with this percentage varying across domains, from a low of 2% of trainees obtaining Stage 2 scores for Relational Skills, Ethical Practice, & Response to Supervision, to a high 26% of trainees receiving Stage 2 scores for clinical assessment skills. Further, about 10% of trainees were judged to be at Stage 1 for Psychometric skills. In addition, as might be expected, across both instruments, frequencies varied among stages with larger numbers of trainees placed in Stages 3 and 4, with these differences being more pronounced for some domains.

To compare whether the two instruments differentiated between earlier and later placements, the cohort was divided into two groups based on whether their competencies

were assessed following the first placement (Novice, n =30) or following the 2nd or 3rd placement (Advanced Beginner, n = 16). These data were subjected to a mixed ANOVA for Groups (Novice, Advanced Beginner) x Instruments (CΨPRS, VMP) x Domains with repeated measures for the Domain factor. For Domain, planned contrasts were conducted comparing each domain score with the mean across 9 domains. The results are presented in graphic form in Figure 2.

Insert Figure 2 about here

Main effects for Group and Instrument were each significant indicating higher competency scores for Advanced Beginners, $F(1,44) = 17.36$, $p < .001$), and higher scores for the CΨPRS instrument, $F(1,44) = 39.90$, $p < .001$). The leniency effect associated with the CΨPRS was qualified by two- and three-way interactions. Specifically, differences between the instruments were more marked for the Clinical Formulation, $F(1,44) = 17.87$, $p < .001$, Ethical Practice $F(1,44) = 30.72$, $p < .001$, and the Scientist-Practitioner domains, $F(1,22) = 4.14$, $p = .05$. These results were further qualified by a Group interaction that approached significance, with the results showing that the CΨPRS better differentiated the Novice and Advanced Beginner groups on the Clinical Formulation domain $F(1,44) = 3.85$, $p = .06$, whereas the vignette instrument better differentiated the groups on the Ethics domain $F(1,44) = 3.46$, $p = .07$.

Across instrument and groups, and compared with mean domain scores, trainees obtained higher scores for Ethical Practice $F(1,44) = 30.72$, $p = .001$, and Professional Skills, $F(1,44) = 10.40$, $p < .01$. They obtained lower competency scores on Clinical Assessment, $F(1,44) = 4.92$, $p < .05$; Clinical Formulation, $F(1,44) = 17.87$, $p < .001$; and Psychometric skills, $F(1,22) = 6.08$, $p < .05$.

To examine halo biases, correlations matrices for the 9 domains were compared between the two evaluation methods (Table 5).

Insert Table 5 about here

For the CΨPRS, 35 of 36 between-domain correlation coefficients were higher (ranging from 0.61 to 0.95; $M = 0.79$), than corresponding correlations observed for the VMP (range from 0.01 to 0.81; $M = 0.43$). Such a pattern is extremely improbable (binomial theorem, $p<.0001$). The differences between correlation coefficients derived from the CΨPRS and VMP instruments were also tested by the Fisher's *z*-test (Howell, 2007). Of 36 pair-wise comparisons, 33 were also statistically significant ($p<.05$ or lower; Table 5). Finally, supervisors' opinions about the validity and utility of the VMP in comparison with the CΨPRS were also obtained (Figure 3). The results suggest a positive endorsement of the VMP, with the procedure being rated as being more valid, easier to use, being more capable of capturing trainee performance and less time consuming.

Insert Figure 3 about here

Discussion

The current study describes a multi-site endeavour to design and standardize a catalogue of 41 vignettes to assess clinical psychology competencies in field placements. Results derived from the VMP were then compared to results from a conventional rating method, the CΨPRS. The VMP yielded superior results as discussed below.

Vignette Matching Procedure vs. CΨPRS

A comparison of the two instruments indicated that the VMP yielded distributions that were different from those obtained by the CΨPRS in both pilot studies. Although we did not have access to matched CΨPRS and vignette ratings for the small sample of students reported

in Study 1, we had access to data from the year's cohort of students and reported elsewhere (Bushnell et al., 2011). The distributions derived from the CΨPRS for 2012 (Study 2) replicated the 2011 CΨPRS results (Bushnell et al., 2011). In each of the pilot studies, the VMP yielded distributions that were different from those obtained by the CΨPRS. Whilst supervisors ignored the lower categories (Stages 1 and 2) on the CΨPRS, they were willing to match at least a small percentage of trainees to lower developmental stages depicted by vignettes. Conversely, a large percentage (75% in Study 2) of CΨPRS domain ratings and a much smaller percentage (45% in Study 2) of VMP ratings fell within the competent band (Stage 4).

These CΨPRS-VMP rating differences are interpreted as suggesting a leniency bias affecting field supervisor ratings on the CΨPRS for the following reasons. First, a significant proportion of trainees (65%) rated in Study 2 had completed no more than a single placement (200-300 hours of placement activity) and would be considered novices by training institutions. Mean competence ratings for these students are much higher than expected (Mean scores of 8 and above on a 10-point scale, Figure 2), with this pattern being apparent across most domains. Further, competence ratings of the advanced beginner group also look inflated, with mean ratings across domains falling within the competent band. Secondly, if field-supervisor ratings were taken at face value, it could be argued that the additional training requirements in terms of client and supervision hours prescribed by the accrediting bodies are excessive and unnecessary for competent practice. Finally, the interpretation favouring a leniency effect is consistent with results from a growing body of research that examined field supervisor ratings within psychology (Borders & Fong, 1991; Gonsalvez & Freestone, 2007; Robiner et al., 1997) and other health disciplines (Bogo et al., 2002, 2004). Therefore, the current results emphasise concerns that field placement assessments using

CERF-type rating scales may be systematically skewed. Whereas previous studies reported long-term results from a single training institution (e.g., Gonsalvez & Freestone, 2007), the current study demonstrates that a similar pattern was observed across several training institutions, and despite several improvements made to the rating instrument to enhance outcomes. Thus, the leniency bias may be more pervasive and persistent than previous studies might indicate. Whilst the vignette instrument produced lower competency scores across all domains (Fig. 2), this pattern was somewhat more pronounced on clinical assessment, formulation, and scientist-practitioner skills. These results might reflect that the vignette instrument is more sensitive at discriminating developmental stages on some domains. Alternatively, the results might represent differences between criterion experts and field supervisors in terms of the way they believed competence should be defined for these specific domains.

The pattern of correlations observed among domains within the same instrument and for the same domains (diagonal values, Table 5) between the two instruments suggests that halo biases affect the CΨPRS ratings. First, the magnitude of the correlations for the CΨPRS are very high (above 0.80), and within-domain correlations for the CΨPRS are higher than same domain correlations between the two instruments (Table 5, diagonal values). In contrast, between-domain correlations for the vignette instrument are much lower, more variable, and more consistent with the assumption that domain ratings would demonstrate a moderate level of independence. For instance, on the VMP, the low correlation (r = .01) between Relational Skills (Domain 1) and Ethical Practice (Domain 5) is consistent with the notion that competence in a skills area may not reflect attitudes and adherence to ethical codes. Conversely, Ethical Practice shares common features with and correlates more highly with Professional Skills (r = 0.67). The suggestion that halo biases may systematically affect practicum ratings is consistent with observations from previous research that found high

correlations among domains rated by the same supervisor and low between-supervisor ratings (Gonsalvez & Freestone, 2007).

Overall, an encouraging result was that the vignette procedure yielded data suggestive of reduced leniency and halo effects. The vignette's ability to provide a descriptive and concrete comparison may underpin its ability to help reduce rater bias. Specifically, the vignette forces the rater to make a comparison of the student's competency profile against a contoured portrait depicted in the vignette rather than to a hypothetical, subjective, and poorly defined notion of competence in the rater's mind. Thus, the concrete and featured characteristics of the vignette may make it less vulnerable to distortion.However, the results failed to show that the vignettes better differentiated novice and advanced beginner trainees. The CΨPRS better differentiated the two groups on the formulation domain whereas the VMP better differentiated the groups on the Ethical Practice domain. The reduced statistical power associated with small sample sizes may have contributed to these 'negative' results. Also, a longer series of standardised vignettes within each domain (e.g., 5 or 6) may help enhance discrimination between developmental stages.

Finally, field supervisors who trialled the VMP gave it a positive endorsement. Compared to the CΨPRS, the VMP was evaluated by supervisors as having better face-validity, as being easier to use, as better capturing trainee competencies, and as not requiring more time to complete (Fig. 3). The likely reasons for leniency and halo effects affecting rating scales like the CΨPRS have been discussed in the literature and include supervisor-supervisee relationship affecting ratings, supervisor role conflicts between switching from formative-supportive supervisory interventions to summative-assessment roles, supervisor perceptions that low ratings may reflect negatively on their supervisory capabilities, and additional demands on time and energy to justify low ratings (Bogo et al., 2002; 2004 ; Borders & Fong, 1991; Gonsalvez & Freestone, 2008; Robiner et al., 1997).

Because the superiority of the VMP over CERF ratings has to be replicated in future studies with larger samples, the potential reasons offered for its merits are no more than tentative. It is possible that adequately standardized vignettes provide supervisors with a rich profile of features that serve to reduce ambiguity and help establish better defined mile stones along the developmental continuum towards competence. It is also possible that matching student competencies to vignettes is more akin to a ranking process (in which a student is compared with a vignette) that is less vulnerable to rating biases than Likert-based rating instruments.

Vignette Development and Calibration

The crafting of vignettes proved to be a complex and painstaking process. The task was made more difficult because the project had to navigate across uncharted territory, with limited assistance from the scientific literature to help determine which methods would yield the best results. A process that yielded satisfactory outcomes combined three elements -- expert-author, informed peer-review, and a person/panel that performed the role of an editor. Notably, this process is not unlike the editorial process to which new manuscripts are subjected.

The high correlations among the group of criterion experts for each of the domains are reassuring. However, these correlations may not be optimally sensitive to leniency-stringency effects that uniformly affect ratings of all vignettes within a domain, and should be viewed within the context that a moderate level of variability around the mean was observed even among the criterion experts (see SDs in Table 3). The initial expectation that there would be much better concordance between-experts (in terms of where, on a 10-point visual analogue scale, a designed vignette was best anchored) proved unrealistic. Consequently, the recruitment of a larger number of experts to establish ‘calibration’ scores for each vignettes was required. Most (73%) vignettes satisfied the apriori criteria set by the subcommittee to

determine adequacy for vignettes (See Table 3), indicating that the process adopted by the current approach was reasonably successful.

The capability of our approach to generate normative calibration scores for each vignette, along with a measure of its variability, is an important advantage over previous approaches that have used vignettes. When normed by a valid criterion group, vignette calibration scores serve as reference points against which an individual's competence may be graded. Thus, for an individual, calibration scores provide a set of relatively stable anchors that help mark progress (or lack thereof) over time. At a macro level, calibrated vignettes provide a framework to benchmark performance of cohorts of trainees across institutions and across time. Further, because sets of vignettes could be normed separately for different contexts or countries in the same way as intelligence tests, calibration scores give the VMP versatility and impact. The decision to not adopt the prototype model for the vignettes has another important advantage. The independence of the various competence-domains has been preserved and the method is capable of providing a cross-domain profile of developmental needs and strengths to fulfil both formative and summative ends. The key drawback is the major investment of resources in the arduous standardisation process. Given that training in field settings is a central aspect of psychology education and training, and that the competency paradigm is only as strong as the reliability of the metric that measures competence, such an investment appears justified.

Limitations and Conclusions

Although it is reassuring that the results of both the pilot and the follow-up study are consistent, the distributions associated with the VMP in both studies are derived from small numbers (N=20 and N=57), limiting the confidence with which these results can be generalized. Also, based on predetermined criteria, 11 of 41 vignettes used for the current study violated one or more of the four criteria for an acceptable vignette. It could be argued

that the unsatisfactory vignettes compromised the results of the study. This concern is allayed by the fact that VMP yielded results that were better than the conventional rating scale method, despite the inadequacies of some vignettes. Nevertheless, a follow-up study that replicates and extends these results with a full set of standardised vignettes would be of value. Because the validity of the VMP was yet to be established, supervisors were informed that judgements based on the VMP would not be incorporated into their summative assessments. It is therefore possible that observed differences between the CΨPRS and the VMP reflect differences in supervisor readiness to report less-positive judgments in formal versus informal contexts. In other words, the gains associated with the VMP may disappear when the VMP is employed as a summative assessment device. Although the endorsement of the vignette approach by supervisors is reassuring, this possibility should be examined in future research. In summary, preliminary results from the current program of research are encouraging and indicate that the VMP has the potential to reduce leniency and halo biases that affect conventional Likert-style instruments.

Implications for Practice

First, there is an urgent need for innovative assessment methods to assess practice-based competencies in psychology. The results from the use of the VMP are encouraging in terms of better differentiating levels of competence. The method's capability to provide a matrix of relatively stable anchor points across diverse domains and developmental levels against which competencies can be judged, has major implications. These implications include idiographic merits that enable the monitoring and tracking of an individual's attainment of competencies, and normative applications that help benchmark outcomes across cohorts and training programs. For instance, standardised vignettes could be used to help compare the developmental trajectory of trainees across the various foundational and functional competency domains (Fouad et al., 2009). It will also be worthwhile to determine

whether vignette scores derived from a previous placement could predict student performance on subsequent placements. Should the method prove successful in psychology, it will also have valuable cross-disciplinary potential. Second, the current study adds to the growing body of evidence that CERF-type ratings by university and field supervisors are affected by systematic leniency and halo biases. The obvious implication is that the justification for sole reliance on these ratings is weak.. Consistent with best-practice guidelines, assessment of practicum competencies must be multitrait and employ multimodal methods of assessment (Kaslow et al., 2007). In addition, having key competencies assessed by experts other than one's supervisor, appears an essential and urgent reform.

References

Baird, B., N. (2005). *The internship practicum, and field placement handbook: A guide for the helping professions* (4th ed.). New Jersey: Pearson-Prentice Hall.

Blackburn, L. M., James, I. A., Milne, D. L., Baker, C., Standart, S. H., Garland, A., & Reichelt, F. K. (2001). The revised cognitive therapy scale (CTS-R): psychometric properties. *Behavioural and Cognitive Psychotherapy, 29*, 431-446.

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*, 385-401.

Bogo, M., Regehr, C., Hughes, J., Power, R., Woodford, M., & Regehr, G. (2004). Toward new approaches for evaluating student field performance: Tapping the implicit criteria used by experienced field instructors. *Journal of Social Work Education, 40,* 417-426.

Borders, L., & Fong, M. L. (1991). Evaluations of supervisees: Brief commentary and research report. *Clinical Supervisor, 9*(2), 43-51.

Bushnell, J. A., Nicholson Perry, K., Blackman, R., Allan, C., Nasstasia, Y., Knight, R., Shires, A., Deane, F., et al. (2011, October). *Where angels fear to tread? Leniency and the halo effects in practicum-based assessment of student competencies.* Paper presented at the Australian Technology Network of Universities Conference, Curtin University, Perth.

Chambers, M. (1998). Some issues in the assessment of clinical practice: A review of the literature. *Journal of Clinical Nursing, 7*, 201-208.

Dienst, E. R., & Armstrong, P. M. (1988). Evaluation of students' clinical competence. *Professional Psychology: Research and Practice, 19,* 339-341.

Dolan, G. (2003). Assessing student nurse clinical competency: will we ever get it right? *Journal of Clinical Nursing, 12*, 132-141.

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*(Suppl.), S5–S26.

Falender, C. A., Cornish, J. A. E., Goodyear, R., Hatcher, R., Kaslow, N. J., Leventhal, G., et al. (2004). Defining competencies in psychology supervision: A consensus statement. *Journal of Clinical Psychology, 80*, 771–786.

Falender, C. A., & Shafranske, E. P. (2011). The Importance of Competency-based Clinical Supervision and Training in the Twenty-first Century: Why Bother? *Journal of Contemporary Psychotherapy, 41*,1-9.

Falender, C. A., & Shafranske, E. P. (2012). *Getting the most out of clinical training and supervision: A guide for practicum students and interns.* Washington, D.C.: American Psychological Association.

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

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.

Gonsalvez, C. J., Oades, L., & Freestone, J. (2002). The objectives approach to clinical supervision: Towards integration and empirical evaluation. *Australian Psychologist, 37* (1), 68-77

Hatcher, R. L., & Lassiter, K. D. (2007). Initial training in professional psychology: The practicum competencies outline. *Training and Education in Professional Psychology, 1*, 49–63.

Howell, D.C. (2007). *Statistical methods for Psychology*, 6th Edition. Thomson Wadworth.

Kaslow, N. J. (2004). Competencies in professional psychology. *American Psychologist, 59,* 774–781.

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*(Suppl.), S27-S45.

Kaslow, N. J., Rubin, N. J., Bebeau, M., Leigh, I. W., Lichtenberg, J., Nelson, P. D., et al. (2007). Guiding principles and recommendations for the assessment of competence. *Professional Psychology: Research and Practice, 38,* 441–451.

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

Leigh, I. W., Smith, I. L., Bebeau, M., Lichtenberg, J., Nelson, P. D., Portnoy, S., et al. (2007). Competency assessment models. *Professional Psychology: Research and Practice, 38,* 463–473.

Lichtenberg, J., Portnoy, S., Bebeau, M., Leigh, I. W., Nelson, P. D., Rubin, N. J., et al. (2007). Challenges to the assessment of competence and competencies. *Professional Psychology: Research and Practice, 38,* 474–478.

Norcross, J. C., Stevenson, J. F., & Nash, J. M. (1986). Evaluation of internship training: Practices, problems and prospects. *Professional Psychology: Research and Practice, 17,* 280-282.

Pachana, N. A., Sofronoff, K., Scott, T., & Helmes, E. (2011). Attainment of competencies in clinical psychology training: Ways forward in the Australian context. *Australian Psychologist, 46* (2), 67-76.

Regehr, G., Bogo, M., Regehr, C., & Power, R. (2007). Can we build a better mousetrap? Improving the measures of practice performance in the field. *Journal of Social Work education, 43,* 327-343.

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. *The Clinical Supervisor, 16* (2), 49-72.

Roth, A.D. & Pilling, S. (2008). *A competence framework for the supervision of psychological therapies.* Retrieved from University College London website: http://www.ucl.ac.uk/clinical-psychology/core/competence_frameworks.htm

Rubin, N. J., Bebeau, M., Leigh, I. W., Lichtenberg, J., Smith, I. L., Nelson, P. D., et al. (2007). The competency movement within psychology: An historical perspective. *Professional Psychology: Research and Practice, 38,* 452–462.

Townsend, A. H., McIlvenny, S., Miller, C. J., & Dunn, E. V. (2001). The use of an objective structured clinical examination (OSCE) for formative and summative assessment in a general practice: Clinical attachment and its relationship to final medical school examination performance. *Medical Education, 35,* 841-846.

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

Watkins, C.E. (1995). Psychotherapy supervisor and supervisee: Developmental models and research nine years later. *Clinical Psychology Review, 15,* 647-680.

Williams, R. G., Klamen, D. A., & McGaghie, W. C. (2003). Cognitive, social and environmental sources of bias in clinical performance ratings. *Teaching and Learning in Medicine, 15* (4), 270-292.

Appendix

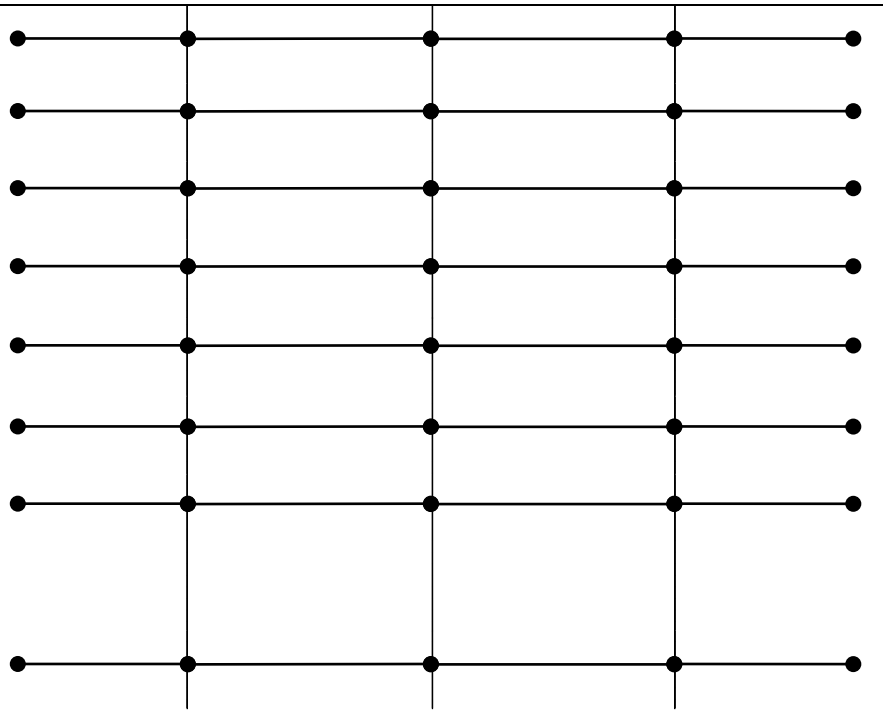
Scientist-Practitioner Vignette, Stage 4: Competent (calibration score = 9.43, SD = 1.09)

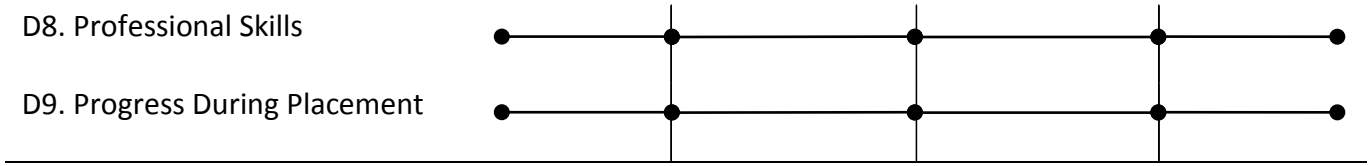
Trainee A consistently demonstrates a commitment to bringing the scientific method to her/his clinical work. She/he uses a systematic hypothesis generation and testing approach in her/his work with clients, appropriately seeking information through interview, observation or psychometric testing to test her/his clinical formulations. She/he seeks evidence of reliability and validity in making decisions about which assessment methods (e.g., tests) to use. She/he routinely accesses scholarly scientific resources (e.g., journals) to guide decisions about the most effective treatments to use. When research is lacking or unclear regarding the best treatment approach, she/he shows the ability to tailor a treatment program for the client based on an analysis of the available evidence or scientific principles. She/he values and usually implements systematic assessment of treatment progress in clients.

Footnotes

[1]Support for this project has been provided by the Australian Learning and Teaching Council Ltd, an initiative of the Australian Government Department of Education, Employment and Workplace Relations.

Table 1. Competence domains, developmental stages, and bands demarcating the four stages.

Domain	Developmental Stage			
	Stage 1[a] Beginner	Stage 2[b]	Stage 3[c]	Stage 4[d] Competent
Numeric Score	0……1……..2……..3……..4…….5…….6…….7…….8…….9…….10			
D1. Relational Skills				
D2. Clinical Assessment Skills				
D3. Case Formulation Skills				
D4a. Intervention Skills – Non-CBT				
D4b. Intervention Skills – CBT				
D5. Psychometric Skills				
D6. Scientist-Practitioner Approach				
D7. Ethical Approach				



Note. [a] Stage 1- Beginner; Knowledge and skills are at an early stage or yet to be developed. Inadequate knowledge and/or difficulty applying knowledge to practice. Several problems or inadequacies occur during sessions. There may be an absence of key features, inability to prioritise issues or to make appropriate judgements. Little awareness of process issues. On par with trainees commencing training without any practicum experience. Regular and intensive supervision required.

[b] Stage 2; Some basic competencies in assessment and intervention, manages narrow range of clients with low levels of severity, using structured therapeutic activities. Performance is variable; major problems may occur occasionally; regular supervision required.

[c] Stage 3; Moderate repertoire of basic competencies in both assessment and intervention leading to management of a wider range of clients. Demonstrates understanding o underlying principles and a moderate ability to generalise these to new cases/situations. Performance can be improved in minor ways; less frequent supervision required.

[d] Stage 4 – Competent; Large repertoire of basic to advanced competencies in both assessment and intervention, applied across range of clients and severity levels. Performance has reached competency levels on a par with a clinical psychology working in their first job upon qualification.

Table 2. Percentage of trainees matched to vignettes by field supervisors using the Vignette-Matching Procedure (VMP) in Study 1.

Domain	N	Stage 1[a]	Stage 2	Stage 3	Stage 4[a]	
D1. Relational Skills	20	5%	20%	55%	20%	
D2. Clinical Assessment Skills	20	5%	15%	60%	20%	
D3. Case Formulation Skills	19	-	32%	58%	10%	
D4b.Intervention Skills – CBT	20	-	35%	55%	10%	
D7. Ethical Practice	20	-	15%	30%	55%	
		Unsat.	**Slow**	**Incon.**	**DWell**	**Excel**
D9. Progress During Placement[b]	20	5%	5%	5%	65%	20%

Note. [a] Stage 1 = Beginner; Stage 4 = Competent. [b]This domain is represented by five vignettes and measures response to supervision and progress during placements. Unsat = Unsatisfactory; Slow = Slow Progress; Incon = Inconsistent Progress; DWell = Developing Well; Excel = Excellent Progress.

Table 3. Intraclass correlations and mean calibration scores (N=15; *SD* values are in parentheses) assigned to vignettes by criterion experts using the visual analogue scale[a].

Domain	ICC[b]	Vignettes numbers within each domain				
		1	2	3	4	5
1. Relational Skills	0.86	1.10 (1.19)	3.27 (1.09)	6.29 (1.68)	8.92 (1.49)	
2. Clinical Assessment Skills	0.88	1.21 (1.04)	3.35 (1.44)	4.73 (1.35)	8.89 (0.97)	
3. Case Formulation Skills	0.84	1.83 (1.37)	3.29 (2.03)	6.12 (1.15)	9.25 (1.03)	
4a. Intervention Skills – Non-CBT	0.89	1.30 (1.24)	2.63 (1.64)	7.74 (1.36)	8.95 (0.83)	
4b. Intervention Skills – CBT	0.90	1.57 (1.45)	2.53 (1.31)	7.63 (1.05)	8.90 (1.01)	
5. Psychometric Skills	0.90	0.83 (0.75	2.79 (1.92)	6.73 (1.05)	9.23 (0.97)	
6. Scientist Practitioner Approach	0.91	0.73 (0.82)	2.77 (1.26)	4.68 (1.39)	9.43 (1.09)	
7. Ethical Practice	0.91	0.33 (0.59)	2.18 (1.37)	6.11 (1.03)	9.28 (1.39)	
8. Professional Skills	0.82	1.90 (1.64)	4.20 (2.23)	7.08 (0.94)	9.45 (0.82)	
9. Progress During Placement	0.88	1.08 (0.96)	2.21 (1.09)	3.29 (1.71)	7.08 (1.29)	9.46 (1.20)

Note. [a] The visual analogue scale ranged from 0 (Unskilled) to 10 (Competent); [b]=Intraclass correlations, $p<.001$ in each instance. The final domain was represented by 5 vignettes. Underlined mean and *SD* values represent vignette scores that violated one or more validation criteria.

Table 4. Percentage of trainees (N = 57) assigned to the four developmental stages based on the CΨPRS and VMP.

Empty cells represent zero values.

	CΨPRS				**Vignette Matching Procedure**			
	Stage 1	**Stage 2**	**Stage 3**	**Stage 4**	**Stage 1**	**Stage 2**	**Stage 3**	**Stage 4**
1. Relational Skills	-	-	29.8%	70.2%	-	1.8%	54.6%	43.6%
2. Clin. Assessment	-	-	32.8%	67.3%	-	26.3%	17.5%	56.1%
3. Case Formulation	-	-	37.4%	62.6%	-	3.6%	62.5%	33.9%
4. Intervention Skills	-	-	30.8%	69.2%	-	3.6%	47.3%	49.1%
5. Psychometrics	-	2.0%	29.1%	68.9%	9.7%	16.1%	64.5%	9.7%
6. S-P Approach	-	-	19.9%	80.1%	-	6.5%	58.1%	35.5%
7. Ethical Practice	-	-	20.4%	79.7%	-	1.8%	29.1%	69.1%
8. Professional Skills	-	0.6%	16.8%	82.7%	-	6.7%	40.0%	53.3%
9. Progress During Placement	-	-	15.6%	84.5%	-	2.0%	41.2%	56.9%
Grand Mean	**-**	**0.3%**	**25.3%**	**74.5%**	**1.1%**	**7.6%**	**46.1%**	**45.3%**

Table 5. Between-domain correlations for the CΨPRS (top right) and the vignette-matching procedure (VMP; bottom left; **bold**).

Shaded cells (diagonal) represent correlations between CΨPRS and the VMP for the same domain.

CΨ/VM	CΨ-D1	CΨ-D2	CΨ-D3	CΨ-D4	CΨ-D5	CΨ-D6	CΨ-D7	CΨ-D8	CΨ-D9	
VM-D1	.50^	.84^	.89^	.84^	.61^	.66^	.73^	.74^	.70^	CΨ-D1
VM-D2	**.39****	.56^	.94^	.94^	.78^	.79^	.78^	.83^	.73^	CΨ-D2
VM-D3	**.41****	**.52^**	.47^	.95^	.73^	.83^	.80^	.81^	.78^	CΨ-D3
VM-D4	**.42****	**.42****	**.47^**	.40**	.77^	.79^	.78^	.79^	.76^	CΨ-D4
VM-D5	**.00**	**.32**	**.28**	**.43****	.65^	.78^	.75^	.69^	.70^	CΨ-D5
VM-D6	**.31**	**.34**	**.63^**	**.35**	**.14**	.60^	.81^	.76^	.80^	CΨ-D6
VM-D7	**.50^**	**.48^**	**.47^**	**.60^**	**.27**	**.53****	.57^	.73^	.87^	CΨ-D7
VM-D8	**.34**	**.43****	**.61^**	**.50****	**.29**	**.81^**	**.63^**	.74^	.85^	CΨ-D8
VM-D9	**.58^**	**.35***	**.46^**	**.45^**	**.32**	**.51****	**.42****	**.50****	.46^	CΨ-D9
	VM-D1	**VM-D2**	**VM-D3**	**VM-D4**	**VM-D5**	**VM-D6**	**VM-D7**	**VM-D8**	**VM-D9**	CY/VM

Note. CΨ = CΨPRS; VM = Vignette Matching Procedure; D = Domain x. *p = .05; ** p = .01; ^p< .001. CΨPRS correlations were significantly higher than corresponding VMP correlations (p <.05 or less) in 33 of 36 pair-wise comparisons. The 3 non-significant comparisons are underlined.

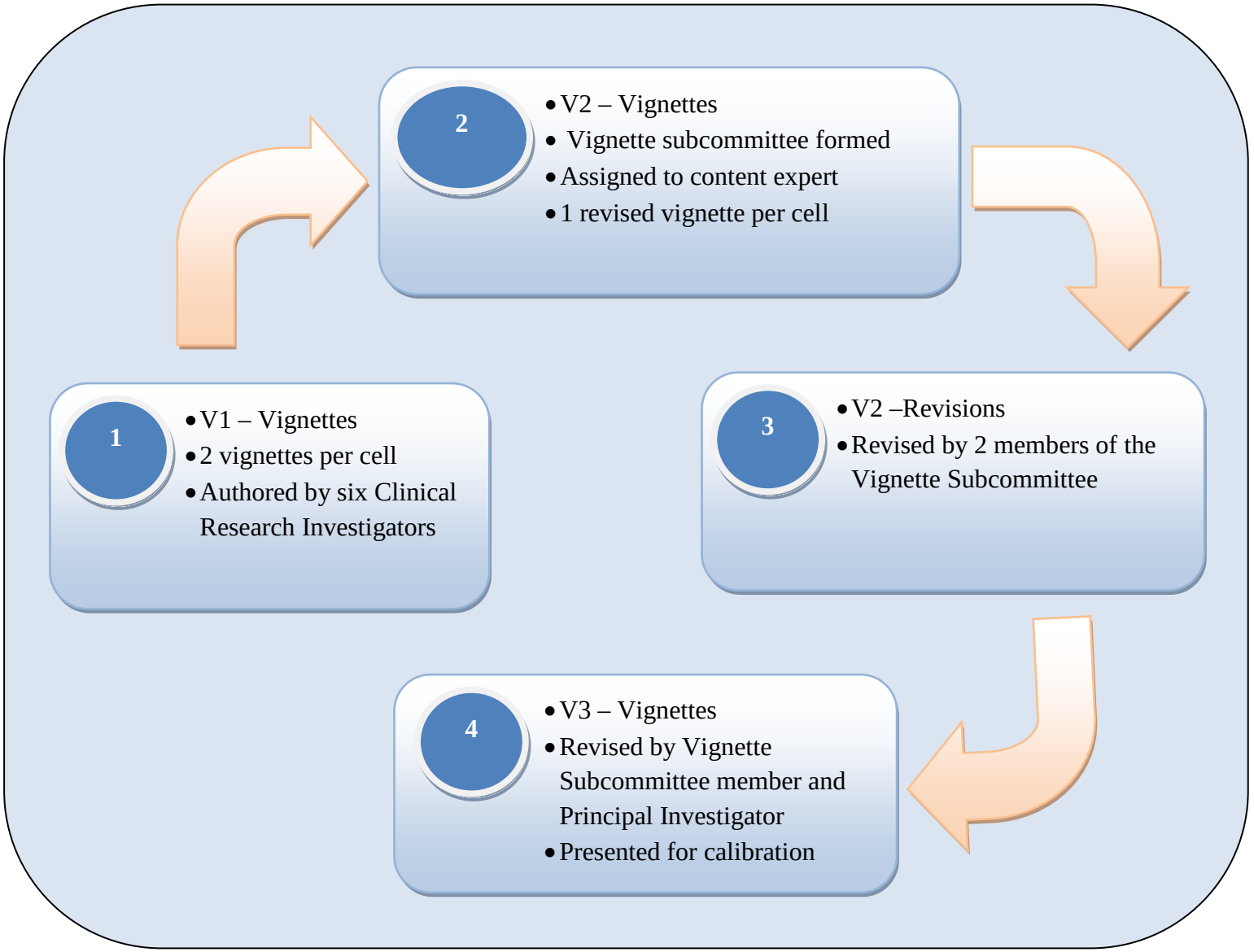


Figure 1. Steps for vignette development.

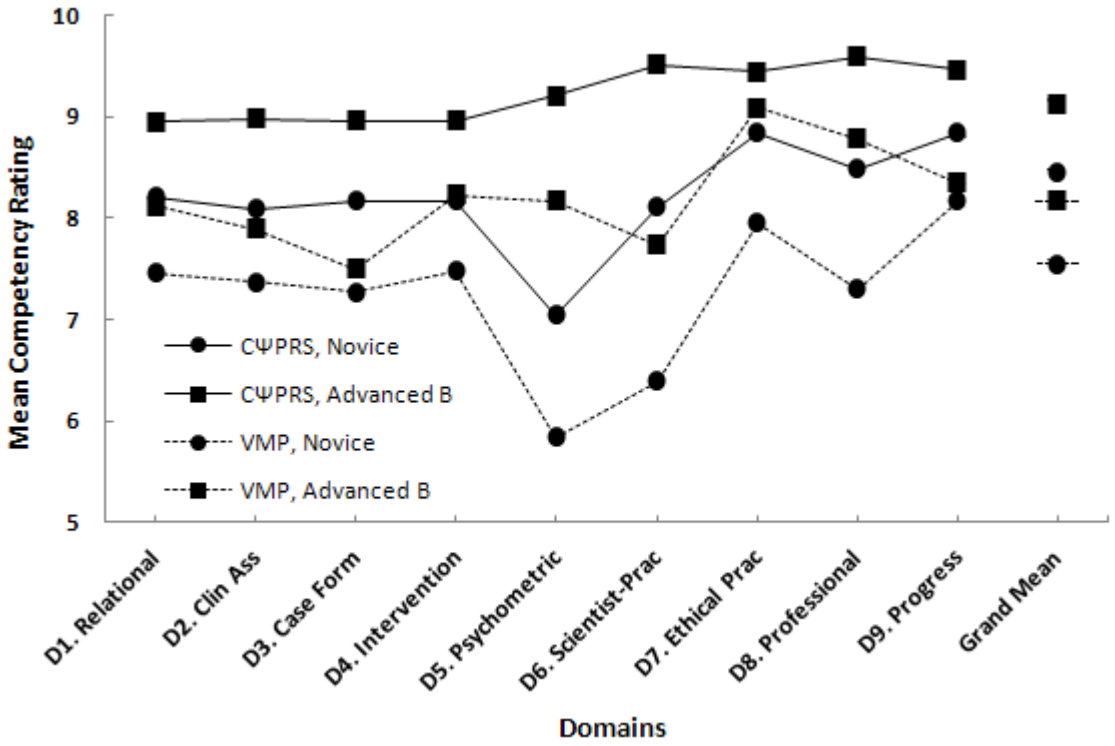


Figure 2. Mean domain scores obtained by novices and advanced beginners on the CΨPRS and the VMP.

Note: Clin Asst = Clinical Assessment, Case Form = Case formulation, Scientist-Prac = Scientist-Practitioner, Ethical Prac = Ethical Practice; Progress = Progress during placement; Advanced B = Advanced Beginner.

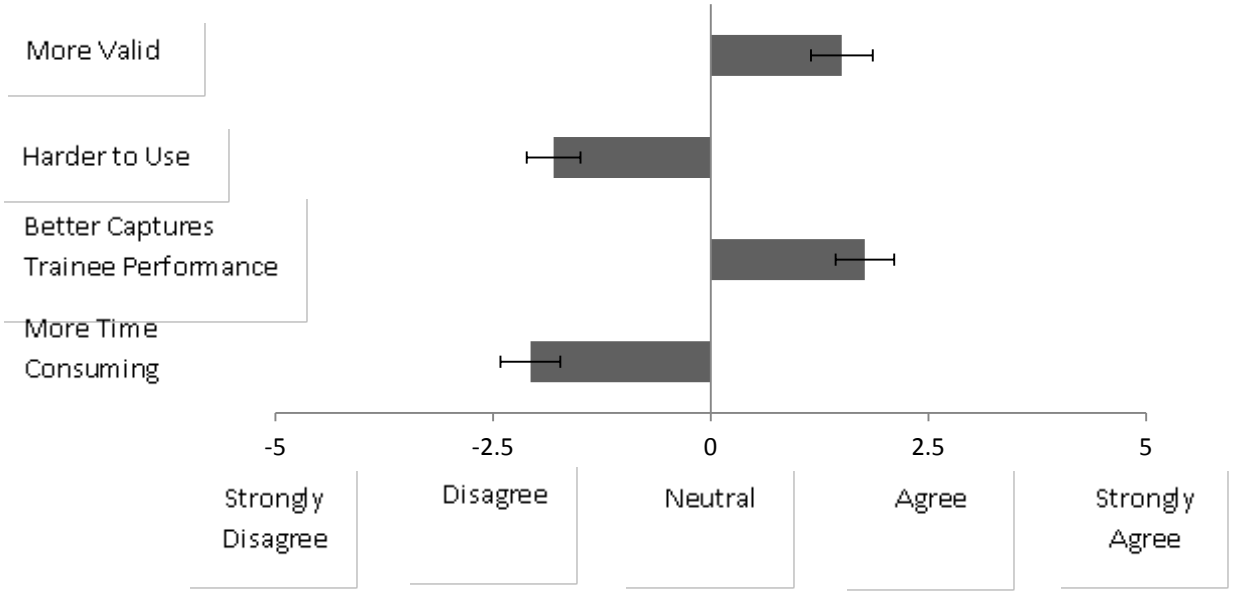


Figure 3. Field supervisors' (N=57) evaluations (means and standard error bars) of the vignette-matching procedure in comparison with the Clinical Psychology Practicum Competencies Rating Scale (CΨPRS).