Adv in Health Sci Educ (2011) 16:405–425
DOI 10.1007/s10459-011-9296-1

REVIEW

Supervisor assessment of clinical and professional competence of medical trainees: a reliability study using workplace data and a focused analytical literature review

D. A. McGill · C. P. M. van der Vleuten · M. J. Clarke

Received: 28 November 2010 / Accepted: 14 April 2011 / Published online: 24 May 2011

Abstract Even though rater-based judgements of clinical competence are widely used, they are context sensitive and vary between individuals and institutions. To deal adequately with rater-judgement unreliability, evaluating the reliability of workplace rater-based assessments in the local context is essential. Using such an approach, the primary intention of this study was to identify the trainee score variation around supervisor ratings, identify sampling number needs of workplace assessments for certification of competence and position the findings within the known literature. This reliability study of workplace-based supervisors' assessments of trainees has a rater-nested-within-trainee design. Score variation attributable to the trainee for each competency item assessed (variance component) were estimated by the minimum-norm quadratic unbiased estimator. Score variance was used to estimate the number needed for a reliability value of 0.80. The trainee score variance for each of 14 competency items varied between 2.3% for emergency skills to 35.6% for communication skills, with an average for all competency items of 20.3%; the "Overall rating" competency item trainee variance was 28.8%. These variance components translated into 169, 7, 17 and 28 assessments needed for a reliability of 0.80, respectively. Most variation in assessment scores was due to measurement error, ranging from 97.7% for emergency skills to 63.4% for communication skills. Similar results have been demonstrated in previously published studies. In summary, overall supervisors' workplace based assessments have poor reliability and are not suitable for use in certification processes in their current form. The marked variation in the supervisors' reliability in assessing different competencies indicates that supervisors may be able to assess some with acceptable reproducibility; in this case communication and possibly overall competence. However, any continued use of this format for assessment of trainee competencies

D. A. McGill (✉)
Department of Cardiology, The Canberra Hospital, Garran, ACT 2605, Australia
e-mail: darryl.mcgill@act.gov.au

C. P. M. van der Vleuten
Department of Educational Research and Development, Maastricht University, Maastricht, The Netherlands

M. J. Clarke
Clinical Trial Service Unit, University of Oxford, Oxford, UK

necessitates the identification of what supervisors in different institutions can reliably assess rather than continuing to impose false expectations from unreliable assessments.

Keywords Assessment · Clinical education · Competence assessment · Performance assessment · Professional judgment · Rater bias · Rating process · Supervisor assessment · Workplace-based assessment

Introduction

Assessments in any form and in all contexts require validation. A relative scarcity of good quality published research on the validation for medical education assessment systems has been highlighted in a number of systematic reviews over the last decade (Hutchinson et al. 2002; Hamdy et al. 2006; Kogan et al. 2009; Miller and Archer 2010). An important aspect of validation is identifying the reliability of an assessment. A well known assertion is that reliability is a necessary although not sufficient requirement for claiming validity (Downing 2004; Streiner and Norman 2009), which is separate from internal consistency and content validity (Sadler 1989; Streiner and Norman 2009); the higher the reliability the higher the maximum possible validity (Streiner and Norman 2009). Importantly: "Although statistical investigations may not be much help in identifying cause, they are able to indicate the presence and likely extent of error or bias" (Sadler 2009).

Detailed knowledge of the problems surrounding observer ratings are well documented in many disciplines, but variation in rater-based assessment of clinical performance is particularly problematic (Kegel-Flom 1975; Streiner 1995; van der Vleuten et al. 2000). Many types of rater bias are commonly manifest when human judgement comprises part of an assessment process of any type (Saal et al. 1980; Williams et al. 2003; Ronan and Prien 1966, 1971; Latham et al. 1975). Raters vary in the behaviours they notice, and vary in how they evaluate those behaviours that do attract their attention, including what they would include in a global assessment (Mazor et al. 2007). The well recognised rater biases for subjective performance assessment appear to remain ubiquitous (Viswesvaran et al. 2005; Williams et al. 2003; van Barneveld 2005). Indeed this bias may be severe. When method bias was corrected in one study, "the results indicated that the dependability of ratings of student performance was almost zero" (van Barneveld 2005). Such rater bias has long been known to be common for the evaluation of an individual's competence by another human's judgment (Thorndike 1920; Saal et al. 1980; King et al. 1980; Levine and McGuire 1971; Kastner et al. 1984). Reducing rater variation and other forms of error will potentially increase the ability of assessments to discriminate between assessment performances (Wherry and Bartlett 1982) and also improve the integrity of the assessment so as to identify true trainee competency achievement (Sadler 2009). Methods exist to help identify such potential sources of variation (Wherry and Bartlett 1982; Cronbach et al. 1972; Saal et al. 1980; Crossley et al. 2007).

The continual improvement in the reliability of rater assessments must be an on-going educational goal (Downing 2004; Streiner and Norman 2009). To achieve this goal the reliability of assessments needs to be evaluated critically, both at a local institutional level and in the national arena, a concept well accepted in higher education (Wainer and Thissen 1996; Joint Committee on Standards for Educational, Psychological Testing of the American Educational Research Association 1999; Koretz 2003; Sadler 2005). Multiple sampling methods to take account of all identifiable unwanted sources of variance in assessments within an assessment programme, rather than relying on an

individual assessment method, offers one way to improve reliability (Swanson et al. 1995; Koretz 2003; van der Vleuten and Schuwirth 2005). Nonetheless, choosing an assessment method inevitably involves compromises which may vary for each specific assessment context (Koretz 2003; van der Vleuten and Schuwirth 2005). Despite the ever-present need for compromise, the ability of any assessment to reliably differentiate trainees' true competence measured against an accepted standard remains an important and prime objective of any method of assessment (Wherry and Bartlett 1982; Streiner and Norman 2009).

Three conceptual assertions about the evaluation of the reliability of rater-based performance assessments will be emphasized. Firstly, variance component estimates rather than reliability coefficients provide more useful information for nested workplace data (Cronbach and Shavelson 2004), and will be used as a primary outcome of reliability. Variance components are the proportion of the total score variation attributable to different influences that might affect the trainee score. Besides differences between trainees (trainee variance), examples include supervisor variation (rater variance) and, for assessments using cases, the types of cases involved (case variance). Nesting in this instance refers to circumstance where repeated assessments for trainees are performed by different supervisors, meaning that supervisors are "nested in" trainees. A second assertion is that the main objective for a reliability evaluation of an assessment method using hierarchical scales is to determine the score variation attributable to the trainee and separate this from variation due to other factors which would be considered error (Wherry and Bartlett 1982; Streiner and Norman 2009). The third assertion is that sampling across contexts (workplaces) and contents (type of competences) is essential for both establishing and improving reliability of rater-based assessments (van der Vleuten and Schuwirth 2005). To be able to evaluate the reliability and sources of variance of rater-based assessments for the workplace, methods are needed to be able to do so easily and repeatedly at a local institutional level, which will also allow benchmarking of the findings with other comparable institutions and contexts.

The specific aims of this evaluation study are to: (1) identify the trainee score variation around supervisor ratings for end-of-term workplace assessments of junior trainees in a local institutional network; (2) identify what the trainee score variation means for sampling judgments in the context of workplace assessment; (3) compare the findings with other publications of similar rater-based assessments using a focused analytical literature review; and (4) present a simple reproducible methodology for reliability evaluations of multiple rater-based workplace assessments.

Methods

Context and population

All assessments that were available and used to judge the ability of a first year postgraduate trainee in the Australian medical system (trainee) to become unconditionally registered as a medical practitioner in the Australian Capital Territory (ACT) were included. The sample consisted of all trainees for the years 2007 and 2008 at The Canberra Hospital in the ACT and the three secondment hospitals. There were no exclusion criteria and no exclusion of any assessment. Every assessment performed was included for all trainees, all supervisors assessing the trainee, and for all competency items assessed.

Trainee performance rating process

The exact 2007–2008 version of the form and the method of application as recommended by the then New South Wales Institute of Medical Education and Training (IMET) personnel were used for all assessments (Appendix). The assessment form is meant to be completed during a formative assessment half way through the 10 week term (training rotation), and again as a summative assessment at the end of each 10 week term. Each competency item identified on the summative assessment forms for each trainee comprises the unit of analysis for this evaluation. The form is completed by the trainee's nominated supervisor for that particular term. The basic process is similar to the Global Ratings described in the Accreditation Council for Graduate Medical Education (ACGME) *Toolbox of Assessment Methods* (Accreditation Council for Graduate Medical Education 2000).

The trainee assessment form consists of a number of items representing different types of constructs of clinical and professional performance to be judged by the assessor. The trainee assessment occurs within different hospitals and within these different hospitals within different terms; usually by the same term-based supervisor for trainees rotating through the term, but not always. There are 5 rotation terms and the trainee has a supervisor for each rotation. Therefore the trainee is assessed by 5 different supervisors. The supervisor often uses information about the trainee's competence and performance from other sources. These include other senior medical staff in the unit and often registrars. Often opinions are also sought from senior nursing staff with whom the trainee works on a daily basis, and sometimes from other professional staff. For relief terms (during which the trainee has a 10 week period when they are seconded to different clinical departments to cover junior doctors away on various forms of leave), the support medical administrator completes the form following consultation with senior staff from all the areas where the trainee worked as relief. The scale descriptors were "requires substantial assistance", "requires further development", "consistent with level of performance", and "performance better than expected" were "scored" as 1, 2, 3 and 4 respectively. "Not applicable/not observed" were treated as missing values and reassessed in a sensitivity analysis (Appendix).

The coded de-identified data from all forms were collated in an SPSS database (Version 16) on a high-level-password protected laptop computer. No analysis is about individuals for this purpose, and information about any individual is not accessible from any public documents produced. This evaluation was undertaken as part of a quality improvement process for the current assessment methods for junior trainees within the network; evaluation being part of the remit of the institutional General Clinical Training Committee. As such, Institutional Ethics Committee Approval was not sought.

Reliability evaluation

Variance components

A reliability study was performed to ascertain the variance components for the assessment of trainees by their supervisors at the end of each clinical rotation. Variance components analysis, a multilevel linear model, was used because the data for participants is organised at more than one level which will involve repeated measurements of individuals, and because of concerns about identifying latent variables (Tabachnick and Fidell 2007). The competency item score for each trainee assessment provided by different supervisors is the dependent variable.

Supervisors are nested with trainees because different supervisors (raters) rated each trainee. This study design was a rater-nested-within-trainee design. Ideally, most of the variance is explained by the trainee, and that variance is equivalent to the "true score" in classical measurement theory. All other variance, that is the variance explained by the raters, all interactions and random effects, is considered error. Since the design is rater-nested-within-trainee there are only two measureable sources of variance: trainee assessment score variance and what would be considered error variance. The error variance includes rater effect, trainee-rater interaction, general error variance and trainee x rater x general error interaction variance. The general error includes the confounding effects by latent unidentified confounding variables and random error.

The variance components were estimated by the minimum norm quadratic unbiased estimator, (MINQUE), analysis of variance type III (ANOVA III), and the restricted maximum likelihood REML (Baltagi et al. 2002). The MINQUE method was chosen because it requires no distributional assumptions and is recommended for unbalanced designs (Baltagi et al. 2002); ANOVA III to estimate the degrees of freedom (Crossley et al. 2007); and REML to confirm the MINQUE results (Baltagi et al. 2002). Since it is possible that some variance components may have negative estimates by MINQUE, the REML method was also used to determine if the competency items having negative variance components were redundant. If redundancy was confirmed, then any such variance components found with ANOVA III or MINQUE(I) were assumed to be zero.

The variance components analysis was undertaken for each of the competency items. Because of the potential for bias in workplace evaluations, all variance components are fully reported and also expressed as percent variance of overall variance. Assessments with missing items were excluded for the primary analysis. All assessments provided by two or more supervisors were included in the primary analysis because the variance of one supervisor cannot be determined from one assessment.

Number needed to improve reliability (NNIR)

The percent variance due to the trainee and the number of observations needed to achieve a reliability coefficient of 0.80 were calculated using the standard reliability formula from Classic Test Theory (Streiner and Norman 2009): Reliability $(R) = \sigma^2_{\text{subjects}}/\{\sigma_{\text{subjects}}{}^2 + (\sigma^2_{\text{error}}/\text{n})\}$ and therefore $\text{n} = \sigma^2_{\text{error}}/\{(\sigma^2_{\text{subjects}}/R) - \sigma^2_{\text{subjects}}\}$ where σ^2 is variance, $\sigma^2_{\text{subjects}}$ is variance due to the subjects assessed (trainees in this case), σ^2_{error} is the variance not due to the trainee, R is 0.80, and therefore n = the number of assessments needed to achieve an R of 0.80.

The design is unbalanced because unequal numbers of supervisors evaluated individual trainees and the supervisors are nested within the trainees. The study design is similar to a Generalisability G-Study identifying the variance components which in this case because of the unbalanced and nested design can only provide unbiased estimates of trainee and error variance. The estimation of the number of assessments needed to provide an adequate level of reliability is similar to a generalisability D-Study.

Sensitivity analysis

To test the robustness of the observations, three sensitivity analyses were performed. The different parameters used should not result in any substantive difference unless the results are influenced by such factors. Because of the potential biasing consequence of missing variables, an examination of their potential effect was performed by re-analysing with the

missing values substituted by an average of the sub-items for each trainee; sub-items being the competency items under the more global constructs "clinical", "communication" and "personal and professional" (Appendix). A further sensitivity analysis was performed by including supervisors who only performed one assessment. Including such assessments where the variance of the supervisor is not measurable should not substantially change the results. However, since the mean-square difference of the trainee is still included in the modelling, it is feasible that the results would be different if the assessments by supervisors performing only one assessment assessed a different trainee population. A third sensitivity analysis will compare the results with the results of the year 2007 alone for variance components. Finally, the results will be compared to those observed in the available literature.

Results

A total of 374 assessments (trainee and supervisor interaction for an end-of-term assessment) were performed involving 74 trainees, and 73 supervisors. Fourteen supervisors were female (19.2%) and contributed 107 assessments (28.6%). The 59 male supervisors (80.8%) provided 267 (71.4%) of the assessments. There were 46 female trainees (62.2%) who had 229 (61.2%) assessments, compared to 28 males (37.8%) having 145 assessments (38.3%). From the 74 trainees, 64 had 5 assessments, 12 handed in 4 and 2 had 3 assessments available. From all assessments nearly one third of supervisors (23 of 73) performed only 1 assessment (6.1% of all assessments), with 93.9% of assessments being performed by 2 or more supervisors. Seven performed 10 or more assessments, there being a wide dispersion of the number performed by any one supervisor.

Mean scores for all competency items were all above 3, 4 of 14 competency items had a median of 4, and the rest 3 (Table 1). No trainee was rated less than optimal for "Professional obligations" and "Professional responsibility". There were negligible low ratings for Procedural skills (0.06%), Team skills (0.05%), Awareness of limitations (0.05%), Teaching and Learning (0.06%), and Medical Records (0.11%). Furthermore, Procedural skills (0.06%), Teaching and Learning (0.06%), and Medical Records (0.11%) have more associated variability in the scores.

The number of assessments with the same score for every item of the "Clinical" construct was 165 from 374 (44%), and for the "Professionalism" construct was 140 from 374 (37%). Seventy six (20.4%) assessments were all scored the same across every item; 35 (9.4%) being all 4 s for every item and 41 (11%) being all 3 s for every item.

Variance components

The variance components analysis was undertaken for each of the competency items. Each competency item was analysed and the variance, percent variance and degrees of freedom are summarised in Table 2.

Error variance is dominant as the main effects for the variance components of the different competency items (Table 2). Apart from communication skills, clinical judgement, teamwork and overall rating, the variance contributions for the trainee were less than 25%. For this type of data-set the interaction effects cannot be separated and are considered part of the residual variance (Cronbach and Shavelson 2004). All variance apart from that due to the trainee is considered error variance.

Table 1 Descriptive statistics for competency items

Competency item	Valid number	Mean (SD)	Median[a]
Knowledge base	373	3.28 (0.500)	3.00
Clinical skills	374	3.39 (0.539)	3.00
Clinical judgement/decision making	373	3.42 (0.601)	3.00
Emergency skills	318	3.20 (0.474)	3.00
Procedural skills	339	3.29 (0.490)	3.00
Communication	374	3.56 (0.543)	4.00
Teamwork skills	374	3.63 (0.501)	4.00
Professional responsibility	374	3.61 (0.488)	4.00
Aware of limitations	372	3.44 (0.503)	3.00
Professional obligations to patients	372	3.46 (0.499)	3.00
Teaching/learning	334	3.29 (0.474)	3.00
Time management skills	372	3.40 (0.547)	3.00
Medical records	370	3.39 (0.506)	3.00
Overall rating	374	3.49 (0.542)	4.00

[a] Inter-quartile range for all items was 3–4

Table 2 Trainee true score and percent variance

Competency	Trainee variance		Error variance		
	% Variance component[a]	Variance component	% Variance component	Variance component	d.f.
Knowledge	18.7	0.046	81.3	0.200	273
Clinical skills	17.8	0.051	82.2	0.235	273
Clinical judgement	29.6	0.105	70.4	0.250	274
Emergency skills	2.3	0.005	97.7	0.211	221
Procedural skills	9.4	0.022	90.6	0.213	245
Communication skills	35.6	0.104	64.4	0.182	274
Teamwork skills	26.9	0.067	73.1	0.182	274
Professional responsibility	18.1	0.043	81.9	0.195	274
Awareness of limitations	18.6	0.047	81.4	0.206	272
Professional obligations	19.7	0.049	80.3	0.200	272
Teaching and learning	13.1	0.028	86.9	0.196	238
Time management	24.3	0.070	75.7	0.220	272
Medical records	21.4	0.054	78.6	0.198	272
Overall rating	28.8	0.084	71.2	0.202	274
Averaged items	20.3		79.7		

d.f. = degrees of freedom obtained from an ANOVA III analysis; [a] all d.f. = 74

Variance components, reliability coefficients and the "number needed to increase reliability" (NNIR)

Continuing with the same line of analysis, it is feasible to use the data to provide a reliability coefficient, not for reporting reliability but to use the same formula for calculation of the number needed to optimise the reliability coefficient to the recommended $\geq$0.80 values. Therefore the number of assessments needed about a trainee's competency can be estimated from known data for the same supervisors (they may change with different context and different supervisors). The number of assessments needed to improve the assessment reliability varies between the competency items substantially, ranging from 7 for assessing communication skills to 169 for assessing emergency skills. Only the assessment of clinical judgement, communication skills and overall rating needed 10 or less assessments (Table 3).

Sensitivity analyses

The initial results remain robust to changing the parameters used in the model. Any variations were small and there were no significant differences for the population size. The re-evaluation using the average of the scores for the missing values demonstrated minor differences in the variance components overall (Table 4) as did the results from the re-analyses of the data using all assessments including those provided by supervisors who undertook only one assessment, but did not change the overall pattern of results (Table 4). There were more differences in the item variance components for the year 2007 compared with the combined years 2007–2008, although still with the same overall pattern of variance components. The main differences were for those competencies that had the worst reliability, namely: "Emergency Skills", "Procedural Skills", "Teaching and Learning" and "Awareness of Limitations", but again they were not statistically significant for the

Table 3 The number of observations of a competency item for a reliability coefficient of 0.80

Observed competency	Number observations to achieve an $R = 0.80$[a]	Reliability coefficient (R)
Knowledge	17.4	0.535
Clinical skills	18.4	0.520
Clinical judgement	9.5	0.677
Emergency skills	168.8	0.106
Procedural skills	38.7	0.344
Communication skills	7.0	0.743
Teamwork skills	10.9	0.650
Professional responsibility	18.1	0.524
Awareness of limitations	17.5	0.534
Professional obligations	16.3	0.550
Teaching and learning	28.0	0.418
Time management	12.6	0.614
Medical records	14.7	0.577
Overall rating	9.6	0.675
Average	27.7[b]	0.533

[a] R (reliability coefficient) $= \{\sigma^2_{\text{subjects}}/(\sigma^2_{\text{subjects}} + \sigma^2_{\text{error}}/n)\}$

$n = 5$ assessments per trainee

[b] 18.8 with "emergency skills" excluded

Table 4 Trainee true score and percent variance for sensitivity analyses

Competency	Sensitivity analysis 1		Sensitivity analysis 2		Sensitivity analysis 3	
	Trainee variance with averaged score added for missing values		Trainee variance for all assessments		Trainee variance for all assessments in 2007 alone	
	% Variance component[a]	Variance component	% Variance component[a]	Variance component	% Variance component[b]	Variance component
Knowledge	15.6	0.039	17.0	0.042	18.2	0.052
Clinical skills	16.1	0.047	18.1	0.052	19.4	0.066
Clinical judgement	29.7	0.108	29.5	0.104	32.9	0.136
Emergency skills	9.6	0.021	5.4	0.012	13.0	0.037
Procedural skills	9.5	0.023	11.4	0.027	18.3	0.049
Communication skills	36.1	0.108	37.2	0.107	40.4	0.124
Teamwork skills	24.6	0.060	24.9	0.061	19.0	0.060
Professional responsibility	14.5	0.034	16.6	0.039	10.0	0.021
Awareness of limitations	13.0	0.033	16.7	0.042	10.7	0.028
Professional obligations	15.6	0.040	15.5	0.044	14.1	0.035
Teaching and learning	11.7	0.026	10.8	0.024	11.7	0.033
Time management	21.2	0.064	21.8	0.063	22.7	0.073
Medical records	17.1	0.042	18.3	0.046	18.3	0.048
Overall rating	26.6	0.082	27.6	0.079	21.8	0.106
Averaged items	18.6		19.3		19.3	

[a] d.f. = 74; [b] d.f. = 39

sample population used and the pattern of variance for each of the competencies remained the same (Table 4).

Discussion

This evaluation of the reliability of supervisors' assessments in one institutional training network provides supporting evidence for continuing problems with rater (supervisor in this context) related assessment of clinical competency. The outcomes described are not merely a local contextual problem. The observations are consistent with what has been identified previously and make it clear that the well known problems with rater judgements continue to "plague" this form of assessment and reduce its reliability (Thorndike 1920; Saal et al. 1980; Williams et al. 2003; Streiner 1995; van der Vleuten et al. 2000). The key issues of sampling adequately, identifying competencies that are inconsistently assessed, and the need for as many assessors as possible (Wass et al. 2001), are once again highlighted. End-of-term supervisors' assessments in the format used in this evaluation do not

provide an acceptable source for an assessment program since an adequate number of assessments are not achievable for individual competencies. When this form of assessment is being used, since contexts, rating scales and supervisors vary between institutions, each institution would need to evaluate the supervisors and processes to determine the reliability and sample number requirements for the local circumstances.

Reliability coefficients

Variance components, although expected to be at least described (Crossley et al. 2007), have not usually been considered a primary outcome for a reliability study in medical education. Yet the developers of Generalisability Theory believed that "The most important function of a Generalisability study is to estimate variance components" (Gleser et al. 1965). Cronbach emphasised the importance of variance components over reliability coefficients: "Coefficients are a crude device that does not bring to the surface many subtleties implied by variance components" (Cronbach and Shavelson 2004). Indeed, "the *generalizability* (G) *study* collects data for the purpose of estimating the components of variance of a measuring procedure; the *decision* (D) *study* collects data for the purpose of making decisions or interpretations" (Gleser et al. 1965).

Although reliability coefficients were not used as a prime measure of reliability, the range obtained (Table 3) are within the predicted value for general job performance ratings (Viswesvaran et al. 1996; Rothstein 1990). The average interrater reliability of supervisory ratings of job performance is 0.52 (Viswesvaran et al. 1996). In job performance assessment using 9,975 first-line supervisors from 79 different organizations, the asymptotic levels of reliability were 0.55 for duty ratings and 0.60 for ability ratings, even with 10–20 years of experience by the rater (Rothstein 1990). A similar range of reliability coefficients were demonstrated in 1959 (Cowles, J. T. & Kubany), and repeated in (Pulito et al. 2007) by studies attempting to improve the measurement of the clinical performance of medical students.

Variance component analysis

The classic psychometric model simply partitions total variance into true variance and error variance, resulting in the potential for variance due to any response bias being regarded as true variance (James et al. 1984). Estimates of interrater reliability will therefore be biased if any of the response errors (halo, leniency and range restriction) are included in the measurements (James et al. 1984). These biases can result in responses to assessment items which do not reflect the competency being assessed for an individual. This will affect the reliability because of variation in true score interpretation between assessors, and possibly a change in the range of individual differences in the score. The impact on the validity of an assessment will therefore be twofold, a reduction in the reliability and an inappropriate interpretation of the item construct (Cronbach 1950). Given the evidence for response bias in this study and in the literature, as discussed below, the use of variance components rather than the potentially biased reliability coefficients for this type of workplace data-set appears useful for evaluating the reliability of rater-based assessment. Moreover, maximising variation in the trainee score (true score variance) and minimising all other contribution to the variance (error variance) should be a primary goal for any assessment method.

The variance components documented in this evaluation are similar to those observed in other medical education studies of various forms of assessment involving rater judgements

of the competence of other people (Kreiter et al. 1998; James et al. 2002; Kreiter et al. 2002; Margolis et al. 2006; de Lima et al. 2007; van Barneveld 2005). The error variance in these studies is also large. For example, the rater variation is much larger than that for the trainee for the mini-CEX (Margolis et al. 2006), the most reliable assessment method by direct observation of competencies (Kogan et al. 2009). Interactions between the rater and trainee, and the residual ("error" component) are substantial for the mini-CEX and other rater-based assessments (Kreiter et al. 1998; James et al. 2002; Kreiter et al. 2002; Margolis et al. 2006; de Lima et al. 2007). The presence of a significant rater-by-ratee interaction has been interpreted as showing the presence of response bias, especially the halo effect (Margolis et al. 2006). The large error variances, of the order of 80% or more in the studies cited, suggest that other confounders might also be affecting the reliability of assessments, and these need to be identified and evaluated given their biasing effects (Norcini et al. 2003). For example confounders could include the setting and context, gender of the trainee and/or supervisor, and professional designation of the rater (Wilkinson et al. 2008). Such bias is not limited to ratings in medical assessments (Dickinson and Tice 1977; Viswesvaran et al. 1996). The presence of such potential confounders could be examined as part of any reliability study in the local workplace context.

The lack of a dominant trainee "main effect", namely small variance components for trainees, is common with other studies with similar assessment processes and contexts (Kreiter et al. 1998, 2002; James et al. 2002; Margolis et al. 2006; de Lima et al. 2007; van Barneveld 2005). This can be interpreted as showing a lack of discriminability between trainees' competence (Saal et al. 1980) and/or an inability to identify authentic competence achievement (Sadler 2009), Although a lack of variation may be due to a true similarity of trainees or a true ceiling effect, it more likely is due to confounding influences and biases such as the response biases of leniency, range restriction and halo, as well as others (Saal et al. 1980), all of which will contribute to the error variance. For example, narrow standard deviations observed in this evaluation may indicate a halo effect and/or range restriction (Saal et al. 1980). Moreover, the ratings by a sizable proportion of supervisors were the same between and across assessments. A large number of assessments were rated the same across the constructs. Nearly 30% of the scores have a median of 4, the maximum score. This lack of variation within individual supervisor's assessments will either artificially increase or decrease the trainee score from the "true" score. Importantly it will also reduce the variance of the overall mean of the trainees' scores if the supervisor has the tendency to rate all the trainees similarly.

The leniency observed in this evaluation is perceived as a pervasive problem in all forms of higher education, including postgraduate medical education (Speer et al. 2000; Thompson et al. 1990; Kreiter and Ferguson 2002; Dudek et al. 2005). The scores for trainee assessment show that the majority or a large minority of trainees perform better than expected in every competence construct, suggesting an elite group of trainees with a ceiling effect, supervisors with low expectations, or other issues of leniency bias (Murphy and Balzer 1989). A large rater main effect if present can be interpreted as indicating leniency (or severity) (Dickinson and Tice 1977; Murphy and Balzer 1989) and will also contribute to the error variance. Large trainee by supervisor interaction indicates that the relative competence of trainees during their ward performance differed across supervisors. However such detailed evaluation of how the error variance components are distributed cannot be accurately determined for our nested type of data-set and so inferences about the causes of any rater error using inter-action effects cannot be made (Cronbach and Shavelson 2004).

As stated, it is feasible that these observations could indicate a number of other possibilities. A few examples that could contribute include: (1) a true ceiling effect for the abilities

of the trainees; (2) little true variation in trainee competence; (3) the competency item meaning may not be understood or differently interpreted by supervisors; and (4) poor scale development. For instance, the "coarseness" of the scale and the fact that the rating-scale has a midpoint that differs from the numerical midpoint can lead to biased scoring (Saal et al. 1980; Streiner and Norman 2009). For discrete response scales, evidence supports the use of 7, or at least between 5 and 9 categories, for many judgment tasks for most judges (James et al. 1984). Rating-system design has an important bearing on the ability of human beings to function as reliable data recording instruments (Hess 1969). Greater reliability may be achieved from methods using many discrete judgements than from methods utilising fewer but more global judgements (Hess 1969). Importantly, the "conceptual precision" about what the competency items means may not be optimised (Saal et al. 1980).

As intimated above, an important limitation for the methodology used for this and other rater-based evaluations relates to competency item construct interpretation, and hence conceptual precision and accuracy about the construct. But this is an old (Remmers et al. 1927) but recurring problem that remains unresolved (Albanese et al. 2008; Govaerts 2008). It is also assumed that one supervisor is as competent to judge the competency of an individual as another, so that an individual rater can be treated as a constant factor (Remmers et al. 1927). The supervisor in different contexts provides a specific assessment which will include their understanding of the construct of the competency being assessed (Turnbull et al. 2000). Although the understanding of the constructs is attributed to each supervisor, there is no measure of how supervisors' interpret the constructs. It is unknown if they are using the same interpretation, whether the judgement can be treated as a constant factor, whether the assessor is as competent as another to make the assessment, and even if the assessor is competent at all to make the assessment. Although this is an issue to be dealt with by test-validity evaluation, identifying whether the issue is a contributing factor to the poor reliability cannot be measured with the methods used for reliability evaluation of rater-based assessments except with a randomised intervention.

Sampling estimation

The required sampling numbers documented in this study are not new: "For individual items the number of observations required for reliable results vary from 10 to 32 with an average of 18" (Carline et al. 1989). The number needed to improve reliability for an assessment (NNIR) is often considered part of the outcome measures for reliability studies of workplace based data sets evaluating comparative reliability. In this evaluation, 5 end-of-term assessments have been used to determine competency. A much larger number of assessments than what is currently required are needed to improve reliability of this form of assessment, a well known requirement for rater-ratee based assessments (James et al. 1984).

Content sampling (different types of competencies) is also important for competency and performance assessments. The assessment method should reliably assess the particular competency-construct stated. Marked variation in the ability of the supervisors to reliably assess different competencies is evident. Clearly some competencies are not easy for supervisors to consistently assess and should be discarded from this type of assessment. The NNIR should therefore be identified for each competency item, not just for global scores. Each individual competency has relevance to patient care and need to be individually identified, developed and assessed for any trainee during their formative years and beyond. Alternative more reliable methods to evaluate these specific competencies need to be identified and used as part of an assessment program, again highlighting the need for assessment programs (van der Vleuten and Schuwirth 2005).

As noted, the NNIR for each competency item varies dramatically between individual competency assessments, from 7 for communication skills to 169 for emergency skills. Many potential reasons exist for these differences, including the classic rater biases discussed earlier (Saal et al. 1980), the inability to rate the competency or simply because the competency was not observed. Clearly the supervisors in our context cannot assess emergency skills without observing the trainee in emergencies. The problem of not observing competencies has been well known for many years (Haber and Avins 1994) and may contribute to bias by under-sampling halo. Under-sampling refers to a rater's insufficient observations of the trainee's behaviour so that small numbers of trainee samples of the behaviour "force raters to rely on a global impression, category-relevant and category-irrelevant observations, linked by beliefs about how categories covary." (Cooper 1981). A number of the competencies were stated by some supervisors as being not applicable and/or not observed. This was indicated by a separate scale "not applicable/not observed" (see Appendix). The competencies which were most frequently stated to be not applicable and/or not observed were emergency skills for 56 assessments, procedural skills for 35, and teaching skills for 40 assessments. These 3 competencies had the highest NNIRs: 169, 39 and 28 respectively. It is therefore feasible that other supervisors did not have sufficient sampling of these behaviours but still provided a score affected by under-sampling halo. As is well known, "halo inflates all within-rater correlations and deflates all between-rater correlations in comparison to values that would be observed in the absence of halo error" (Viswesvaran et al. 2005). A reduction in between-supervisor correlation will reduce the reliability value for any score, resulting in a larger number of samples needed to improve reliability. The only time this would not occur is if all raters "shared the same halo" thus increasing inter-rater correlation giving a biased increased reliability.

Comparison to other studies

The variance component results for this study of supervisors' assessment reliability show similar results to that found with the mini-CEX (Margolis et al. 2006). This is of concern because the mini-CEX has been described as the most valid of the direct observational methods assessing the clinical competence of trainees in internal medicine (Holmboe and Hawkins 1998; Kogan et al. 2009). It is possible that obtaining 20% variance due to the trainee is all that can be expected for this and other types of rater-based assessment. This possibility is supported by the information from the fully crossed study assessing the reliability of the mini-CEX method (Margolis et al. 2006). The measurable variation of supervisors' ratings and the size of the error component are substantial for each competency. The study by Margolis et al. (2006) is useful because the design has a low likelihood of having many latent variables affect the variance components. The design is more likely to give the "true" variance contribution of the trainees to the overall variance with fewer likely confounding factors. The variance due to the individual items attributable to the trainee in the current study is comparable.

The results of other studies of the reliability of the mini-CEX using variance components analysis, usually performed as part of a reliability evaluation using Generalisability Theory methods, also produce trainee and total error variance component results similar to that found in this evaluation (Hill et al. 2009; Weller et al. 2009; Cook et al. 2008; Wilkinson et al. 2008; de Lima et al. 2007; Norcini et al. 1995).

Moreover, a large study of a similar population and with a similar process for assessment, and so with potentially similar latent variables (Kreiter et al. 1998) demonstrated trainee-related variances less than those observed in this evaluation. In that study (Kreiter

et al. 1998), calculation of the percent variance for the averaged item values for Obstetrics and Gynaecology, Internal Medicine and Surgery were 13.8, 5.9 and 11.0% respectively. Trainee score depended on the particular rater that the trainee was assigned and the unique context—the relative rank orderings of the students varied considerably depending on the rater and the context: "… variation between raters overwhelms the true performance differences between the students." (Kreiter et al. 1998).

The reliability of similar types of assessment processes have been variable, although many have claimed adequate reliability (Keck and Arnold 1979; Kwolek et al. 1997; Magzoub et al. 1998; Kreiter et al. 1998; Nasca et al. 2002; Durning et al. 2005; Beckman et al. 2006; Cohen et al. 2009; Kreiter et al. 1998), others have been either equivocal (Cowles and Kubany 1959; Hull et al. 1995; Schwanz et al. 1995; Williams et al. 2004), or found the reliability not acceptable (Levine and McGuire 1971; Davis et al. 1986; Thompson et al. 1990; Metheny 1991; Ryan et al. 1996; Pulito et al. 2007; Searle 2008). A common problem with many of the studies claiming reliability for this form of competency assessment has been the inappropriate use of the alpha coefficient for nested and/or unbalanced designs which appears to be common for workplace-based assessments (Keck and Arnold 1979; Magzoub et al. 1998; Nasca et al. 2002; Durning et al. 2005; Cohen et al. 2009). One study that gave equivocal conclusions and used the alpha coefficient provides insight into the inappropriate use of the coefficient (Hull et al. 1995). The evaluators found that the alpha coefficient from fully crossed reliability study designs for Objective Structured Clinical Examinations (OSCE) and the then "National Board of Medical Examination" (NBME) was much lower than the monthly clinical evaluation form (CEF) when this value was obtained using a nested and unbalanced data-set. The alpha coefficients for the CEF "trait components" were >0.9, while those for the NBME were reported as 0.77 and for the OSCE were 0.57 (range 0.38–0.69) for the average clinical skills assessment and 0.53 (range 0.36–0.64) for the average knowledge assessment. These observations provide an incisive but unrecognized insight by the authors into the problem of biased alpha reliability coefficients when used for nested and/or unbalanced evaluation designs.

Other studies claiming acceptable reliability used composite scores only (Kwolek et al. 1997; Kreiter et al. 1998; Beckman et al. 2006), and one accepted reliability coefficients under 0.70 as adequate (Kreiter et al. 1998). The method used to assess reliability in 2 of these evaluations increased the possibility of bias estimates and did not provide information about individual competencies (Kwolek et al. 1997; Beckman et al. 2006). For example, in an early study trainees were evaluated by different patterns of faculty raters from different rotations (Kwolek et al. 1997). An intra-class correlation and the Spearman–Brown formula were used to assess faculty reliability. An average of 7 faculty members evaluated each resident, using an evaluation form containing 10 specific performance items and an overall summary score (Kwolek et al. 1997). The inter-rater reliability of the mean overall performance rating of the evaluators was 0.82. The reliability of a single overall rating was 0.39 (Kwolek et al. 1997). For the mean overall performance rating 7 raters were needed for a reliability >0.80. Using a similar analysis with our data with a two-way mixed effects design and a Type C intra-class correlation coefficient (ICC) of the mean overall performance rating (Fleiss and Shrout 1978; Shrout and Fleiss 1979), the single measures ICC was 0.51 and the average measures ICC was 0.93; and using the lower bound 95% confidence interval for the average measures ICC, the number needed to achieve an ICC of 0.8, 0.7 and 0.6 were 4.6, 2.7 and 1.7 respectively. This result however is also biased and not a "reliable" reliability coefficient. However, two important observations from the study by Kwolek et al. were a lack of discrimination between the

competency skills assessed and that overall ratings were insensitive to performance deficiencies, similar to the observations in the current study and that by Hull et al. (1995).

Only one study reported variance components, and this evaluation showed only 11% variance attributable to the trainee for the overall mean score (Kreiter et al. 1998). This study used appropriate methods for a Generalisability G-study and D-study but did so for a mean composite score, not for individual competencies. The evaluators concluded that "The reliability of assigning students clerkship grades based on single CEFs (clinical evaluation forms) is unacceptably low. However, CEFs can accurately measure student's clerkship performances if completed by 3 or more raters", using lower reliability coefficients to estimate the NNIR (Kreiter et al. 1998). Our results using an overall computed mean score from all competencies gave a trainee variance component of 35.2%, which gives a reliability coefficient of 0.73 for 5 assessments. To achieve a reliability of 0.80 for a mean composite score, 7.4 assessments are needed. If only a mean composite score was of interest then the current process provides a reliable assessment format. However, all competencies are crucial for each individual practitioner, not just a composite score which could hide important deficiencies. Moreover such averaging will hide the major problems associated with response and other biases. This issue was highlighted by Williams et al. (2004): "These results suggest that a program director should strive to acquire not the average, but rather the maximum number of observations needed to consistently classify resident performance in the least stable performance area (in our case, professional behavior) in the year with the lowest reliabilities. Using this approach, our results suggest that 38 observations of each resident's performance (approximately 3 per month) will ensure a stable classification of competence (reliability coefficient of .80)".

Fully crossed designs are more likely to provide unbiased reliability coefficients for the evaluation of the reliability of rater-based competency assessment. The nature of supervisor assessments make it difficult to undertake fully crossed studies. However, examples of studies that have used fully-crossed designs show that rater-based assessments are problematic in other types of competency assessment methods (Keller et al. 2000; van Barneveld 2005) with similar issues as illustrated for the mini-CEX (Margolis et al. 2006). Over a decade ago Turnbull et al. lamented the problem of poor reliability of clinical clerkship assessments and attempted to improve on the process (Turnbull et al. 2000). When applying a crossed design to a subpopulation of the data from the study by Turnbull et al. (2000) in order to achieve unbiased reliability estimates, van Barneveld demonstrated the biased nature of reliability estimates derived from methodology not suitable for separating the variance components in a nested and unbalanced data-collection design (van Barneveld 2005). Only 4% of total variance for the scores was attributable to differences between the trainees. Keller et al. (2000) also used an evaluation format involving a crossed design with low likelihood of having biased estimates as a primary analysis of all assessments, that is a full sample set. They also included the competency items within the design rather than evaluating individual competencies separately. This study involving 4 expert raters in the examination of 200 medical students on 16 performance items with a 9-point scale related to a computer simulation demonstrated only a 14.9% variance for score differences between medical students. Error variance, that is, variance due to differences in the scores between trainees not due to differences in the trainees, accounted for 85.1% of the total variance, again similar to our observations. These studies document the problem of method bias for rater-based assessments of competency and also highlight the usefulness of reporting variance components as recommended by Cronbach (Cronbach and Shavelson 2004).

Sensitivity analysis

While no significant differences in the pattern of variance components for the competencies are observed in the sensitivity analysis generally, the small differences observed in the least reliably assessed competencies further supports the concept that some competencies cannot be assessed by supervisors. Although the differences were not significant, possibly because of the sample size, they do highlight the importance of investigating causes of potential confounders (latent variables) that influence the reliability of the assessment, in this case the previously identified potential for differences in reliability between the years of sampling (Williams et al. 2004). Furthermore, the overall pattern of variance for the different competencies remains the same, further highlighting the problems with using averaging of competency scores and the loss of information about supervisor ratings of individual competencies.

In summary, for workplace data-sets the variance component for the trainee effect should be the focus of reliability assessments for every competency item thought to be important, in conjunction with the identifying number of assessments needed to improve each competency assessment's reliability within a local context. This form of evaluation may lead to a process that more easily identifies local influences that inappropriately reduce the variance due to the trainee, and provide a basis to test the efficacy of any proposed interventions to improve supervisor assessments. Moreover, these results from a small local evaluation combined with information from previous research again suggest that a marked improvement in rater based assessments in general is still required (Thorndike 1920; Cronbach 1951; Levine and McGuire 1971; Kroboth et al. 1992; Williams et al. 2003; Sadler 1989). The current findings for the reliability of supervisor assessments indicate that when the variance of trainees and the number needed to achieve adequate reliability are used as the measure of reliability (or utility), the supervisors in this study obtain at least the same outcomes as past studies and those using the mini-CEX. This study emphasises the importance of having an adequate number of assessments for any workplace performance evaluation using rater judgements, highlighting the ongoing need for improvement recognised 40 years ago (Levine and McGuire 1971).

Conclusions

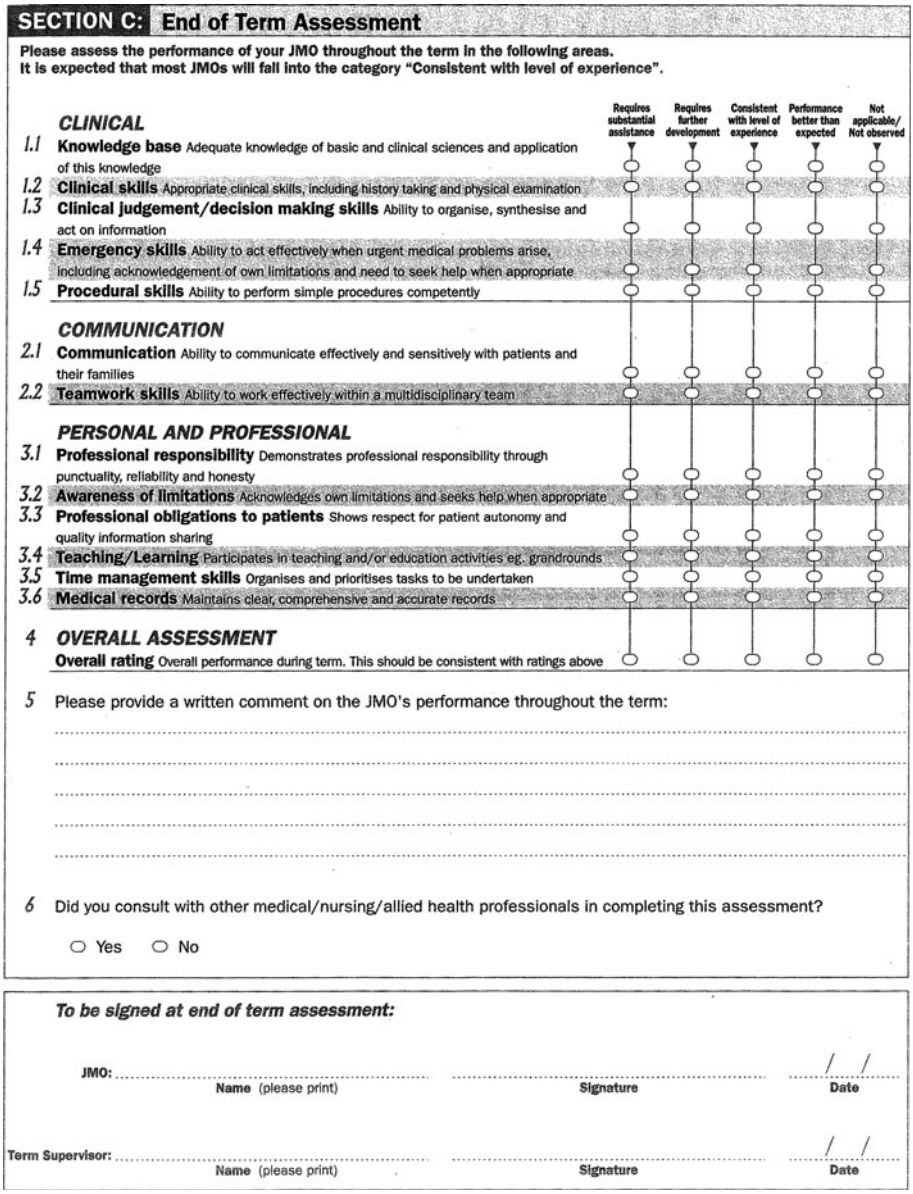
Ideally the variance within any assessment method should be predominantly related to differences between those being assessed. Score variation between trainees is low for end-of-term workplace assessments for the population and context studied. The amount of variation attributable to error is unacceptably large. The number of supervisors' assessments needed to improve reliability in general is also unacceptably large for nearly all competencies. The ability of supervisors to identify some competence constructs is particularly poor.

Continued use of this format for assessment of trainee competencies necessitates the identification of what supervisors' in different institutions can reliably assess rather than continuing to impose false expectations from unreliable assessments. The methodology used in this study allows institutions to estimate the number of assessments needed for minimally acceptable reliability for each competency in their institution with their particular assessors for any assessment process relying on raters judging others' competency in the workplace. If the reliability and sample number is found to be unacceptable, then alternative ways of assessing that competency should be used and validated as part of an assessment program.

The need to establish assessment programs rather than relying on individual types of assessment in the workplace is highlighted by these results. Similarly the need for systematic

evaluation of any workplace based assessment process should be considered mandatory. A process of improvement can then be initiated for any rater-based assessment method chosen to be part of an assessment program. Thus a cycle of evaluation, identification of problems, development of improvement interventions and re-evaluation can be instituted. The simple methodology used in this study provides one potentially useful approach for evaluation of workplace assessment data-sets with nested and unbalanced designs.

Appendix



SECTION C: End of Term Assessment

Please assess the performance of your JMO throughout the term in the following areas.
It is expected that most JMOs will fall into the category "Consistent with level of experience".

		Requires substantial assistance	Requires further development	Consistent with level of experience	Performance better than expected	Not applicable/ Not observed
	CLINICAL					
1.1	**Knowledge base** Adequate knowledge of basic and clinical sciences and application of this knowledge	○	○	○	○	○
1.2	**Clinical skills** Appropriate clinical skills, including history taking and physical examination	○	○	○	○	○
1.3	**Clinical judgement/decision making skills** Ability to organise, synthesise and act on information	○	○	○	○	○
1.4	**Emergency skills** Ability to act effectively when urgent medical problems arise, including acknowledgement of own limitations and need to seek help when appropriate	○	○	○	○	○
1.5	**Procedural skills** Ability to perform simple procedures competently	○	○	○	○	○
	COMMUNICATION					
2.1	**Communication** Ability to communicate effectively and sensitively with patients and their families	○	○	○	○	○
2.2	**Teamwork skills** Ability to work effectively within a multidisciplinary team	○	○	○	○	○
	PERSONAL AND PROFESSIONAL					
3.1	**Professional responsibility** Demonstrates professional responsibility through punctuality, reliability and honesty	○	○	○	○	○
3.2	**Awareness of limitations** Acknowledges own limitations and seeks help when appropriate	○	○	○	○	○
3.3	**Professional obligations to patients** Shows respect for patient autonomy and quality information sharing	○	○	○	○	○
3.4	**Teaching/Learning** Participates in teaching and/or education activities eg. grandrounds	○	○	○	○	○
3.5	**Time management skills** Organises and prioritises tasks to be undertaken	○	○	○	○	○
3.6	**Medical records** Maintains clear, comprehensive and accurate records	○	○	○	○	○
4	***OVERALL ASSESSMENT***					
	Overall rating Overall performance during term. This should be consistent with ratings above	○	○	○	○	○

5 Please provide a written comment on the JMO's performance throughout the term:

.....

.....

.....

.....

.....

6 Did you consult with other medical/nursing/allied health professionals in completing this assessment?

○ Yes ○ No

To be signed at end of term assessment:

JMO: .... / /
Name (please print) **Signature** **Date**

Term Supervisor: .... / /
Name (please print) **Signature** **Date**

References

Accreditation Council for Graduate Medical Education (ACGME). (2000). *ACGME/ABMS joint initiative toolbox of assessment methods version 1.1 September 2000* http:\\www.abim.org (Accesed 7th March 2007): Accreditation Council for Graduate Medical Education and American Board of Medical Specialties.

Albanese, M. A., Mejicano, G., Mullan, P., Kokotailo, P., & Gruppen, L. (2008). Defining characteristics of educational competencies. *Medical Education, 42*, 248–255.

Baltagi, B. H., Song, S. H., & Jung, B. C. (2002). A comparative study of alternative estimators for the unbalanced 2-way error component regression model. *Econometrics Journal, 5*, 480–493.

Beckman, T. J., Cook, D. A., & Mandrekar, J. N. (2006). Factor instability of clinical teaching assessment scores among general internists and cardiologists. *Medical Education, 40*, 1209–1216.

Carline, J. D., Wenrich, M., & Ramsey, P. G. (1989). Characteristics of ratings of physician competence by professional associates. *Evaluation & the Health Professions, 12*, 409–423.

Cohen, S. N., Farrant, P. B., & Taibjee, S. M. (2009). Assessing the assessments: UK dermatology trainees' views of the workplace assessment tools. *British Journal of Dermatology, 161*, 34–39.

Cook, D. A., Dupras, D. M., Beckman, T. J., Thomas, K. G., & Pankratz, V. S. (2008). Effect of rater training on reliability and accuracy of mini-CEX scores: A randomized, controlled trial. *Journal of General Internal Medicine, 24*, 74–79.

Cooper, W. H. (1981). Ubiquitous halo. *Psychological Bulletin, 90*, 218–244.

Cowles, J. T., & Kubany, A. J. (1959). Improving the measurement of clinical performance of medical students. *Journal of Clinical Psychology, 15*, 139–143.

Cronbach, L. J. (1950). Further evidence on response sets and test design. *Educational and Psychological Measurement, 10*, 3–31.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Pyschometrica,* pp. 297–333.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability of scores and profiles*. New York: Wiley.

Cronbach, L. J., & Shavelson, R. J. E. (2004). My current thoughts on coefficient alpha and successor procedures. *Educational and Psychological Measurement, 64*, 391–418.

Crossley, J., Russell, J., Jolly, B., Ricketts, C., Roberts, C., Schuwirth, L., et al. (2007). 'I'm pickin' up good regressions': the governance of generalisability analyses. *Medical Education, 41*, 926–934.

Davis, J. K., Inamdar, S., & Stone, R. K. (1986). Interrater agreement and predictive validity of faculty ratings of pediatric residents. *Journal of Medical Education, 61*, 901–905.

de Lima, A. A., Barrero, C., Baratta, S., Costa, Y. C., Bortman, G., Carabajales, J., et al. (2007). Validity, reliability, feasibility and satisfaction of the mini-clinical evaluation exercise (Mini-CEX) for cardiology residency training. *Medical Teacher, 29*, 785–790.

Dickinson, T. L., & Tice, T. E. (1977). The discriminant validity of scales developed by retranslation. *Personnel Psychology, 30*, 217–228.

Downing, S. M. (2004). Reliability: On the reproducibility of assessment data. *Medical Education, 38*, 1006–1012.

Dudek, N. L., Marks, M. B., & Regehir, G. (2005). Failure to fail: The perspectives of clinical supervisors. *Academic Medicine, 80*, S84–S87.

Durning, S. J., Pangaro, L. N., Lawrence, L. L., Waechter, D., McManigle, J., & Jackson, J. L. (2005). The feasibility, reliability, and validity of a program director's (supervisor's) evaluation form for medical school graduates. *Academic Medicine, 80*, 964–968.

Fleiss, J. L., & Shrout, P. E. (1978). Approximate interval estimation for a certain intraclass correlation coefficient. *Psychometrika, 43*, 259–262.

Gleser, G. C., Cronbach, L. J., & Rajaratnam, N. (1965). Generalizability of scores influenced by multiple sources of variance. *Psychometrika, 30*, 395–418.

Govaerts, M. J. B. (2008). Educational competencies or education for professional competence? *Medical Education, 42*, 234–236.

Haber, R. J., & Avins, A. L. (1994). Do ratings on the American Board of Internal Medicine resident evaluation form detect differences in clinical competence? *Journal of General Internal Medicine, 9*, 140–145.

Hamdy, H., Prasad, K., Anderson, M. B., Scherpbier, A., Williams, R., Zwierstra, R., et al. (2006). BEME systematic review: Predictive values of measurements obtained in medical schools and future performance in medical practice. *Medical Teacher, 28*, 103–116.

Hess, J. W. (1969). A comparison of methods for evaluating medical student skill in relating to patients. *Journal of Medical Education, 44*, 934–938.

Hill, F., Kendall, K., Galbraith, K., & Crossley, J. (2009). Implementing the undergraduate mini-CEX: A tailored approach at Southampton University. *Medical Education, 43*, 326–334.
Holmboe, E. S., & Hawkins, R. E. (1998). Methods for evaluating the clinical competence of residents in internal medicine: A review. *Annals of Internal Medicine, 129*, 42–48.
Hull, A. L., Hodder, S., Berger, B., Ginsberg, D., Lindheim, N., Quan, J., et al. (1995). Validity of three clinical performance assessments of internal medicine clerks. *Academic Medicine, 70*, 517–522.
Hutchinson, L., Aitken, P., & Hayes, T. (2002). Are medical postgraduate certification processes valid? A systematic review of the published evidence. *Medical Education, 36*, 73–91.
James, R. J., Demnaree, R. G., & Wolf, G. (1984). Estimating within-group interrater reliability with and without response bias. *Journal of Applied Psychology, 69*, 85–98.
James, P. A., Kreiter, C. D., Shipengrover, J., & Crosson, J. (2002). Identifying the attributes of instructional quality in ambulatory teaching sites: A validation study of the MedEd IQ. *Family Medicine, 34*, 268–273.
Joint Committee on Standards for Educational, Psychological Testing of the American Educational Research Association. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
Kastner, L., Gore, E., & Novack, A. H. (1984). Pediatric residents' attitudes and cognitive knowledge, and faculty ratings. *The Journal of Pediatrics, 104*, 814–818.
Keck, J. W., & Arnold, L. (1979). Development and validation of an instrument to assess the clinical performance of medical residents. *Educational and Psychological Measurement, 39*, 903–908.
Kegel-Flom, P. (1975). Predicting supervisor, peer, and self-ratings of intern performance. *Journal of Medical Education, 50*, 812–815.
Keller, L. A., Mazor, K. M., Swaminathan, H., & Pugnaire, M. P. (2000). An investigation of the impacts of different generalizability study designs on estimates of variance components and generalizability coefficients. *Academic Medicine, 75*, S21–S24.
King, L. M., Schmidt, F. L., & Hunter, J. E. (1980). Halo in a multidimensional forced-choice evaluation scale. *Journal of Applied Psychology, 65*, 507–516.
Kogan, J. R., Holmboe, E. S., & Hauer, K. S. (2009). Tools for direct observation and assessment of clinical skills of medical trainees: A systematic review. *Journal of the American Medical Association, 302*, 1316–1326.
Koretz, D. (2003). Using multiple measures to address perverse incentives and score inflation. *Educational Measurement: Issues and Practice, 22*, 18–26.
Kreiter, C. D., & Ferguson, K. J. (2002). The empirical validity of straight-line responses on a clinical evaluation form. *Academic Medicine, 77*, 414–418.
Kreiter, C. D., Ferguson, K., Lee, W. C., Brennan, R. L., & Densen, P. (1998). A generalizability study of a new standardized rating form used to evaluate students' clinical clerkship performances. *Academic Medicine, 73*, 1294–1298.
Kreiter, C. D., James, P. A., Stansfield, R. B., & Callaway, M. R. (2002). An empirical validity study of a preceptor evaluation instrument. *Academic Medicine, 77*, S70–S72.
Kroboth, F. J., Hanusa, B. H., Parker, S., Coulehan, J. L., Kapoor, W. N., Brown, F. H., et al. (1992). The inter-rater reliability and internal consistency of a clinical evaluation exercise. *Journal of General Internal Medicine, 7*, 174–179.
Kwolek, C. J., Donnelly, M. B., Sloan, D. A., Birrell, S. N., Strodel, W. E., & Schwartz, R. W. (1997). Ward evaluations: Should they be abandoned? *Journal of Surgical Research, 69*, 1–6.
Latham, G. P., Wexley, K. N., & Pursell, E. D. (1975). Training managers to minimize rating errors in the observation of behavior. *Journal of Applied Psychology, 60*, 550–555.
Levine, H. G., & McGuire, C. H. (1971). Rating habitual performance in graduate medical education. *Academic Medicine, 46*, 306–311.
Magzoub, M. E. M. A., Schmidt, H. G., Abdel-Hameed, A. A., Dolmans, D., & Mustafa, S. E. (1998). Student assessment in community settings: A comprehensive approach. *Medical Education, 32*, 50–59.
Margolis, M. J., Clauser, B. E., Cuddy, M. M., Ciccone, A., Mee, J., Harik, P., et al. (2006). Use of the mini-clinical evaluation exercise to rate examinee performance on a multiple-station clinical skills examination: A validity study. *Academic Medicine, 81*, S56–S60.
Mazor, K. M., Zanetti, M. L., Alper, E. J., Hatem, D., Barrett, S. V., Meterko, V., et al. (2007). Assessing professionalism in the context of an objective structured clinical examination: An in-depth study of the rating process. *Medical Education, 41*, 331–340.
Metheny, W. P. P. (1991). Limitations of physician ratings in the assessment of student clinical performance in an obstetrics and gynecology clerkship. *Obstetrics and Gynecology, 78*, 136–141.
Miller, A., & Archer, J. (2010). Impact of workplace based assessment on doctors' education and performance: A systematic review. *British Medical Journal, 341*, c5064. doi:10.1136/bmj.c5064.

Murphy, K. R., & Balzer, W. K. (1989). Rater errors and rating accuracy. *Journal of Applied Psychology, 74*, 619–624.

Nasca, T. J., Gonnella, J. S., Hojat, M., Veloski, J., Erdmann, J. B., Robeson, M., et al. (2002). Conceptualization and measurement of clinical competence of residents: A brief rating form and its psychometric properties. *Medical Teacher, 24*, 299–303.

Norcini, J. J., Blank, L. L., Arnold, G. K., & Kimball, H. R. (1995). The mini-CEX (clinical evaluation exercise): A preliminary investigation. *Annals of Internal Medicine, 123*, 795–799.

Norcini, J. J., Blank, L. L., Duffy, F. D., & Fortna, G. S. (2003). The mini-CEX: A method for assessing clinical skills. *Annals of Internal Medicine, 138*, 476–481.

Pulito, A. R., Donnelly, M. B., & Pylmale, M. (2007). Factors in faculty evaluation of medical students' performance. *Medical Education, 41*, 667–675.

Remmers, H. H., Shock, N. W., & Kelly, E. L. (1927). An empirical study of the validity of the Spearman-Brown formula as applied to the Purdue rating scale. *The Journal of Educational Psychology, 18*, 187–195.

Ronan, W. W., & Prien, E. P. (1966). *Toward a criterion theory: A review of research and opinion*. Greensboro, NC: Creativity Research Institute, Smith Richardson Foundation.

Ronan, W. W., & Prien, E. P. (1971). *Perspectives on the measurement of human performance*. New York: Appleton Century Crofts.

Rothstein, R. H. (1990). Interrater reliability of job performance ratings: Growth to asymptote level with increasing opportunity to observe. *Journal of Applied Psychology, 75*, 322–327.

Ryan, J. G., Mandel, F. S., Sama, A., & Ward, M. F. (1996). Reliability of faculty clinical evaluations of non-emergency medicine residents during emergency department rotations. *Academic Emergency Medicine, 3*, 1124–1130.

Saal, F. E., Downey, R. G., & Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin, 88*, 413–428.

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science, 18*, 119–144.

Sadler, D. R. (2005). Interpretations of criteria-based assessment and grading in higher education. *Assessment & Evaluation in Higher Education, 30*, 175–194.

Sadler, D. R. (2009). Grade integrity and the representation of academic achievement. *Studies in Higher Education, 34*, 807–826.

Schwanz, R. W., Donnelly, M. B., Sloan, D. A., Johnson, S. B., & Strodel, W. E. (1995). The relationship between faculty ward evaluations, OSCE, and ABSITE as measures of surgical intern performance. *The American Journal of Surgery, 169*, 414–417.

Searle, G. F. (2008). Is CEX good for psychiatry? An evaluation of workplace-based assessment. *Psychiatric Bulletin, 32*, 271–273.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*, 420–428.

Speer, A. J., Solomon, D. J., & Fincher, R.-M. E. (2000). Grade inflation in internal medicine clerkships: Results of a national survey. *Teaching and Learning in Medicine, 12*, 112–116.

Streiner, D. L. (1995). Clinical ratings—ward rating. In S. Shannon & G. Norman (Eds.), *Evaluation methods: A resource handbook* (pp. 29–32). Hamilton: Program for Educational Development McMaster University.

Streiner, D. L., & Norman, G. R. (2009). *Health measurement scales. A practical guide to their development and use* (4th ed.). Oxford: Oxford University Press.

Swanson, D. B., Norman, G. R., & Linn, R. L. (1995). Performance-based assessment: Lessons from the health professions. *Educational Researcher, 24*, 5–11–35.

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). Boston: Pearson Allyn and Bacon.

Thompson, W. G., Lipkin, M, Jr., Gilbert, D. A., Guzzo, R. A., & Roberson, L. (1990). Evaluating evaluation: assessment of the American Board of Internal Medicine Resident Evaluation Form. *Journal of General Internal Medicine, 5*, 214–217.

Thorndike, E. L. (1920). A constant error in psychological ratings. *Journal of Applied Psychology, 4*, 25–29.

Turnbull, J., MacFadyen, J., van Barneveld, C., & Norman, G. (2000). Clinical work sampling: A new approach to the problem of in-training evaluation. *Journal of General Internal Medicine, 15*, 556–561.

van Barneveld, C. (2005). The dependability of medical students' performance ratings as documented on in-training evaluations. *Academic Medicine, 80*, 309–312.

van der Vleuten, C. P. M., Scherpbier, A. J. J. A., Dolmans, D. H. J. M., Schuwirth, L. W. T., Verwijnen, G. M., & Wolfhagen, H. A. P. (2000). Clerkship assessment assessed. *Medical Teacher, 22*, 592–600.

van der Vleuten, C. P., & Schuwirth, L. W. (2005). Assessing professional competence: From methods to programmes. *Medical Education, 39*, 309–317.

Viswesvaran, C., Ones, D. S., & Schmidt, F. L. (1996). Comparative analysis of the reliability of job performance ratings. *Journal of Applied Psychology, 81*, 557–574.

Viswesvaran, C., Schmidt, F. L., & Ones, D. S. (2005). Is there a general factor in ratings of job performance? A meta-analytic framework for disentangling substantive and error influences. *Journal of Applied Psychology, 90*, 108–131.

Wainer, H., & Thissen, D. (1996). How is reliability related to the quality of test scores? What is the effect of local dependence on reliability? *Educational Measurement: Issues and Practice, 15*, 22–29.

Wass, V., Van der Vleuten, C., Shatzer, J., & Jones, R. (2001). Assessment of clinical competence. *The Lancet, 357*, 945–949.

Weller, J. M., Jolly, B., Misur, M. P., Merry, A. F., Jones, A., Crossley, J. G., et al. (2009). Mini-clinical evaluation exercise in anaesthesia training. *British Journal of Anaesthesia, 102*, 633–641.

Wherry, S., & Bartlett, C. J. (1982). The control of bias in ratings: A theory of rating. *Personnel Psychology, 35*, 521–551.

Wilkinson, J. R., Crossley, J. G., Wragg, A., Mills, P., Cowan, G., & Wade, W. (2008). Implementing workplace-based assessment across the medical specialties in the United Kingdom. *Medical Education, 42*, 364–373.

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*, 270–292.

Williams, R. G., Verhulst, S., Colliver, J. A., & Dunnington, G. L. (2004). Assuring the reliability of resident performance appraisals: More items or more observations? *Surgery, 137*, 141–147.