Medical Teacher

ISSN: 0142-159X (Print) 1466-187X (Online) Journal homepage: informahealthcare.com/journals/imte20

Quality evaluation reports: Can a faculty development program make a difference?

Nancy L. Dudek, Meridith B. Marks, Timothy J. Wood, Suzan Dojeiji, Glen Bandiera, Rose Hatala, Lara Cooke & Leslie Sadownik

To cite this article: Nancy L. Dudek, Meridith B. Marks, Timothy J. Wood, Suzan Dojeiji, Glen Bandiera, Rose Hatala, Lara Cooke & Leslie Sadownik (2012) Quality evaluation reports: Can a faculty development program make a difference?, Medical Teacher, 34:11, e725-e731, DOI: 10.3109/0142159X.2012.689444

To link to this article: https://doi.org/10.3109/0142159X.2012.689444

Published online: 12 Nov 2012.

2012; 34: e725–e731 MEDICAL TEACHER

WEB PAPER

Quality evaluation reports: Can a faculty development program make a difference?

NANCY L. DUDEK[1], MERIDITH B. MARKS[1], TIMOTHY J. WOOD[1], SUZAN DOJEIJI[1], GLEN BANDIERA[2], ROSE HATALA[3], LARA COOKE[4] & LESLIE SADOWNIK[3]

[1]University of Ottawa, Canada, [2]University of Toronto, Canada, [3]University of British Columbia, Canada, [4]University of Calgary, Canada

Abstract

Background: The quality of medical student and resident clinical evaluation reports submitted by rotation supervisors is a concern. The effectiveness of faculty development (FD) interventions in changing report quality is uncertain.

Aims: This study assessed whether faculty could be trained to complete higher quality reports.

Method: A 3-h interactive program designed to improve evaluation report quality, previously developed and tested locally, was offered at three different Canadian medical schools. To assess for a change in report quality, three reports completed by each supervisor prior to the workshop and all reports completed for 6 months following the workshop were evaluated by three blinded, independent raters using the Completed Clinical Evaluation Report Rating (CCERR): a validated scale that assesses report quality.

Results: A total of 22 supervisors from multiple specialties participated. The mean CCERR score for reports completed after the workshop was significantly higher (21.74 ± 4.91 versus 18.90 ± 5.00, $p = 0.02$).

Conclusions: This study demonstrates that this FD workshop had a positive impact upon the quality of the participants' evaluation reports suggesting that faculty have the potential to be trained with regards to trainee assessment. This adds to the literature which suggests that FD is an important component in improving assessment quality.

Introduction

Assessing the clinical competence of medical students and residents is an essential role of medical schools and residency training programs. The majority of this assessment is done by clinical supervisors using in-training evaluation (ITE). Supervisors document their evaluation on an in-training evaluation report (ITER; Turnbull & Van Barneveld 2002). ITERs, which are also referred to as, among other terms, clinical performance reports, performance assessment forms, clinical performance progress reports and end of clinical rotation reports, usually consist of a list of items on a checklist or rating scale and written comments.

Despite the importance of this type of assessment, there is evidence to suggest that the final evaluation (i.e. pass vs. fail) written on the ITER is not always consistent with the evaluator's actual judgment of the performance, especially for the poorly performing resident (Cohen et al. 1993; Speer et al. 1996; Hatala & Norman 1999). As well, the literature suggests that current ITER style tools do not effectively assess the various individual competencies required by residency accreditation bodies (Silber et al. 2004; Lurie et al. 2009).

Several authors, including the Advisory Committee on Educational Outcome Assessment (Swing et al. 2009), have proposed that assessor training is a key component of high quality assessment in residency programs, with some suggesting that rater training may be the "missing link" in improving the quality of trainee assessments (Holmboe et al. 2011).

Practice points

- ITER quality is a concern.
- FD strategies to improve ITER quality have had mixed success.
- This study describes a FD program designed to improve ITER quality by focusing on the quality of narrative comments.
- Using a previously validated tool, we demonstrated an improvement in ITER quality following workshop participation suggesting that faculty may indeed be trainable in regards to the quality of their ITER comments.

Other authors have reported that the supervisors themselves identify a need for faculty development (FD) programs to help them improve their ability to complete ITERs (Dudek et al. 2005).

However, there is remarkable controversy regarding the effectiveness of FD for improving rater-based evaluation. While, there is some evidence to suggest that faculty can be trained to improve the quality of their assessments (Holmboe et al. 2004; Littlefield et al. 2005), evidence also exists to suggest that such training is largely ineffective (Newble et al. 1980; Cook et al. 2008), leading several authors to suggest that faculty might be largely untrainable in this regard (Newble et al. 1980; Williams et al. 2003; Cook et al. 2008).

Correspondence: N. Dudek, Faculty of Medicine, University of Ottawa, The Rehabilitation Centre, 505 Smyth Road, Room 1105D, Ottawa, ON K1H 8M2, Canada. Tel: 613 737 7350 ext. 75596; fax: 613 737 9838; email: ndudek@ottawahospital.on.ca

ISSN 0142–159X print/ISSN 1466–187X online/12/110725–7 © 2012 Informa UK Ltd.
DOI: 10.3109/0142159X.2012.689444

With few exceptions (Littlefield et al. 2005), the research in this area has focused heavily on the psychometric properties of the tools. There has been substantially less attention paid to other properties such as the narrative information value and how that might be useful for learners and decision makers (Dudek et al. 2008a). In an effort to redress this gap regarding the utility of the narrative information on assessments, Dudek et al. (2008a) have looked at a broader set of criteria for higher quality completed ITERs and have developed a reliable scale to evaluate these properties. The objective of this study is to expand the debate about the value of FD for improving rater-based assessment using this broader set of criteria to assess the effectiveness of a FD workshop targeted at improving the quality of completed ITERs.

Method

This project was approved by all participating institutions' research ethics boards. A multi-site, uncontrolled, pre-post design was used to assess the effectiveness of the FD workshop.

Faculty development workshop

We developed a workshop entitled "Completing quality ITERs: What every supervisor needs to know" to meet clinical supervisors' perceived and observed need for help to improve the quality of their completed ITERs (Dudek et al. 2005, 2008b). Designed to facilitate active learning, it enables participants to: (1) describe the importance of well-completed ITERs in supporting trainee learning, (2) discuss the features of a well-completed ITER and (3) identify challenges and potential solutions to enhance the quality of the ITER completed within their current education system. The workshop was piloted locally and demonstrated a positive effect on ITER quality (Dudek et al. 2008b).

This 3-h workshop was then provided at three separate Canadian medical schools (Universities of British Columbia, Calgary and Toronto) during 2008 and 2009. The outline, detailing the content and the amount of time allocated to each area for the workshop appears in Table 1. The same set of slides was used for each workshop to insure that the basic content was addressed. To allow maximal applicability for the participants, examples discussed in the workshops varied with the situations presented by the participants. The principal investigator participated as a co-facilitator for all workshops along with the co-investigator from each local site. This model insured that the workshop content was consistent across sites yet acknowledged the different cultures of the three institutions.

Participants

Physicians who supervise and evaluate medical trainees on clinical rotations were invited to participate in the workshop. Recruitment occurred through various means: university FD strategies (emails to all teaching faculty, website announcement); direct emails and letters to program directors; and, presentation of the workshop opportunity at multiple program director meetings. Program directors were invited to participate and were asked to advertise the workshop to their faculty members. Only clinical supervisors who agreed to complete all segments of the program, including the pre-post evaluations, were permitted to enroll. In addition, only supervisors who used ITER forms satisfying the criteria of the evaluation tool (see below) could participate.

Outcome measures

First, we assessed the impact of the FD program on participants' application of newly acquired knowledge and skills by measuring the quality of the ITERs that they submitted pre- and post-workshop. All study participants were asked to submit the three most recent ITERs that they completed for medical trainees prior to the FD workshop. This number was chosen because we felt that it would be representative of the ITERs that they had recently completed. We also needed to be practical and acknowledge that supervisors in different programs may evaluate vastly different numbers of trainees. We felt that the majority of potential participants would have completed three ITERs in the 6 months prior to the workshop. Participants were also asked to submit all ITERs completed during the 6 months following the workshop. This timeline was chosen because not all supervisors would complete an

Table 1. Outline of the ITER FD workshop.

Time allocation	Activity
15 min	Introductions
20 min	Interactive discussion regarding the need for well-completed ITERs and the challenges and concerns in completing ITERs
15 min	• Review features of a well-completed ITER • Review of the CCERR
20 min	• Use the CCERR to evaluate four ITERs (ITERs blinded for author and trainee) • Follow-up discussion of issues noted with ITERs
15 min	Break
20 min	• Use the CCERR to evaluate one of the participants' own ITERs • Follow-up discussion on issues noted • Participants to identify two areas for improvement in their own ITERs
40 min	Tips to facilitate completion of ITERs (included how to identify specific behaviors, how to document behaviors, and an opportunity to practice writing effective comments); presented by facilitators with input from group
20 min	View TV clips of residents performing clinical tasks and practice writing appropriate ITER comments
15 min	Review challenges and summarize key points – insure all have been addressed

ITER in the first month or two after the workshop depending on when they were next asked to supervise a trainee. However, it was felt that 6 months would be long enough for most supervisors to have completed ITERs. The ITERs were collected monthly from all participants. ITERs were blinded for timing (pre vs. post), authorship, location and trainee. ITER quality was measured using the Completed Clinical Evaluation Report Rating (CCERR; Dudek et al. 2008a). The CCERR provides a reliable rating of the quality of ITERs completed by clinical supervisors (Dudek et al. 2008a). Nine items are rated on five point scales (where a rating of 3 is defined as acceptable) resulting in a total score that ranges from 9 to 45. The descriptions for each of the items can be found in Table 3. A full description of the tool along with its development and validation has been previously published (Dudek et al. 2008a). In brief, the CCERR was developed using a focus group to determine key features of high quality completed ITERs. These features were used to create the CCERR. It was pilot tested locally, analyzed, modified and then tested on a national level. The reliability of the CCERR was 0.82 in the national field test. Evidence for validity was demonstrated by the CCERR's ability to differentiate between groups of completed ITERs previously judged by experts to be of high, average and poor quality. The CCERR can be used on any style of ITER form provided that it has a list of items to be evaluated on a checklist or rating scale and a space for comments.

Second, as is typical for FD programs, we assessed participants' satisfaction with the workshop. This was determined through feedback from participants using a validated 9-item, seven-point continuing medical education (CME) satisfaction measure (Wood et al. 2005). This tool also provides a space for comments.

Raters

Physicians who supervise medical students and residents were recruited from an additional medical school to evaluate the ITERs using the CCERR. Previous work has shown that clinical supervisors can reliably use the CCERR without the need for any additional rater training beyond reading the brief instructions provided with the CCERR (Dudek et al. 2008a). Our pilot study data suggested that for adequate reliability, we would need a minimum of two physician raters per ITER (Dudek et al. 2008b). To be sure, we recruited three raters for this project. All raters evaluated all ITERs. One of the physician raters had used the CCERR before (rater 1) and the other two had not.

Analysis

For each participant, a total CCERR score for each ITER was calculated by summing the ratings across items. Next, in order to minimize the variability in the number of ITERs that a participant may have submitted (particularly post-workshop), the pre-workshop scores for individual items and for the total CCERR scores were averaged across the total number of ITERs submitted. Similarly, a mean total score for each post-workshop ITER was determined. To determine a measure of inter-rater reliability, an intra-class correlation was calculated on the pre-workshop total CCERR score and on the post-workshop total CCERR score. Of interest would be whether the reliabilities differ from pre- to post-workshop despite the raters being blinded to the time of the rating. To determine if the workshop differentially influenced specific items, the item ratings and mean total scores were averaged over the three raters, and a repeated measures ANOVA with time (pre- vs. post-workshop) and items (1–9) was conducted.

A mean score for each item on the program satisfaction scale was calculated. Conventional content analysis was used to code the written comments from the program satisfaction scale. These comments were iteratively and independently read by two of the investigators (Nancy Dudek and Rose Hatala). Coding categories were developed to note common themes and consensus was reached regarding these categories. This coding structure was used to analyze all of the written comments on the program satisfaction scale.

Results

A total of 22 clinical supervisors (site one – 12, site two – 7, and site three – 3) participated. Several specialties were represented in this group: Family Medicine, Internal Medicine (including various subspecialties), Plastic Surgery, Obstetrics and Gynecology, Pediatrics, Physical Medicine and Rehabilitation, Neurology, and Radiation Oncology.

Inter-rater reliability

The inter-rater reliability across all raters was 0.88 (pre-workshop CCERR scores) and 0.92 (post-workshop CCERR scores) indicating that the raters were consistent despite being blinded as to whether the ITERs were pre- or post-workshop.

ITER quality

Table 2 illustrates the number of ITERs submitted by each participant and their mean total CCERR score for both pre- and post-workshop. A total of 64 pre-workshop ITERs and 171 post-workshop ITERs were submitted for an average of 2.91 pre-workshop ITERs and 7.77 post-workshop ITERs per participant.

In addition, the mean total CCERR score of all participants' ITERs pre-workshop was 18.90 ± 5.00 (Table 2). The mean CCERR score of all participants' ITERs post-workshop was 21.74 ± 4.91. There was a significant difference between pre- and post-scores ($F(1,21) = 6.54$, $p = 0.02$). The effect size was $\eta_p^2 = 0.24$ which would be considered moderate to large in size.

Table 3 illustrates the mean pre- and post-workshop scores for each item on the CCERR. As shown in this table, there were some differences in the mean score of individual items ($F(8,168) = 112.22$, $p < 0.001$, $\eta_p^2 = 0.84$), but the interaction between items and time did not differ significantly ($F(8,168) = 1.57$, $p > 0.05$, $\eta_p^2 = 0.07$) indicating that the pre/post differences were consistent for all items. This latter finding suggests that our workshop contributed to overall ITER quality improvement by making small improvements in all items rather than making larger improvements for only certain items.

Table 2. Participant mean total CCERR scores by time (pre- and post-workshop).

Participant	Number of pre-workshop ITERs	Mean CCERR score (±standard deviation) pre-workshop	Number of post-workshop ITERs	Mean CCERR score (±standard deviation) post-workshop
1	3	24.89 ± 8.86	10	17.80 ± 5.65
2	3	28.00 ± 8.82	4	23.67 ± 8.13
3	3	15.56 ± 1.92	3	17.67 ± 3.38
4	3	17.56 ± 4.82	14	17.12 ± 4.68
5	3	14.33 ± 3.06	3	20.33 ± 6.93
6	3	11.22 ± 1.35	14	13.81 ± 2.25
7	3	18.44 ± 3.85	10	19.00 ± 5.72
8	3	21.22 ± 7.15	11	17.61 ± 4.39
9	3	19.00 ± 5.24	6	22.44 ± 6.18
10	3	20.44 ± 6.19	5	29.27 ± 9.47
11	3	18.89 ± 3.67	6	28.22 ± 10.44
12	3	16.44 ± 3.60	2	24.67 ± 9.67
13	3	12.56 ± 1.07	6	19.33 ± 5.51
14	1	23.67 ± 9.07	4	24.67 ± 7.34
15	3	15.11 ± 3.75	3	22.67 ± 7.86
16	3	12.78 ± 2.17	8	18.79 ± 6.35
17	3	17.89 ± 4.25	5	27.13 ± 6.93
18	3	23.78 ± 7.60	1	32.33 ± 10.79
19	3	28.44 ± 9.05	6	20.56 ± 6.88
20	3	24.89 ± 9.75	2	27.50 ± 6.50
21	3	17.33 ± 5.36	8	18.17 ± 5.28
22	3	13.33 ± 2.03	40	15.52 ± 3.72
Total	64	18.90 ± 5.00	171	21.74 ± 4.91

Note: The total CCERR score has a range of 9–45.

Table 3. Individual CCERR item scores by time for entire group ($n = 22$).

	Item	Pre-workshop mean item CCERR score (±standard deviation)	Post-workshop mean item CCERR score (±standard deviation)
1	Checklist/numeric ratings show sufficient variability to allow identification of relative strengths and weaknesses of the trainee	3.04 ± 0.44	3.06 ± 0.50
2	Comments are balanced providing both strengths and areas for improvement	1.75 ± 0.84	2.15 ± 0.84
3	The trainee's response to feedback and/or remediation during the rotation is described in the comments	1.44 ± 0.46	1.58 ± 0.50
4	Comments justify the ratings provided	2.45 ± 0.60	2.77 ± 0.49
5	Clearly explained examples of strengths using specific descriptions (not generalizations) are provided in the comments	2.23 ± 0.72	2.71 ± 0.72
6	Clearly explained examples of weaknesses using specific descriptions (not generalizations) are provided in the comments	1.59 ± 0.73	1.91 ± 0.78
7	Concrete recommendations for the trainee to attain a higher level of performance are provided	1.58 ± 0.70	1.94 ± 0.76
8	Comments are provided in a supportive manner	2.82 ± 0.62	3.19 ± 0.42
9	Overall, this ITER provides enough detail for an independent reviewer to clearly understand the trainee's performance on the rotation	2.01 ± 0.68	2.43 ± 0.70

Note: The item CCERR scores have a range of 1–5.

Program satisfaction

The individual items on the CME scale ranged from 5.88 to 6.27 (where 7 indicates outstanding, 6 excellent and 5 very good Wood et al. 2005). The overall quality of the workshop was rated at 6.04. Written comments from participants indicating the useful aspects of the workshop fell into two categories: (1) structure of the workshop and (2) content. Cited strengths of the workshop included the degree of interactivity and the varied learning methods. The practical suggestions and the provision of examples with which to practice were noted as the most useful content.

Discussion

Based on the traditional measure of participant satisfaction, we developed a successful workshop. However, more importantly our FD workshop had a positive impact on the demonstrated ability of clinical supervisors to complete quality ITERs as measured using the CCERR. This suggests that faculty can be trained with a relatively modest intervention (our workshop was only 3 h in length) and adds to the literature that has found success with rater training for improving the quality of trainee assessments (Holmboe et al. 2004; Littlefield et al. 2005).

The importance of a statistically significant pre/post difference of 2.84 points on the CCERR may be questioned, given that the total on the CCERR can range from 9 to 45 points. The effect size for this measure was moderate to large and we believe that the difference is also educationally significant. Previous work demonstrated that ITERs evaluated by experts and identified as "poorly completed" (mean CCERR score ~16) versus those they rated as "average" (mean CCEER score ~24) differed by 8 points on the CCERR. A similar difference of 8 points was found between the expert rated "average" and "excellently" completed ITERs (Dudek et al. 2008a). Expecting participation in one workshop to change clinical supervisors' practices enough to have their ITERs go from poor to average or average to excellent would be unrealistic and inconsistent with educational theory which tells us that in order to develop expertise significant deliberate practice is required (Ericsson 2004; Cook et al. 2008). However, making over 30% of the improvement towards the next level of performance after one 3-h workshop can be argued to be educationally significant.

It is important to note that we made no effort in this study to examine the reliability of the checklist ratings assigned by supervisors on the ITERs. This was a deliberate choice based on previous research which found that a significant part of the problem in failing to report poor clinical performance is that supervisors often do not know what to document when completing an ITER (Dudek et al. 2005). It was noted that the problem was not limited to the actual form but included issues related to the use of the form. For example, most supervisors were aware that simply stating that the resident is incompetent will not provide adequate evidence to support a failing grade. However, they did not know how to complete a report that reflected their reasoning for why that resident performed at an incompetent level (Dudek et al. 2005). The specific components of a high quality completed ITER (i.e. not the form itself but rather the *completed* form) have been described in a previous study (Dudek et al. 2008a). Nine features were determined to be important. Eight of these nine features dealt with the comments (Dudek et al. 2008a). Therefore, the workshop focused on improving these features. This approach is in line with recent discussions in the literature which suggest that we move beyond focusing solely on numeric rating scales and incorporate more qualitative types of assessment (Holmboe et al. 2011). The use of this approach might have contributed to the success of our study. Perhaps faculty are more trainable when it comes to improving the quality of their comments on evaluation reports.

Our study also demonstrates that it is possible to assess a FD program's ability to create behavior change. Traditionally, FD programs have been assessed using satisfaction surveys, self-reported ratings of confidence, or at best, pre/post knowledge tests. While some groups have assessed lecturing skills in a controlled classroom setting (D'Eon 2004; Pandachuck et al. 2004) or the quality of clinical supervisors' feedback skills (Marks et al. 2008), *objective* assessment of teaching and evaluation skills in clinical environments is rarely considered (Steinert et al. 2006; Marks et al. 2008). The dearth of FD program evaluation at this level in part relates to a lack of reliable, valid and objective outcome measures to assess for change in specific teaching behaviors (Steinert et al. 2006). The use of the CCERR, a previously developed objective tool, allowed us to go beyond the usual outcome measures of participant satisfaction and knowledge gains (Steinert et al. 2006), and demonstrate a change in behavior in actual practice. This level of assessment corresponds to Kirkpatrick's (1998) third level of program evaluation: behavior change or application of knowledge and skills.

Our study has limitations. Despite aggressive recruitment strategies, we were only able to enroll 22 individuals in our study. This prevented us from doing more detailed analyses, such as determining whether workshop site had an impact on ITER quality improvement, as it would be inappropriate to compare sites when the numbers were so small. We chose the workshop format to address the participants' ITER learning needs as interactive workshops are a favored method of providing FD programs. As well, although not often assessed at this level, some workshops have been shown to result in professional practice change (Wilkerson & Irby 1998; O'Brien et al. 2001; D'Eon 2004; Dudek et al. 2008b). Workshops provide an interactive environment in which participants can share their ideas and concerns. This is a particularly important concept when we wish to have an impact on attitudes. Given the extended period of time that many supervisors have been completing ITERs, convincing some supervisors of the importance and need for quality ITERs was thought to be essential to promote behavior change in this domain.

However, the difficulties in recruiting physicians to participate in FD initiatives are well documented (Rubeck & Witzke 1998). As well, the addition of an evaluation component seemed to deter some interested faculty from participating despite the anonymous nature of the analysis. Many potential participants also cited a concern with the time it might take them to collect their ITERs. We took many steps to insure that the time it would take faculty to participate in the study would be minimal. Those who participated in the study indicated that the time commitment was very small.

Ideally, each site would also have a control group where the participants would submit their ITERs to be evaluated on the same time-line as the intervention group. A control group was attempted. Participants from that site were asked to sign-up to participate in the workshop either at the study onset, or 6 months later. The workshop participants in the later group were to serve as the control group, as they were not to receive the FD workshop until the completion of the study but were to submit ITERs on the same time-line as the intervention group. Only two individuals signed up for the control group which prevented us from having a comparison group. It is possible that simply being aware that their ITERs were going to be evaluated made the participating faculty complete these forms to a higher standard. However, our study does show that, regardless of the effects of monitoring, participants are *able to* improve ITER quality upon completion of the FD workshop. As well, we feel the change in behavior demonstrated in this study is legitimate (beyond simply the observation bias) because we chose to use ITERs completed in the course of normal trainee supervision. Supervisors improved the manner

in which they completed real ITERs that affect trainee records, rather than simulated ITERs or 'parallel' ITERs that did not count towards the trainee records.

The combined results of participant satisfaction and an objective evaluation of improved performance suggest that supervisors benefited from participating in the workshop. Therefore, given that a relatively small proportion of medical faculty attend FD workshops, we are exploring offering the same workshop content using alternative learning methods (take home FD program, on-line workshop, etc.) that do not rely on faculty coming to a particular location for a set period of time to participate in a workshop. In this program, we also plan to explore the impact of repeated/refresher interventions to determine if larger improvements in ITER quality can be made overtime. The effectiveness of those methods can be compared to the results from this project in future studies.

Our study demonstrates that participating in a FD workshop enabled supervisors to complete evaluation reports to a higher standard suggesting that faculty may indeed be trainable. This improvement may have occurred in part due to our focus on the non-psychometric aspects of clinical performance assessment. This approach suggests an additional aspect of rater training to consider when designing FD programs aimed at improving the assessment of our medical trainees' clinical competence.

Acknowledgements

The authors would like to thank Ms Jeanie Zeiter for her assistance with the project management and preparation of the manuscript.

Declaration of interest: Funding support for this study was obtained from an Academy for Innovation in Medical Education Education Research Grant. The authors report no conflicts of interest.

Notes on contributors

NANCY DUDEK, MD, MEd, is an associate professor, Faculty of Medicine, University of Ottawa and the director for the University of Ottawa's Physical Medicine and Rehabilitation Residency Program, Ottawa, Canada.

MERIDITH MARKS, MD, MEd, is a professor, Faculty of Medicine, and the assistant dean of the Academy for Innovation in Medical Education (AIME), University of Ottawa, Ottawa, Canada.

TIMOTHY WOOD, PhD, is an assistant professor, Department of Medicine, University of Ottawa, and a PhD educator with the Academy for Innovation in Medical Education (AIME), Ottawa, Canada.

SUZAN DOJEIJI, MD, MEd, is an associate professor, Faculty of Medicine, University of Ottawa, and the physiatrist-in-chief The Ottawa Hospital Rehabilitation Centre (TOHRC), chief of Department of Physical Medicine and Rehabilitation Bruyère Continuing Care, and Chair of Division of Physical Medicine and Rehabilitation in the Faculty of Medicine, Ottawa, Canada.

GLEN BANDIERA, MD, MEd, is an associate professor of Medicine, associate dean, Postgraduate Medical Education, University of Toronto, and Chief of Emergency Medicine, St. Michael's Hospital, Toronto, Canada.

ROSE HATALA, MD, MEd, is a clinical associate professor of Medicine and an associate program director for the Internal Medicine Residency Program, Faculty of Medicine, University of British Columbia, Vancouver, Canada.

LARA COOKE, MD, MSc, is an assistant professor of Clinical Neurosciences and the program director, Neurology Residency Training Program, University of Calgary, Calgary, Canada.

LESLIE SADOWNIK, MD, MEd, is an assistant professor in the Department of Obstetrics and Gynaecology and the project lead for the Office for Faculty Development and Educational Support, Faculty of Medicine University of British Columbia, Vancouver, Canada.

References

Cohen GS, Blumberg P, Ryan NC, Sullivan PL. 1993. Do final grades reflect written qualitative evaluations of student performance? Teach Learn Med 5(1):10–15.

Cook DA, Dupras DM, Beckman TJ, Thomas KG, Pankratz S. 2008. Effect of rater training on reliability and accuracy of mini-CEX scores: A randomized controlled trial. J Gen Intern Med 24(1):74–79.

D'Eon MF. 2004. Evaluation of a teaching workshop for residents at the University of Saskatchewan: A pilot study. Acad Med 79(8):791–797.

Dudek N, Marks M, Lee C, Wood T. 2008a. Assessing the quality of supervisors' completed clinical evaluation reports. Med Educ 42(8):816–822.

Dudek NL, Marks MB, Regehr G. 2005. Failure to fail: The perspectives of clinical supervisors. Acad Med 80(10):S84–S87.

Dudek N, Marks M, Wood T, Dojeiji S. 2008b. Clinical evaluation report quality: Faculty development can make a difference. Abstract presented at the Research in Medical Education (RIME) Conference Oct 31–Nov 5, San Antonio TX.

Ericsson KA. 2004. Deliberate practice and the acquisition and maintenance of expert performance in medicine and related domains. Acad Med 79(10):S70–S81.

Hatala R, Norman GR. 1999. In-training evaluation during an internal medicine clerkship. Acad Med 74(10):S118–S120.

Holmboe ES, Hawkins RE, Huot SJ. 2004. Effects of training on direct observation of medical residents' clinical competence: A randomized trial. Ann Intern Med 140(11):874–881.

Holmboe ES, Ward DS, Reznick RK, Katsufrakis PJ, Leslie KM, Patel VL, Ray DD, Nelson EA. 2011. Faculty development in assessment: The missing link in competency-based medical education. Acad Med 86(4):460–467.

Kirkpatrick DL. 1998. Evaluating training programs: The four levels. 2nd ed. San Francisco: Berrett-Koehler.

Littlefield JH, DaRosa DA, Paukert J, Williams RG, Klamen DL, Schoolfield JD. 2005. Improving resident performance assessment data: Numeric precision and narrative specificity. Acad Med 80(5):489–495.

Lurie SJ, Mooney CJ, Lyness JM. 2009. Measurement of the general competencies of the accreditation council for graduate medical education: A systematic review. Acad Med 84(3):301–309.

Marks MB, Wood TJ, Nuth J, Touchie C, O'Brien H, Dugan A. 2008. Assessing change in clinical teaching skills: Are we up for the challenge? Teach Learn Med 20(4):288–294.

Newble DI, Hoare J, Sheldrake PF. 1980. The selection and training of examiners for clinical examinations. Med Educ 14(5):345–349.

O'Brien MA, Freemantle N, Oxman AD, Wolf F, Davis DA, Herrin J. 2001. Continuing education meetings and workshops: Effects on professional practice and healthcare outcomes. Cochrane Database of Syst Rev 1, Art. no. CD003030.

Pandachuck K, Harley D, Cook D. 2004. Effectiveness of a brief workshop designed to improve teaching performance at the University of Alberta. Acad Med 79(8):798–804.

Rubeck RF, Witzke DB. 1998. Faculty development: A field of dreams. Acad Med 73(9):S32–S37.

Silber CG, Nasca TJ, Paskin DL, Eiger G, Robeson M, Veloski JJ. 2004. Do global rating forms enable program directors to assess the ACGME competencies? Acad Med 79(6):549–556.

Speer AJ, Solomon DJ, Ainsworth MA. 1996. An innovative evaluation method in an internal medicine clerkship. Acad Med 71:S76–S78.

Steinert Y, Mann K, Centeno A, Dolmans D, Spencer J, Gelula M, Prideaux D. 2006. A systematic review of faculty development initiatives

designed to improve teaching effectiveness in medical education: BEME guide no. 8. Med Teach. 28(6):497–526.

Swing SR, Clyman SG, Holmboe ES, Williams RG. 2009. Advancing resident assessment in graduate medical education. J Grad Med Educ 1(2):278–286.

Turnbull J, Van Barneveld C. 2002. Assessment of clinical performance: In-training evaluation. In: Norman GR, Van der Vleuten CPM, Newble DI, editors. International handbook of research in medical education. London: Kluwer: Academic Publishers. pp 793–810.

Wilkerson L, Irby DM. 1998. Strategies for improving teaching practices: A comprehensive approach to faculty development. Acad Med 73(4):387–396.

Williams RG, Klamen DA, McGaghie WC. 2003. Cognitive, social and environmental sources of bias in clinical performance ratings. Teach Learn Med 15(4):270–292.

Wood T, Marks M, Jabbour M. 2005. Development of a participant questionnaire to assess continuing medical education presentations. Med Educ 39(6):568–572.