Global clinical performance rating, reliability and validity in an undergraduate clerkship

H.E.M. Daelmans[1*], H.H. van der Hem-Stokroos[2], R.J.I. Hoogenboom[3], A.J.J.A. Scherpbier[4], C.D.A. Stehouwer[5,6], C.P.M. van der Vleuten[3]

Departments of [1]Skills Training, [2]Surgery and [6]Medicine, Vrije universiteit Medical Centre, Amsterdam, the Netherlands, tel.: +31 (0)20-444 80 31, fax: +31 (0)20-444 80 30, e-mail: hem.daelmans@vumc.nl, [3]Department of Educational Development and Research, [4]Faculty of Medicine/Institute for Medical Education, Maastricht University, the Netherlands, [5]Department of Medicine, Maastricht University Hospital and Maastricht University, the Netherlands, [*]corresponding author

ABSTRACT

Background: Global performance rating is frequently used in clinical training despite its known psychometric drawbacks. Inter-rater reliability is low in undergraduate training but better in residency training, possibly because residency offers more opportunities for supervision. The low or moderate predictive validity of global performance ratings in undergraduate and residency training may be due to low or unknown reliability of both global performance ratings and criterion measures. In an undergraduate clerkship, we investigated whether reliability improves when raters are more familiar with students' work and whether validity improves with increased reliability of the predictor and criterion instrument.

Methods: Inter-rater reliability was determined in a clerkship with more student-rater contacts than usual. The in-training assessment programme of the clerkship that immediately followed was used as the criterion measure to determine predictive validity.

Results: With four ratings, inter-rater reliability was 0.41 and predictive validity was 0.32. Reliability was lower and validity slightly higher than similar results published for residency training.

Conclusion: Even with increased student-rater interaction, the reliability and validity of global performance ratings were too low to warrant the usage of global performance ratings as individual assessment format. However, combined with other assessment measures, global performance ratings may lead to improved integral assessment.

KEYWORDS

Clerkship, disattenuation, global clinical performance rating, inter-rater reliability, predictive validity

INTRODUCTION

Evaluation of clinical performance typically takes the form of a global rating by a supervisor, halfway or at the end of a clinical rotation, covering learners' performance on a number of clinically relevant competencies over a certain period of time. Hereafter, we will refer to this type of rating as global performance rating (GPR). Despite the availability of new assessment methods, GPRs continue to be frequently used in both undergraduate and residency training, most probably due to the combined advantage of feasibility and face validity (the assessed performance represents the performance domain of interest). In undergraduate training, GPRs are often the primary determinant of the final grade a student receives at the end of a clerkship.[1,2] Moreover, despite measures to increase the reliability of GPRs, such as rater training, in practice most assessors are not trained. At best the items on a scale are anchored to descriptors of criterion behaviour. In the last two decades several studies have examined the reliability and validity of GPRs by untrained assessors in both undergraduate and residency training.

For inter-rater agreement among members of staff as a measure of the reliability of GPRs in undergraduate training the findings varied, with inter-rater agreement ranging from 0.29-0.42.[3,4] Studies performed in residency training have consistently demonstrated higher inter-rater agreement (0.79-0.87) than studies among undergraduate students.[5-8]

A possible explanation for this difference may be that clinical staff, who typically evaluate students' and residents' performance, generally have more opportunity to supervise the work of residents than that of students because residency rotations last longer than clerkship rotations.[9,10] Moreover, because residents treat patients, supervision of residents is necessary to ensure the provision of appropriate patient care. Staff members have a strong professional stake in an adequate performance by residents, because they may be held liable for adequate supervision.[11] Assuming that the reliability of assessment benefits from increased supervision, we designed a study in which we measured inter-rater agreement on GPRs in a setting where staff members supervised students' work more frequently than is customary in undergraduate clerkships. If our assumption is correct, we would expect inter-rater agreement in this setting to be moderate to quite high.

Studies have investigated both concurrent and predictive validity of GPRs by untrained assessors. Both concurrent and predictive validity indicate the extent to which GPRs predict scores on a selected criterion that is not directly measured by the assessment but that is assumed to be parallel. For concurrent validity the criterion measurement is performed at the same time, for predictive validity it is performed at some point in the future. *Concurrent validity* has been studied in both undergraduate and residency training by correlating GPRs with more objective performance measures for the same training period, such as written examinations, OSCEs or simulated patient exams.[6,12-14] Correlations ranged from 0.19 to 0.33 for undergraduate training and from 0.29 to 0.56 for residency training. Fewer studies have addressed the *predictive validity* of GPRs. In undergraduate training predictive validity has been determined by comparing GPRs of student performance in different rotations with GPRs of end-of-clerkship performance or of performance in residency training. Predictive validity ranged from 0.17 to 0.44.[12,15] Callahan examined the predictive validity of GPRs in clerkships for the results on United States Medical Licensing Examinations (USMLE) steps 2 and 3. The maximum predictive validity was 0.29 for USMLE step 2 and for USMLE step 3 it was 0.20.[15] In residency training the predictive validity of GPRs for performance at in-training and American Board of Internal Medicine certification exams was reported to be moderate for overall competence (0.19) and for specific competencies on a global performance rating scale (ranging from 0.11 to 0.41).[8,16] The validity coefficients reported in these studies suffered from attenuation, i.e. low or unknown reliabilities in predictor and criterion variables introduce inaccuracy into the calculation. When a measurement error is present in one or both of the variables that are being correlated, the correlation coefficient that is obtained will be attenuated. This implies that the observed correlation coefficient between less than perfectly reliable scores will tend to underestimate the true level of co-variation between the predictor and criterion variables.[17] Therefore, if the reliability of either the predictor or the criterion variables is low, validity coefficients will also be low. This effect might have been even stronger in studies in undergraduate training settings, where the reliability of global ratings was typically low and the reliability of the criterion variable mostly unknown. That is why we considered it worthwhile to examine the predictive validity of GPRs in undergraduate training using a criterion variable of known and acceptable reliability. If staff members have more opportunities to supervise students' performance the reliability of GPRs in undergraduate medical training might benefit.

We thus sought to answer the following research questions: what is the reliability of GPRs in an undergraduate clerkship with increased rater-student interactions? And, because of the inaccuracy in validity estimates of GPRs due to the low or unknown reliabilities of predictor and criterion variables (attenuation): what is the validity of GPRs when the reliabilities of both predictor and criterion variables are assumed to be perfect (disattenuation correction)?

We addressed the research questions by determining the inter-rater agreement for GPRs in an undergraduate clerkship with extensive interaction (detailed below) of staff members (raters) and students. These GPRs were then compared with a valid and reliable performance measure for the competencies demonstrated by the same students in the next clerkship that immediately followed, with less student-staff interaction. The performance measure used in the second rotation was the overall score on an in-training assessment programme (ITA) consisting of several assessments of clinical competence.

MATERIALS AND METHODS

Educational background

At the Vrije Universiteit Medical Centre (VU Medical Centre), Amsterdam, the Netherlands, four years of preclinical medical education are followed by two years of rotations in the major clinical disciplines. The clinical phase starts in year 5 with a three-week introductory clerkship in which the students are closely supervised by clinical staff. The students' main tasks are history taking and physical examination, medical record writing and practising skills in pathophysiological thinking and clinical reasoning in structured discussions both in groups supervised by a member of staff and in writing. Staff members are scheduled to supervise the group discussion, discuss the medical records and observe (parts of) history taking and physical examination. Every day a different staff member supervises the students in their daily structured group discussion, which lasts about an hour. The supervisor asks several students to elaborate on their findings, interpret data, formulate differential diagnoses and propose additional investigations. Over the course of the clerkship, students are supervised twice by a member of staff while performing a (scheduled) complete patient interview and physical examination. Afterwards student and staff member discuss the student's performance. For morning and evening reports, radiology meetings, interdisciplinary meetings et cetera, students are not linked to residents but to members of staff, who are thus more focussed on students' contributions. In most cases, a student is supervised by six to seven members of staff during the three-week rotation. At the end of the three weeks, a staff member to whom this task has been assigned determines a GPR for the student's performance during the rotation. Next, the students move on to the ten-week internal medicine rotation in the VU Medical Centre or in one of the affiliated hospitals. This rotation is scheduled immediately after the introductory clerkship. In this rotation the students are usually supervised by residents instead of members of staff during student-patient interactions and for medical records. This rotation involves more participation by the students in day-to-day clinical practice, including multidisciplinary meetings and on call duties. In order to better monitor students' performance in the internal medicine clerkship, a programme of systematic observation and documentation of students' actual performance (detailed below) has been introduced in the rotation in the VU Medical Centre.

Global clinical performance rating

Eight supervisors of the introductory clinical rotation were approached during a staff meeting and asked to participate in the study. Participation entailed giving a global performance rating on a five-point Likert scale (1 = fail, 2 = borderline, 3 = pass, 4 = high pass, 5 = excellent) for every student doing the rotation in the study period. The raters received a brief description to be used in rating students' performance. The description mainly focused on the comparison of the student's performance with that of an average student in his/her first three weeks of clinical rotation. On a student's last day of this rotation, the members of staff received a form together with a scanned picture of the student concerned. The members of staff who had interacted with the student were asked to complete and return the form. The members of staff who had not supervised the student could return the form without filling it in. The entire procedure was computerised. We used a single-item rating (global performance) to preclude the use by raters of only one or two items (dimensions of performance) of a larger scale to judge global performance.[5,6,16,18,19] Each participating staff member was asked to complete one evaluation form for each student during the study period. In this way students could receive a maximum of eight GPRs from different examiners.

In-training assessment

All students proceed from the introductory rotation to the internal medicine rotation. In the internal medicine rotation in the VU Medical Centre a fully integrated ITA programme is used. ITA implies systematic observation and documentation of the learners' actual performance using several formats.[20] ITA in undergraduate clinical training has been described as a feasible assessment format that has reasonable reliability and good content validity.[20-22] The ITA programme used in this study consisted of observation and documentation of students' actual performance in five test formats.[22] A minimum frequency per student over the entire clerkship was specified for each test format, resulting in a required total of 19 assessments: three single-sample formats (student-patient encounter, critical appraisal session and case presentation) and two multiple-sample formats (12 case write-ups and four structured long cases). The student-patient encounter, critical appraisal session, case presentation and structured long cases were assessed by staff and the case write-ups by residents. All tests were rated on the five-point Likert scale and an overall score was obtained by calculating the mean of the scores and rounding it off to the nearest integer (1-5). The assessors of the ITA programme were not specifically informed about the current study.

Subjects: student participants

From April 2001 to October 2002, 91 students received global ratings of their performance in the introductory clerkship. We collected ITA scores for 48 of these 91 students. These 48 students did the subsequent ten-week internal medicine rotation in the VU Medical Centre, whereas the other students went to affiliated hospitals, where the ITA programme had not yet been implemented. A t-test on the means of the GPRs showed no differences between the GPRs of the students participating in the study and the students assigned to the affiliated hospitals.

Data analysis

First we counted the total number of GPRs per student. For further calculations we used the balanced dataset of the group of students for whom at least four GPRs were available.[23] For the analysis we used random samples of four GPRs per student.

We calculated means and standard deviations for the GPRs. Inter-rater reliability was estimated based on the generalisability theory. We used a one-facet design with raters (or GPRs) nested within persons (students) to estimate variance components. Subsequently, reliability coefficients, i.e. dependability coefficients, were calculated as a function of the number of examiners (or GPRs). In the clerkship studied, each member of staff supervised students in two to three group discussions and most probably several times during reports and meetings. Only two members of staff witnessed complete student-patient contacts (interview and physical examination). As a result, these two staff members may have developed significantly different judgements than did other staff members. However, the fact that different selections of four members of staff did not yield significantly different ratings suggests that this was not the case.

We calculated means and standard deviations of the ITA scores. A similar generalisability design was used to estimate the reliability of the ITA programme, with observations across test formats nested within students.
The predictive validity of the GPRs was determined by correlating the mean GPR with the mean ITA score. We estimated the disattenuated correlation (the estimated correlation when both predictor and criterion measures have perfect reliability) using the reliability coefficient of the GPRs with four raters and the reliability of the ITA programme.

RESULTS

Of the 91 students whose performance was rated in the introductory clerkship, 87 received four or more GPRs (*table 1*). Four students were rated by fewer than four staff members. Each of the eight members of staff who participated in the study contributed to the GPRs throughout the duration of the study. Having had no interaction with the student, holidays and illness were the main reasons given by staff members for not having witnessed a student's performance and thus being unable to provide a GPR.
The mean GPR was 3.19 (SD 0.37). Inter-rater reliability with four GPRs per student was 0.41 (n=87). Twenty-five GPRs per student would be needed to reach sufficient reliability (0.8) (*table 2*).

Table 1 *The number of global performance ratings (GPRs) per student*

Number of GPRs	2	3	4	5	6	7	8
Number of students	1	3	13	30	16	23	5

Table 2 *Reliability coefficients as a function of the number of examiners or global performance ratings (GPRs)*

Number of GPRs	4	6	10	25
Reliability coefficient	0.41	0.51	0.63	0.81

Means and standard deviations of the different ITAs ranged from 3.61 (SD 0.65) for case write-ups to 4.35 (SD 0.65) for case presentations (*table 3*).
The reliability of the ITA programme was 0.71. The observed predictive validity of the GPRs for the ITA programme was 0.32 (p<0.05) and the disattenuated predictive validity was 0.59.

Table 3 *Mean scores (standard deviations; SD) for the different tests in the in-training assessment (ITA) programme*

ITA	Mean (SD)
Student-patient encounter	3.70 (.70)
Critical appraisal session	4.00 (.67)
Case presentation	4.35 (.65)
Write-up	3.61 (.65)-4.04 (.71)
Structured long case	4.00 (.67)-4.26 (.63)

DISCUSSION

Global ratings have some well-known disadvantages. They are often only given at the end of a rotation when assessors may have forgotten details of student's performance. In addition, they may be biased due to a halo effect, i.e. the phenomenon that an impression created by a student's good or poor performance in one area affects assessors' judgements of that student's performance in another area.[24]
In the introductory clerkship, we could easily have used structured assessment with rating forms, such as in-training assessment. However, we purposely used global ratings, because in this study we set out to investigate the possibility of improving the reliability and validity of such ratings, as they continue to be much used in undergraduate and graduate training. With improved reliability and validity, global ratings could make a truly positive contribution to assessment of clinical performance, the more so since they can cover more competencies than assessment formats focused on specific items.[25]
We investigated the reliability and the predictive validity of GPRs in undergraduate training. We studied the reliability of GPRs in an introductory clerkship where the members of staff who rated the students supervised students' performance more frequently than is customary in undergraduate clerkships. We expected that this would yield a better, i.e. moderate to high, reliability than is generally reported for GPRs in undergraduate clinical training.
We observed an inter-rater reliability of 0.41, which is comparable with the literature on undergraduate inter-rater agreement. We speculate that the potentially positive influence of increased supervision of students' clinical work by staff may have been mitigated by the limited duration of the clerkship as compared with residency rotations.[26]

A relatively shorter period during which staff are in a position to supervise students may lead to a correspondingly diminished accuracy of the perceived levels of students' performance. Hence increased supervision did not result in improved reliability. Furthermore, reliability may have (slightly) suffered on account of staff not having participated in assessment training before this study was conducted.[27,28]
The results showed that 25 GPRs from different examiners would be needed to achieve adequate reliability. Other studies have yielded estimates of between 7 and 14 ratings to attain a reliability of 0.80.[5,29,30] However, the assessment formats on which the GPRs in those studies were based included aspects that might potentially improve reliability, such as a long duration of the student-staff work relationship (up to one year), a highly detailed description of the behaviour associated with the low and high scale points on the rating scale and raters who were better acquainted with students' performance (e.g. resident ratings). We suspect that more ratings will be needed to reach acceptable reliability in undergraduate settings, where the working relationship of staff and students lasts only a short time and raters thus witness less of the students' work and have to judge performance without guidance from concrete descriptions of the behaviours corresponding to the different scale points.

The validity measure derived from the predictive validity of the GPRs for the scores on the ITA programme in the subsequent clerkship was 0.32. Although slightly higher than the predictive validity reported for overall competence in studies in both undergraduate and residency training, it is still quite low.[8,15] The GPRs in this study were based on staff members' evaluations of students at the end of a three-week rotation. Despite frequent student-staff interaction in these three weeks, details of the interactions can be lost quite quickly.[31,32] In contrast, the evaluations in the ITA programme were recorded immediately after the activity or behaviour that was evaluated and according to a checklist. Despite the more than usually intensive interaction between staff and students in the initial clerkship, the fact that the GPRs were based on less detailed information about students' clinical performance than the ITA scores may offer an explanation for the low predictive validity of the GPRs. Disattenuated predictive validity was 0.59, however, which is much higher. Our findings implicate that GPRs, despite being based on less detailed information, can still make a positive contribution to the evaluation of students' performance. In a recently published study, Kreiter and Ferguson found comparable disattenuated predictive validity when they compared global ratings of clinical clerkship performance with former measures of physical examination performance provided by simulated patients (SP) using ratings and checklists, and with SP ratings of rapport and communication.[33] They conclude that measures of skills by global ratings are correlated with other clinical performance measures and discuss that more studies of this topic are needed to conclude that global ratings make a positive contribution to students' evaluation and thus contributes to the conclusion that global rating can positively contribute to students' evaluation. The evidence in our study points in the same direction.
This study has one major drawback. We compared our findings to findings in the literature and not to those of a control group. The circumstances in which research in the presented literature was performed were certainly different from the circumstances of our study. However, it was practically not feasible to have a control group in the same clerkship at the same time due to staff shortage and the practical impossibility to have two educational programmes performed by the same members of staff during the same period of time.

Our results indicate that even when conditions in an undergraduate rotation are positively manipulated, reliability and validity of GPRs remain low. However, the reliability and validity we reached were not lower than those found for other assessment formats performed over a short testing time.[34,35] This means that GPRs can contribute to the assessment of undergraduate students' clinical competencies as long as they are sampled on many occasions and by many assessors. Nevertheless, sufficient reliability and validity are likely to be hard to achieve. In a recent review, Williams *et al.* concluded that GPRs by themselves were an insufficient measure of students' clinical competence, even though they might be an important source of information about it.[36] These authors recommended that GPRs should be supplemented with ratings of students' performance in standardised clinical encounters and assessment protocols. The results of our study point to a similar recommendation, i.e. to combine GPRs with more specific and reliable assessment formats, such as the ITA programme in this study, to arrive at an integrated assessment programme. Further studies will have to examine whether such an assessment programme can provide reliable and valid measures of students' competencies.

REFERENCES

1. Magarian GJ, Mazur DJ. Evaluation of students in medicine clerkships. Acad Med 1990;65:341-5.
2. Kassebaum DG, Eaglen RH. Shortcomings in the evaluation of students' clinical skills and behavior in medical school. Acad Med 1999;74:842-9.
3. Dielman TE, Hull A, Davis WK. Psychometric properties of clinical performance ratings. Evaluation and the Health Professions 1980;3:103-17.
4. Maxim BR, Dielman TE. Dimensionality, internal consistency and interrater reliability of clinical performance ratings. Med Educ 1987;21:130-7.
5. Haber RJ, Avins AL. Do ratings on the American Board of Internal Medicine resident evaluation form detect differences in clinical competence. J Gen Int Med 1994;9(3):140-5.

6. Kwolek CJ, Donnelly MB, Sloan DA, Birrell SN, Strodel WE, Schwartz RW. Ward evaluations: should they be abandoned? J Surg Res 1997;69(1):1-6.
7. Davis JD. Comparison of faculty, peer, self and nurse assessment of obstetrics and gynecology residents. Obstet Gynecol 2002;99:647-51.
8. Durning SJ, Cation LJ, Jackson JL. The reliability and validity of the American Board of Internal Medicine monthly evaluation form. Acad Med 2003;78:1175-82.
9. Remmen R, Denekens J, Scherpbier A, et al. An evaluation study of the didactic quality of clerkships. Med Educ 2000;34:460-4.
10. Busari JO, Scherpbier AJJA, van der Vleuten CPM, Essed GGM. The perceptions of attending doctors of the role of residents as teachers of undergraduate clinical students. Med Educ 2003;37:241-7.
11. Kachalia A, Studdert DM. Professional liability issues in graduate medical education. JAMA 2004;292:1060-1.
12. Keynan A, Friedman M, Benbassat J. Reliability of global rating scales in the assessment of clinical competence of medical students. Med Educ 1987;21:477-81.
13. DaRosa DA, Dawson-Saunders B, Folse R. A comparison of objective and subjective measures of clinical competence. Evaluations and Program Planning 1985;8:327-30.
14. Adusumilli S, Cohan RH, Korobkin M, Fitzgerald JT, Oh MS. Correlation between radiology resident rotation performance and examination scores. Acad Radiol 2000;7:920-6.
15. Callahan CA, Erdmann JB, Hojat M, et al. Validity of faculty ratings of students' clinical competence in core clerkships in relation to scores on licensing examinations and supervisors' ratings in residency. Acad Med 2000;75:S71-3.
16. Norcini JJ, Webster GD, Grosso LJ, Blank LL, Benson JA Jr. Ratings of residents' clinical competence and performance on certification examination. J Med Educ 1987;62:457-62.
17. Pedhazur EJ. Multiple Regression in Behavioral Research – Explanation and Prediction. 2nd Edition. New York: Holt, Rinehart and Winston, 1982. p. 112-4.
18. Metheny WP. Limitations of physician ratings in the assessment of student clinical performance in an obstetrics and gynecology clerkship. Obstet Gynecol 1991;78:136-41.
19. Streiner DL. Global rating scales. In: Neufield VR, Norman GR (eds). Assessing Clinical Competence. New York: Springer Publishing Company, 1985. p.114-41.
20. Turnbull J, van Barneveld C. Assessment of clinical performance: in-training evaluation. In: Norman GR, van der Vleuten CPM, Newble DI (eds). International Handbook of Research in Medical Education. Dordrecht/Boston/London: Kluwer Academic Publishers, 2002. p. 793-810.
21. Turnbull J, MacFadyen J, van Barneveld C, Norman G. Clinical work sampling: a new approach to the problem of in training evaluation. J Gen Intern Med 2000;15:556-61.
22. Daelmans HEM, van der Hem-Stokroos HH, Hoogenboom R, Scherpbier AJJA, Stehouwer CDA, van der Vleuten CPM. Feasibility and reliability of an in-training assessment programme in an undergraduate clerkship. Med Educ 2004;12:1270-7.
23. Kreiter DC, Ferguson K, Lee W, Brennan RL, Densen P. A generalizability study of a new standardized rating form used to evaluate students' clinical clerkship performances. Acad Med 1998;73:1294-8.
24. McKinstry BH, Cameron HS, Elton RA, Riley SC. Leniency and halo effects in marking undergraduate short research projects. BMC Med Educ 2004;4:28.
25. Van Luijk SJ, van der Vleuten CPM. A comparison of checklists and rating scales in performance-based testing. In: Hart IR, Harden RM, DesMarchais J (eds). Current Developments in Assessing Clinical Competence. Montreal: Can Health Publications, 1992. p. 357-82.
26. Rothstein HR. Inter-rater reliability of job performance ratings: growth to asymptote level with increasing opportunity to observe. J Appl Psychol 1990;75(3):322-7.
27. Noel GL, Herbers JEJ, Caplow MP, Cooper GS, Pangaro LN, Harvey J. How well do internal medicine faculty members evaluate the clinical skills of residents? Ann Int Med 1992;80:1294-8.
28. Newble DI, Hoare J, Sheldrake PK. The selection and training of examiners for clinical examinations. Med Educ 1980;14:345-9.
29. Carline JD, Paauw DS, Thiede KW, Ramsey PG. Factors affecting the reliability of ratings of students' clinical skills in a medicine clerkship. J Gen Int Med 1992;7:506-10.
30. 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.
31. Heneman RL. The effects of time delay in rating and amount of information observed on performance rating accuracy. Academy of Management Journal 1983;26:677-86.
32. Kassin S, Tubb V, Hosch H, Memon A. On the "general acceptance" of eyewitness testimony research: a new survey of the experts. Am Psychol 2001;56:405-16.
33. Ferguson KJ, Kreiter CD. Using a longitudinal database to assess the validity of preceptors' ratings of clerkship performance. Adv Health Sci Educ Theory Pract 2004;9:39-46.
34. Swanson DB. A measurement framework for performance-based tests. In: Hart I, Harden RJ (eds). Further Developments in Assessing Clinical Competence. Montreal: Can-Health Publications, 1987. p. 13-45.
35. Petrusa ER. Clinical performance assessment. In: Norman GR, Van der Vleuten CPM, Newble DI (eds). International Handbook of Research in Medical Education. Dordrecht/Boston/London: Kluwer Academic Publishers, 2002. p. 673-709.
36. Williams RG, Klamen DA, McGaghie WC. Cognitive, social and environmental sources of bias in clinical performance ratings. Teach Learn Med 2003;15:270-92.