Global overall rating for assessing clinical competence: what does it really show?

Rosângela C L Domingues,[1] Eliana Amaral[1] & Angélica M B Zeferino[2]

CONTEXT Single-item global overall ratings are often used as a method of assessing learners' clinical competence at the end of a clerkship. The purpose of this study was to identify which aspects of clinical competence are assessed through these ratings.

METHODS At the end of a clinical clerkship in primary health units, 106 Year 4 students are routinely assessed by faculty staff of three disciplines (obstetrics and gynaecology, internal medicine, paediatrics), using a single global numeric rating (on a scale of 0–10). Faculty scores across disciplines for each learner are averaged to produce a global overall rating (GOR). In this study, the same students were assessed by the same faculty staff 2 weeks later using a newly developed, more detailed form composed of 13 domains, of which six related to technical skills and seven to humanistic skills, each scored on a scale of 0–10. Scores for each domain across disciplines were averaged as global itemised ratings (GIRs). Statistical analysis included Cronbach's α coefficient and Pearson's correlation coefficients. Statistical significance was set at $P \leq 0.05$.

RESULTS The internal consistency of GIR items was high (α coefficient = 0.935). Global overall rating scores were higher than most technical domains of GIRs and lower than the humanistic domains of GIRs. The highest significant correlations were found between the GOR and the technical domains of the GIR.

CONCLUSIONS When faculty staff attribute a global single-item overall rating to a student's clinical competence, they tend to focus more on technical skills.

*Medical Education 2009: **43**: 883–886*
doi:10.1111/j.1365-2923.2009.03431.x

[1]Department of Obstetrics & Gynaecology, State University of Campinas, Campinas, São Paulo, Brazil
[2]Department of Pediatrics, State University of Campinas, Campinas, São Paulo, Brazil

Correspondence: Rosângela Curvo Leite Domingues, Rua Luciano Venére Decourt 858, Campinas, São Paulo, 13083-740, Brazil. Tel: 00 55 19 3289 2182; Fax: 00 55 19 3289 8338; E-mail: rosangela@cpdec.com.br

INTRODUCTION

The assessment of medical students during clinical training is a complex undertaking. Among the different methods used to assess clinical competence, a single-item global performance rating given at the end of a clinical clerkship continues to be commonly employed.[1] The comparison between a more general single-item rating with a multiple domain rating will provide information that can lead to a better understanding of what is being measured in the assessment. As clinical performance requires the consideration of a variety of specific competencies, it is important to identify which of these is captured by this score.[2]

The present study sought to answer the following research questions: what is the relationship between a single-item rating and a specific multiple-item rating, and which competencies do faculty members really focus on when assessing through a single score?

METHODS

In the institution in which the study was conducted, the medical curriculum lasts 6 years. The clinical phase starts in Year 4 with a 9-month clinical clerkship in which students undertake approximately 200 hours of supervised clinical practice at the primary level of care in three disciplines: obstetrics and gynaecology; internal medicine, and paediatrics. At the end of this period, the respective faculty staff provide a summative overall global rating, comprising a single-item score ranging from 0 to 10, which summarises the student's performance. Faculty overall global scores across the three disciplines are then averaged to form a single aggregated score for each student, defined as a global overall rating (GOR). The GOR has routinely represented one of the components of the learner's final assessment in the clerkship.

At the end of October 2005, 2 weeks after securing GOR ratings for all students, the same faculty members were asked to assess the same students using a newly developed rating form, the global itemised rating (GIR). This form encompassed 13 domains, of which six related to technical skills (quality of history, physical examination, medical knowledge, clinical judgement, problem-solving skills, work habits), and seven to humanistic skills (interpersonal and communication skills, respect for patients, self-reflection skills, compassion, relationships with peers, relationships with faculty members, relationships with staff or other health professionals). Items were selected on the basis of current literature[3] and on Year 4 learning objectives by a core panel of faculty members. Raters were given brief oral instructions before completing the questionnaire.

Faculty scores for each domain across disciplines were averaged and correlated to GOR. Reliability determined by the internal consistency across the 13-item form (GIR) was measured by Cronbach's α coefficient. Pearson's correlation coefficients were used to evaluate the statistical associations between the 13 domains of the GIR and GOR, and within the 13 domains of the GIR. Statistical significance was set at $P \leq 0.05$.

RESULTS

A total of 106 students were scored by 19 faculty members. Data were missing for at least one GIR item for three students, leaving 103 valid student scores. Each faculty member assessed a group of five to six learners. The mean age of the students was 22.8 years (standard deviation [SD] 0.2) and 50% were female. Most faculty staff were women (67%) and 58% were aged > 40 years.

Mean scores were high on the GOR (8.85, SD 0.58) and for all domains on the GIR (ranging from 8.43, SD 0.72 for *Medical knowledge* to 9.57, SD 0.48 for *Respect for patients*). Scores on the GOR were higher than all GIR technical domain scores, except those for *Quality of history* and *Work habits*, and lower than all GIR humanistic domain scores excluding that for *Self-reflective skills*. Within the GIR, humanistic domains were scored higher than technical domains (Table 1).

The internal consistency of the GIR form was high, with an α-coefficient of 0.935. We found positive and significant correlations between the GOR and each of the 13 items of the GIR, with the strongest correlations between the GOR and the technical items (Table 1). Positive and significant correlations between the 13 GIR domains were also found. The highest correlations were observed among the technical domains ($0.75 < r < 0.84$), except for *Work habits*. Among the humanistic domains the correlations ranged from 0.51 to 0.76 (data not shown).

DISCUSSION

In this study, faculty assigned higher ratings on the GOR and on the humanistic domains of the GIR,

Table 1 Mean, standard deviation, and Pearson correlations (r) between GOR and GIR ratings (n=103)

GOR	Mean	SD	r*
	8.85	0.580	
GIR for technical skills			
1. Quality of history	8.73	0.770	0.804
2. Physical examination	8.67	0.732	0.722
3. Medical knowledge	8.43	0.722	0.805
4. Clinical judgement	8.43	0.752	0.835
5. Problem-solving	8.60	0.764	0.808
6. Work habits	9.28	0.579	0.594
GIR for humanistic skills			
7. Interpersonal and communication skills	9.26	0.610	0.583
8. Respect to patients	9.57	0.482	0.314
9. Self-reflective skills	9.03	0.676	0.618
10. Compassion	9.12	0.635	0.600
11. Relationship with peers	9.34	0.532	0.398
12. Relationship with faculty	9.48	0.437	0.424
13. Relationship with staff/ other health professionals	9.13	0.547	0.470

* All values are significant (p<0.05).

compared with technical domains. This finding is consistent with the results of other studies, in which the mean ratings of students were significantly higher for humanistic aspects of clinical competence than for technical aspects.[4] There are some possible explanations for these results. Firstly, there seems to be a tendency among faculty to simplify humanistic skills and overemphasise the importance of technical skills, particularly at this phase of clinical training. Secondly, faculty may be somewhat lenient when assessing humanistic skills. These skills are not only more difficult to assess, but scores are also harder to justify. Faculty staff may feel more uncomfortable about having to explain low scores for personal qualities than they do when they are required to point out and explain to a student that he or she needs to improve his or her technical skills.[5] Thirdly, assessors in our study may have had too high a level of expectation of learners' technical skills, which led them to be stricter when evaluating these domains. Finally, as the learners were novice clinicians, they may have demonstrated a lack of technical skills appropriate to clinical practice.

There was a significant correlation between GOR and GIR scores. The fact that faculty assigned both GOR and GIR scores may partially explain this correlation (memory bias). However, as GIR forms were filled in 2 weeks after GOR forms, the effects of such lack of independence between these two assessment methods may have been reduced.[4] The strong correlations between GOR and GIR scores on technical domains provides some evidence for the supposition that faculty staff focus more on technical than humanistic skills when they attribute a GOR score.[4]

With a single overall rating, faculty staff tend to assign grades based on their own weighting systems, which may cause some rating bias if one specific skill is favoured. This tendency also compromises the meaningfulness of detailed feedback given to students. By contrast, gathering technical and humanistic ratings may help to avoid such bias, as well as the likelihood that a learner's strengths in one area may compensate for deficiencies in another.

Our results led us to conclude that when faculty staff wish to express a global impression of learners, technical domains appear to prevail. The use of an instrument that gathers ratings on technical and humanistic domains and the training of raters are recommended.

Contributors: RCLD participated in the development of the project, including the instrument and its implementation, prepared the draft manuscript and incorporated the contributions of the co-authors. EA and AMBZ contributed to the development and implementation of the project. EA also made significant contributions to the conceptualisation and presentation of this project. All authors contributed to the editing of and gave final approval to the manuscript accepted for publication.
Acknowledgements: the authors appreciate the statistical support and advice of Sirlei Siani Morais and Armando Mario Infante, and previous version inputs from Vanessa Burch and John Boulet.
Funding: none.
Conflicts of interest: none.
Ethical approval: this study was approved by the Ethical and Research Committee of the State University of Campinas (# 581/2005).

REFERENCES

1 Daelmans HEM, van der Hem-Stokroos HH, Hoogenboom RJL, Scherpbier AJJA, Stehouwer CDA, van der Vleuten CPM. Global clinical performance rating,

reliability and validity in undergraduate clerkship. *Neth J Med* 2005;**63**:279–84.
2 Solomon DJ, Szauter K, Rosebraugh CJ, Callaway MR. Global ratings of student performance in a standardised patient examination: is the whole more than the sum of the parts? *Adv Health Sci Educ Theory Pract* 2000;**5**:131–40.
3 Ramsey PG, Wenrich MD, Carline JD, Inui TS, Larson EB, LoGerfo JP. Use of peer ratings to evaluate physician performance. *JAMA* 1993;**269**:1655–60.
4 Morgan PJ, Cleave-Hogg D, Guest CB. A comparison of global ratings and checklist scores from an undergraduate assessment using an anaesthesia simulator. *Acad Med* 2001;**76**:1053–5.
5 Pulito AR, Donnelly MB, Plymale M. Factors in faculty evaluation of medical students' performance. *Med Educ* 2007;**41**:667–75.

Received 12 December 2008; editorial comments to authors 12 February 2009, 8 May 2009; accepted for publication 28 May 2009