Faculty Development- Is Some Better Than None? [version 1]

Kelsey Anne Crawford, Timothy J Wood, Karl-André Lalonde, Nancy Dudek
University of Ottawa

Abstract
This article was migrated. The article was marked as recommended.

Introduction: With the advent of competency-based medical education there is an emphasis on formative workplace based assessment. The quality of these assessments is a concern for medical educators and their trainees. Faculty development (FD) strategies to improve assessment quality have resulted in some success. However, few faculty participate, and those who do are likely more motivated to improve, making it difficult to demonstrate a conclusive benefit. To address these weaknesses, we designed a FD initiative to improve the quality of completed in-training evaluation reports (ITERs). All faculty within a division participated. We hypothesized that clinical supervisors would improve their ITER quality based on feedback, regardless of their own motivation to do so, with a simple, point-in-time intervention.

Methods: In this three-phase study, two independent raters used the Completed Clinical Evaluation Report Rating (CCERR) to assess the quality of ITERs completed by all faculty in the Division of Orthopedic Surgery at the University of Ottawa. In phase one, ITERs from the previous nine months were evaluated. In phase two, the participants were aware that their ITERs were being evaluated, but they did not receive feedback. In phase three, participants received regular feedback on their performance in the form of their mean CCERR scores. Mean CCERR scores from the different phases of the study were compared.

Results: CCERR scores were similar for all three phases (one: 17.56 ± 1.02, two: 17.65 ± 0.96, three: 17.54 ± 0.75, p=0.98).

Discussion and Conclusions: There was no evidence in our study that participants’ improved their ITER quality despite being aware that they were being evaluated and/or receiving feedback. Potentially, this was related to a lack of motivation. Alternatively, the intensity and/or
frequency of the feedback may have been inadequate to create change. These results raise concerns that some faculty development may not necessarily be better than none.

**Keywords**
Assessment, Faculty Development, In-Training Assessment, Workplace Based Assessment, Feedback
**Introduction**

Medical education training programs, both at the undergraduate and postgraduate level, need to assess the clinical performance of their trainees to ensure that they are competent to move to the next level of training or into independent practice. In-training evaluation (ITE) by physician preceptors is a common component of many training programs’ assessment process. This assessment is recorded on an In-Training Evaluation Report (ITER). ITERs are also referred to as clinical performance reports, performance assessment forms, clinical performance progress reports and end of clinical rotation reports. ITERs follow the typical format of many workplace based assessment (WBA) tools in that they consist of a list of items on a checklist or rating scale and written comments. Unfortunately, ITERs are often poorly completed, particularly in the case of the poorly performing resident (Cohen *et al.*, 1993; Speer, Soloman and Ainsworth, 1996; Hatala and Norman, 1999). There is also evidence that clinical supervisors lack knowledge regarding what to document on ITERs and that this is in part responsible for their failure to report unsatisfactory clinical performance (Dudek, Marks and Regehr, 2005).

With the advent of competency-based medical education (CBME) there is a substantial emphasis on WBA. Although ITERs are less likely to be used in a CBME program of assessment given their summative focus, similar concerns exist surrounding other formative WBA tools with a comparable format (checklist or rating scale with space for written comments). For example, the Daily Encounter Card is commonly used in many emergency medicine programs. Unfortunately, it has similar concerns with regards to quality (Cheung *et al.*, 2016).

Faculty development (FD) is commonly used to address concerns with regards to the teaching and assessment skills of physicians who supervise medical trainees. FD is anticipated to be essential for a successful transition to CBME (Dath and Iobst, 2010; Royal College of Physicians and Surgeons of Canada 2017). The Advisory Committee on Educational Outcome Assessment (Swing *et al.*, 2009) has proposed that assessor training is a key component in addressing the problem of quality assessments in residency programs. Clinical supervisors have also indicated that they want FD programs to help them improve their ability to complete evaluation reports (Dudek *et al.*, 2005). It seems logical that rater training would improve report quality. However, the literature on this is mixed with success (Holmboe, Hawkins and Huot, 2014; Littlefield *et al.*, 2005; Dudek *et al.*, 2012; Dudek, Marks and Dojeiji, 2013) and failures (Newble, Hoare and Ainsworth, 1996; Speer, Soloman and Ainsworth, 1996; Sheldrake, 1980; Cook *et al.*, 2009).

In addition to the controversy regarding the “trainability” of faculty, there are other concerns with a rater training approach. First, there is some concern that positive improvements noted with regards to rater training might be the result of the participants knowing that they are being observed, a type of Hawthorne effect (Holden, 2001). This concern is raised because many studies of FD strategies lack a suitable control group (Dudek *et al.*, 2012; Steinert *et al.*, 2006), which raises the question of whether it was the FD intervention or the monitoring that resulted in the observed improvement. Second, there is the constant concern of FD recruitment (Rubeck and Witzke, 1998). Notoriously, few faculty participate in FD so the impact of FD is not large. Third, there is the motivation issue. Potentially, those who chose to participate in FD may be more motivated to change; therefore, even if these strategies are proven successful they may not be universally applicable (i.e. to unmotivated faculty).

To address the dual concerns of recruitment and possible Hawthorne effect, Dudek et al. developed a FD program designed to improve completed ITER quality. This program provided similar content to an existing FD workshop that demonstrated improved ITER quality (Dudek *et al.*, 2012; Dudek, Marks and Dojeiji, 2013), but used an “at-home” format incorporating a feedback component (Dudek *et al.*, 2013a) and a control group. An improvement in ITER quality was noted in the intervention group, but not in the control group, suggesting that the information was creating a change in ITER quality. As well, this “at home” program had greater participation than the “in-person” workshop. However, recruitment remained an issue and limited the power of the study. In addition, the Dudek *et al.* study did not address the motivation issue, as participation was voluntary. Therefore, it can be assumed that those who agreed to participate were motivated to improve their ITER quality.

To build on this previous work, the current study was designed to address these two limitations: low participation rate and possible motivational differences. First, we made participation in the study extremely simple as participants were no longer responsible for sending research team their ITERs for assessment. Their ITERs were simply provided to the research team by the program’s administrative assistant. Second, we had all supervisors in one postgraduate residency participate so that we would be offering this FD initiative to everyone and not just those who chose to participate.

The objectives of this study were to determine if clinical supervisors improved their ITER quality when they knew that they were being monitored and whether they could improve their ITER quality based on feedback, regardless of their own motivation to do so. If successful, this practical intervention to improve ITER quality could be applied in any residency program.
Methods
This study involved assessing the quality of the ITERs completed in the Orthopedic Surgery residency-training program at the University of Ottawa. Two raters using the Completed Clinical Evaluation Report Rating (CCERR) assessed the quality of the completed ITERs. The project was divided into 3 phases: 1) participants were unaware that their ITER quality would be assessed; 2) participants were aware that their ITER quality would be assessed but they did not receive any feedback and 3) participants received feedback in the form of their mean CCERR scores, a copy of the CCERR tool and a copy of the ITERs that they had completed as per the protocol used by Dudek, Marks, Bandiera et al. (2013). In this previous study there were two feedback groups: one group that received the items listed above and a second that included those items plus a feedback guide that provided additional information about how to improve their ITERs. No difference was noted between the groups (Dudek et al., 2013a). Therefore, given that we wanted the intervention in our study to be as simple as possible for faculty to participate in we chose to exclude the feedback guide for the present study. The mean CCERR scores from each phase were compared.

Participants
All Orthopedic Surgeons affiliated with The Ottawa Hospital Orthopedic Surgery residency program who completed ITERs between December 2012 and August 2015 were participants in this study. They provided consent at the start of phase two (see description below). This consent included looking at their past ITERs (phase 1 - see below). There was no additional work required outside of their typical educational responsibilities with the exception of reviewing the feedback sent via e-mail in phase three.

Experimental Design
Phase One (December 2012 - August 2013)
Participants were informed of the study and told that their ITERs were being collected and evaluated for the next two years starting in September 2013. They were also informed that phase one involved rating the preceding 9 months of their completed ITERs. Nine months was chosen, as participants would typically have completed 2 to 3 ITERs during that time period. This served as a baseline assessment of participants’ ITER quality when they did not think that their ITER quality was being assessed (a control phase).

| Participant | Number of Phase 1 ITERs | Mean CCERR score (± standard error) Phase 1 | Number of Phase 2 ITERs | Mean CCERR score (± standard error) Phase 2 | Number of Phase 3 ITERs | Mean CCERR score (± standard error) Phase 3 |
|-------------|--------------------------|---------------------------------------------|--------------------------|---------------------------------------------|--------------------------|---------------------------------------------|
| 1           | 5                        | 23.80 (± 2.10)                             | 3                        | 26.17 (± 1.34)                             | 5                        | 24.20 (± 0.42)                             |
| 2           | 3                        | 18.50 (± 2.24)                             | 8                        | 13.13 (± 2.22)                             | 4                        | 17.13 (± 1.64)                             |
| 3           | 3                        | 17.67 (± 1.80)                             | 4                        | 17.75 (± 1.64)                             | 2                        | 16.75 (± 1.24)                             |
| 4           | 3                        | 16.67 (± 2.00)                             | 7                        | 16.14 (± 1.14)                             | 3                        | 20.83 (± 1.57)                             |
| 5           | 3                        | 22.17 (± 2.67)                             | 6                        | 22.50 (± 1.62)                             | 11                       | 18.81 (± 1.24)                             |
| 6           | 1                        | 17.50 (± 2.00)                             | 2                        | 17.25 (± 1.24)                             | 1                        | 17.00 (± 1.64)                             |
| 7           | 4                        | 9.63 (± 1.57)                              | 7                        | 14.71 (± 1.64)                             | 5                        | 15.70 (± 1.14)                             |
| 8           | 4                        | 20.13 (± 2.00)                             | 6                        | 18.75 (± 2.22)                             | 2                        | 19.00 (± 1.24)                             |
| 9           | 3                        | 12.83 (± 1.80)                             | 5                        | 14.10 (± 1.64)                             | 1                        | 17.00 (± 1.64)                             |
| 10          | 6                        | 15.08 (± 2.00)                             | 7                        | 17.64 (± 1.64)                             | 4                        | 13.88 (± 1.64)                             |
| 11          | 3                        | 17.67 (± 2.10)                             | 3                        | 18.00 (± 1.42)                             | 3                        | 14.83 (± 1.24)                             |
| 12          | 4                        | 17.75 (± 1.80)                             | 7                        | 17.07 (± 1.24)                             | 2                        | 17.00 (± 1.64)                             |
| 13          | 3                        | 18.83 (± 2.00)                             | 9                        | 16.28 (± 1.24)                             | 4                        | 15.86 (± 1.64)                             |
| Total (of eligible participants, n=13) | 47 | 17.56 ± 1.02 | 45 | 17.65 ± 0.96 | 79 | 17.54 ± 0.75 |

Note that the total CCERR score has a range of 9 to 45.
**Phase Two (September 2013 - November 2014)**
During this phase, participants were aware that their ITERs were being studied, but they did not receive their CCERR scores. This phase addressed whether simply being aware that they were in a study affected participant performance. To help minimize bias, ITERs from phases one and two were rated at the same time and raters were blinded to phase.

**Phase Three (December 2014 - August 2015)**
Participants received feedback on their ITER quality in the form of the mean CCERR score for each item on the CCERR and the total CCERR score for each ITER that they completed. This was provided to them on an actual CCERR tool so that they were informed of what each item was assessing. They also received a copy of the corresponding ITER, for reference. This protocol, which was developed for the previous study (Dudek et al., 2013a), was designed based on commentary provided by past workshop participants that the CCERR is so self-explanatory that after just reviewing it along with their own ITERs it was easy to know how to improve their ITER quality (Dudek et al., 2013). Phase three data collection started after they had received their first feedback in December 2014. The first set of feedback included CCERR scores from ITERs completed from July to November 2014. Note that the CCERR results from July to November 2014 were included in phase two (as they had not yet received this feedback). This phase addressed whether participants improve their performance based on feedback, regardless of their motivation to do so.

Given that the CCERR score was to be relayed back to the participant in time for them to complete further ITERs, the raters were not blinded to study phase during this time. This phase was intended to evaluate the impact of feedback on their performance.

**ITERs**
ITERs were printed in paper form and de-identified (both preceptor and resident) by an administrative assistant before they were given to the raters for review.

**CCERR**
Results from the Completed Clinical Evaluation Report Rating (CCERR) have demonstrated evidence for validity (Dudek et al., 2008; Dudek et al., 2012; Dudek et al., 2013a). It is used to rate the quality of ITERs completed by physician supervisors regardless of the style of the form provided, it includes a list of items to be evaluated on a checklist or rating scale and a space for comments (Dudek et al., 2008). Nine items are rated on a five-point scale (where a three is defined as acceptable) resulting in a total score that ranges from nine to forty-five. For reference, a total CCERR score of ~16 corresponded with ITERs rated as poor quality by experts, a score of ~24 corresponded to average quality, and a score of ~32 was considered high quality (Dudek et al., 2008).

**Raters**
Two physician raters scored each ITER using the CCERR. Previous research demonstrated that two physician raters per ITER are sufficient for adequate reliability (Dudek et al., 2008). Physician raters can reliably use the CCERR without the need for additional rater training beyond following the written instructions on the CCERR (Dudek et al., 2008). Physician raters were blinded to the staff and resident associated with each ITER.

Phase three of our study required that the ITERs be evaluated in a timely fashion so that participants could receive feedback continually. Given that one of our objectives was to create an intervention that could be replicated in any residency program, the physician rater in this scenario would be aware that this intervention was occurring. This raises a concern for bias, as raters would know that the supervisors were getting feedback and may be expecting improvement. To fully re-create this scenario, and then evaluate for bias, we had the two physician raters be two of our co-investigators (who were aware of the intervention phase). At the conclusion of the study, all ITERs were rated by two additional independent raters who were blinded to participant, resident and phase to determine if there were differences between the blinded-to-phase and unblinded-to-phase ratings.

**Analysis**
Two physicians rated each ITER using the CCERR. A total CCERR score was determined for each submitted ITER by summing the ratings on each item of the CCERR assigned by a specific rater. For each participant in a phase, a mean total CCERR score was determined by averaging across the forms a participant submitted. These total CCERR scores for each participant were analyzed with phase (phase one, phase two, phase three) as a repeated measures factor. Correlations between raters for each phase were also determined.

To explore the potential effect of rater bias, all ITERs were evaluated by two separate independent physician raters who were blinded to study phase, as well as to resident and physician, at the completion of the study. The analysis above was
repeated. In addition, the CCERR ratings from the blinded raters were compared to the original CCERR scores using rater type (blind, unblind) as a repeated measures factor and correlations

Results/Analysis
There were nineteen members of the Division of Orthopedic Surgery who were involved in resident ITER completion at the initiation of the study. This included co-investigator KL; however, he was not eligible to participate, as he was a rater in the study. Two surgeons had only recently joined the division and were deemed ineligible because they did not have phase one ITERS available. All sixteen remaining members of the Division of Orthopedic Surgery at The Ottawa Hospital agreed to participate, resulting in a 100% participation rate for eligible participants. However, three participants did not submit ITERs in all phases of the study and they were therefore not included in the study analysis.

ITER Quality
The thirteen participants completed one hundred and seventy one ITERs in total. Forty-seven ITERs were submitted in phase one, forty-five ITERs in phase two, and seventy-nine ITERs in phase three. Table 1 illustrates the number of ITERs submitted by each participant and their mean total CCERR score for each phase.

Comparative Data
The mean CCERR score for phase one was 17.56 ± 1.02, phase two was 17.65 ± 0.96 and phase 3 was 17.54 ± 0.75. There was no significant improvement in CCERR scores (F(2,24) = .012, p=.98, η²p = .001). However, there was a significant interaction between rater and phase (F(2,24) = 5.013, p=.02, partial eta = .295). This interaction appears to be as a result of a difference in phase three, as compared to phase one and two. Raters were highly correlated for phases one and two (r(11) = 0.95, p<.001 and r(11)= 0.97, p<.001, respectively). In phase three there was a moderately high correlation of r(11)=0.68, p=.01. The raters were not blinded during this time, so the lower correlation raises the question of rater bias in phase three.

To address the issue of blinding, two independent blinded raters with previous experience using the CCERR rated all ITERs at once. The mean CCERR scores were 14.21 ± 0.71 (phase one), 15.11 ± 0.60 (phase two) and 14.46 ± 0.64 (phase three). The blinded raters also showed no difference in mean CCERR score between the phases. F(2,24)=1.23, p=0.31, η²p= 0.09. Unlike the unblinded raters, there was no interaction between rater and phase (F(2,24) =.532, p=.59, η²p= .04).

The CCERR scores from the blinded raters were lower than from the unblinded raters. F(1,12) = 114.08, p<.001, η²p= 0.91. However, despite the fact that they rated the ITERs lower on the CCERR than the unblinded raters, they were rating them consistently as evidenced by the strong correlations by the two groups of raters (phase 1: r(11) = 0.94, p<.001, phase 2: r(11) = 0.93, p<.001 and phase 3: r(11) = 0.93, p<.001).

Overall, there was no evidence of improvement from participants’ knowing they were being monitored (no improvement phase one versus phase two) or from the feedback provided (phase one or two versus phase three).

Discussion
The purpose of this study was to determine if a practical, low-intensity intervention provided to an entire division of clinical supervisors, with potentially differing motivations to improve, would improve ITER quality. Quite simply, it did not demonstrate any effect. No significant improvement was found from participants knowing they were being monitored or from receiving feedback in the form of their CCERR score.

Why didn’t this intervention work when studies with similar content have been successful (Dudek et al., 2012; Dudek et al., 2013a)? Is it possible that this group was simply not as good at completing quality ITERs and, therefore, not able to use the feedback like previously studied groups? The results do not support this possible explanation. The mean CCERR scores were in the poor to average quality range (poor ~16 and average ~24) (Dudek et al., 2008). However, this level of performance is consistent with the quality of ITERs completed by participants in all previous studies (ie. poor to average range) where an improvement was noted following the faculty development intervention (Dudek et al., 2008; Dudek et al., 2012; Dudek, Marks and Dojeiji 2013; Dudek et al., 2013a).

A more likely possibility is that the feedback was not timely enough. It is well known that in order for feedback to be useful it must be provided in a timely fashion (Hunt, 1992). There must also be an opportunity to use the feedback soon after receiving it (Doran, 1981). Our feedback was provided at three time points over a nine-month period; however, not all participants received three chunks of feedback, as ITER submission was variable. The variability may have occurred because they did not have a resident for a part of the year or the ITER was not completed in a timely fashion. In fact, some...
participants clumped the completion of all their ITERs for that time period together on the same day and therefore would not have had an opportunity to improve. In other situations, they may have received the feedback reasonably soon after they completed their ITER but did not have the opportunity to complete the ITER for several months and hence forgot some of the observations that may have been made on how they could improve.

There is also a chance that the amount of feedback was not adequate. First, we are assuming that the participants actually read their e-mail and reviewed their results and this may not be true. Even if they did review the CCERR, their CCERR scores and their ITERs, it is possible that it was simply not enough information to create change. A previous study demonstrated improvement in ITER quality with participants receiving the exactly the same information, so, we had anticipated similar results with this study (Dudek et al., 2013a). However, in the previous study the intervention group was mixed with some participants receiving only the above information and others receiving that information plus a generic feedback guide on how to improve ITER quality. No difference was noted between these groups. Therefore, in this study we did not include the feedback guide, as we wanted to keep things simple for the participants. It is practically important to determine the minimally effective dose of feedback since time is often cited as a reason for individuals to not participate in FD and is therefore a potential barrier when trying to deliver FD to the masses. It is notable that the previous study may have been underpowered to detect a difference between the groups. If a difference does exist it is possible that we may have been too “streamlined” in the amount of information provided.

Participants’ underlying motivation to improve may also have had an impact on the results. It is known that clinical supervisors’ underlying motivation can affect their teaching quality (Cate, Kusurkar and Williams, 2011), so why not their ITER quality? Participants seven and nine improved during phase three. Although our study was not powered to determine statistical significance at the individual level, it is possible that these differences may be related to underlying participant motivation (ie. those who improved were motivated to do so). Improving ITER quality may not have been seen as a priority for the majority of our participants amongst the many pressures in a residency program. In addition, faculty are aware that there is currently a shift toward CBME in Canadian residency programs. They may believe that other forms of assessment will ultimately replace ITERs, leading them to be less interested and motivated to change.

The self-determination theory (Deci and Ryan, 2000) proposes that a person’s behavior is determined not only by the level of motivation, but also by the type of motivation, intrinsic or extrinsic. Intrinsic motivation is when an individual pursues an activity out of personal interest and extrinsic motivation is when the activity is pursued in order to obtain a reward or avoid loss or punishment (Deci and Ryan, 2000). In this study, we did not try and specifically increase extrinsic motivation (ie. no external reward for improvement or punishment for lack of improvement) and therefore relied on participants’ baseline intrinsic motivation. It is possible that if the results were not anonymous (ie. the program director or director of assessments was aware of the individuals performance), or that there was an external incentive or punishment related to performance, that the participants’ would be more motivated to improve, potentially leading to improved results. We attempted to answer some of these questions through a post study email survey. However, we did not receive any responses.

It is relevant to note that our blinded raters were more stringent and provided lower CCERR scores than the unblinded raters; however, the numerical value of this difference was not large (ie. 2-3 points on the scale). Despite being lower, their scores were highly correlated, suggesting that they rated the ITERs consistently regardless of blinding. For formative faculty development purposes unblinded raters should be adequate.

Ultimately, there was no additional work required by the participants to participate in this intervention other than looking at their CCERR scores. The low-commitment nature of the intervention potentially did appear to overcome the issue of recruitment, a commonly cited barrier to FD participation (Rubeck and Witzke, 1998), as everyone agreed to participate.

Given the shift in medical education towards competency based medical education curricula there is a substantial need for faculty development (Dath and Iobst, 2010). As this shift involves all clinical faculty there has been a push for short bursts of point-in-time faculty development (Royal College of Physicians and Surgeons of Canada 2017). Although the ITER will likely not be a dominant part of trainee assessment, CBME curricula relies heavily on workplace-based assessment tools that follow a very similar format to ITERs (items on a checklist and narrative comments) which are known to suffer from some of the same quality concerns as ITERs (Royal College of Physicians and Surgeons of Canada 2017). As a result faculty development on assessment quality will be key. Strategies to access all clinical supervisors, rather than only those who are already interested in improving and volunteering for FD, are needed (Steinert et al., 2006). Decreasing the time commitment required to participate seems logical; however, there is likely a minimal required intensity of faculty development in order to observe positive change.
Conclusion
Overall, the results of our study raise the concern that short bursts of feedback directed at clinical teachers on relevant activities is not necessarily going to be successful. It is imperative that medical educators stop and think as we move forward with faculty development initiatives for our new CBME curricula. It is undoubtedly going to be a challenge to capture all clinical teachers and have them improve their assessment quality. As this study clearly demonstrates, it is not going to be so simple as some faculty development is not necessarily better than no faculty development. Future studies will need to determine the optimal, effective approaches to faculty development that are practical to employ with all of our clinical teachers.

Take Home Messages
- Faculty development (FD) strategies to improve assessment quality have resulted in some success but questions remain regarding the necessary motivation and time commitment by faculty to see these results.
- A simple, point-in time feedback intervention was given to all faculty in one division.
- Faculty did not improve their In Training Evaluation Report (ITER) quality despite knowing they were being monitored and despite receiving feedback.
- These results suggest that some faculty development may not necessarily be better than none.
- Medical educators should be cautious in their expectations regarding the impact of small FD strategies to a general population of faculty.

Notes On Contributors
Dr. Crawford is a Lecturer, Faculty of Medicine, University of Ottawa, Ottawa, Ontario, Canada. Orcid ID https://orcid.org/0000-0001-7794-4358

Dr. Dudek is a Professor, Faculty of Medicine, University of Ottawa, Ottawa, Ontario, Canada. Orcid ID https://orcid.org/0000-0002-6178-406X

Dr. Wood is a Professor, Department of Innovation in Medical Education (DIME), University of Ottawa, Ontario, Canada. Orcid ID https://orcid.org/0000-0001-9177-704X

Dr. Lalonde is an Assistant Professor, Department of Surgery, University of Ottawa, Ottawa, Ontario, Canada. Orcid ID https://orcid.org/0000-0002-1634-6474

Declarations
The author has declared that there are no conflicts of interest.

Ethics Statement
The Ottawa Health Science Network Research Ethics Board approved this project (20130300-01H).

External Funding
Funding support for this study was obtained from the Department of Innovation and Medical Education (DIME), University of Ottawa, Education Research Grant.

Acknowledgments
Funding support for this study was obtained from a Department of Innovation and Medical Education (DIME), University of Ottawa, Education Research Grant.

We would like to thank Johanne Samson, Dr. Vidya Sreenivasan and Dr. Scott Wiebe for their assistance with this study.

This was presented in part at the Canadian Conference on Medical Education 2017.

Crawford, KA, Dudek, ND, Lalonde, KA and Wood, TJ. Faculty Development: Is Some Always Better than None? Oral presentation. Canadian Conference on Medical Education. Winnipeg, Manitoba. 2017.
Bibliography/References

Cate, O. T. T., Kusrkara, R. A. and Williams, G. C. (2011) How self-determination theory can assist our understanding of the teaching and learning processes in medical education. AMEE Guide. No. 59. Medical Teacher. 33(12), pp. 961–973. Reference Source

Cheung, W. J., Dudek, N., Wood, T. J. and Frank, J. R. (2016) Daily encounter cards- evaluating the quality of documented assessments. Journal of Graduate Medical Education. 8(4), pp. 601–604. Reference Source

Cohen, G. S., Blumberg, P., Ryan, N. C. and Sullivan, P. L. (1993) Do final grades reflect written qualitative evaluations of student performance? Teaching and Learning in Medicine. 5(4), pp. 10–15. Reference Source

Cook, D. A., Dupras, D. M., Beckman, T. J., Thomas, K. G., et al. (2009) Effect of rater training on reliability and accuracy of mini-CEX scores: a randomized controlled trial. Journal of General Internal Medicine. 24(1), pp. 74–79. Reference Source

Dorset, D. and Jobst, W. (2010) The importance of faculty development in the transition to competency-based medical education. Medical Teacher. 32, pp. 683–686. Reference Source

Deci, E. L. and Ryan, R. M. (2000) The "what" and "why" of goal pursuits: Human needs and the self-determination of behaviour. Psychological Inquiry. 11(4), pp. 227–268. Reference Source

Doran, G. T. (1981) There's a S.M.A.R.T. way to write management's goals and objectives. Management Review (AMA FORUM). 70(11), pp. 35–36. Reference Source

Dudek, N. L., Marks, M. B. and Regehr, G. (2005) Failure to fail - the perspectives of clinical supervisors. Academic Medicine. 80(10), 584–587. Reference Source

Dudek, N. L., Marks, M. B., Wood, T. J. and Lee, A. C. (2008) Assessing the quality of supervisors' completed clinical evaluation reports. Medical Education. 42(8), pp. 816–822. Reference Source

Dudek, N. L., Marks, M. B., Wood, T. J., Dojeiji, S., et al. (2012) Quality evaluation reports - can a faculty development program make a difference? Medical Teacher. 34(11). Reference Source

Dudek, N. L., Marks, M. B., Bandiera, G., White, J., et al. (2013) Quality in training evaluation reports - does feedback drive faculty performance? Academic Medicine. 88(3), pp. 1129–1134. Reference Source

Dudek, N. L., Marks, M. B. and Dojeiji, S. (2013) Completing a quality evaluation report: what clinical supervisors need to know - a faculty development workshop. MediPORTAL Publications. Accessed October 13 2017. Reference Source

Hatala, R. and Norman, G. R. (1999) In-training evaluation during an internal medicine clerkship. Academic Medicine. 74(10), pp. 5118–5120. Reference Source

Holden, J. D. (2001) Hawthorne effects and research into professional practice. Journal of Evaluation in Clinical Practice. 7(1), pp. 65–70. Reference Source

Holumbo, E. S., Hawkins, R. E. and Huot, S. J. (2004) Effects of training on direct observation of medical residents' clinical competence - a randomized trial. Annals of Internal Medicine. 140(7), pp. 874–881. Reference Source

Hunt, D. D. (1992) Functional and dysfunctional characteristic of the prevailing model of clinical evaluation systems in North American medical schools. Academic Medicine. 67(6), pp. 254–259. Reference Source

Littlefield, J. H., DaRosa, D. A., Paukert, J., Williams, R. G., et al. (2005) Improving resident performance assessment data: numeric precision and narrative specificity. Academic Medicine. 80(5), pp. 489–495. Reference Source

Newble, D. I., Hoare, J. and Sheldrake, P. F. (1980) The selection and training of examiners for clinical examinations. Medical Education. 14(5), pp. 345–349. Reference Source

Rubeski, R. F. and Witzke, D. B. (1998) Faculty development: a field of dreams. Academic Medicine. 73(Supplement), pp. S33–S37. Reference Source

Royal College of Physicians and Surgeons of Canada. (2017) Competence by design. Faculty development. Available at: Reference Source (Accessed: October 16, 2017).

Speer, A. J., Solomon, D. J. and Arnsworth, M. A. (1996) An innovative evaluation method in an internal medicine clerkship. Academic Medicine. 71, pp. 576–578. Reference Source

Swing, S. R., Clyman, S. G., Holmboe, E. S. and Williams, R. G. (2009) Advancing resident assessment in graduate medical education. Journal of Graduate Medical Education. 1(2), pp. 278–286. Reference Source

Steinert, Y., Mann, K., Centeno, A., Dolmans, D., et al. (2006) A systematic review of faculty development initiatives designed to improve teaching effectiveness in medical education: BMJ guide no. 8. Medical Teacher. 28(6), pp. 497–526. Reference Source
Open Peer Review

Migrated Content

Version 1

Reviewer Report 29 April 2019

https://doi.org/10.21956/mep.19753.r29639

© 2019 Deshpande S. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Saee Deshpande
Maharashtra University of Health Sciences, Nashik - VSPM Dental College Nagpur

This review has been migrated. The reviewer awarded 5 stars out of 5

A very interesting and brave paper. I often experience this at our institution, faculty do not take up assessment and for that matter teaching seriously or more student centric so to say. Its disheartening in the beginning but I think its an iterative process. We should try and investigate Why's of the situation..we cant always blame on lack of intrinsic motivation, other factors such as institutional environment, peer characteristics matter as well.A well written paper, please follow up with faculty for qualitative analysis..

Competing Interests: No conflicts of interest were disclosed.

Reviewer Report 12 March 2019

https://doi.org/10.21956/mep.19753.r29640

© 2019 Van Schalkwyk S. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Susan Van Schalkwyk
Stellenbosch University

This review has been migrated. The reviewer awarded 4 stars out of 5

As faculty development continues to find traction in the context of health professions education, studies such as the one described in this article are important. It is incumbent on those who plan, design and facilitate faculty development initiatives to be critically reflective on their work, not only because it is
important to know that what they are doing is indeed ‘better’, but also because often those who engage in these initiatives do so in the context of multiple other competing demands on their time and attention. The authors are to be commended for demonstrating such criticality and for reporting ‘negative’ results – in so doing reminding us that faculty development activities are in and of themselves not necessarily always ‘good’. I would agree with the other reviewers who call for further, possibly qualitative, work in this area although it does appear that the study reported on builds on a series of early investigations. I found the framing of the intervention as ‘faculty development’ most interesting, given that its approach seemed to be largely based on the provision of feedback on practice. We are reminded of the many different ways in which our practice as health professions educators can be strengthened.

**Competing Interests:** No conflicts of interest were disclosed.

---

Thank you for inviting me to review this interesting article. The quality of workplace-based assessment is an issue indeed. In my country there are not many tutors who have some kind of pedagogical education and we do not have a tradition of giving formative feedback, either. Thus, assessment of clinical performance evokes anxiety in the clinical educators. The authors were concerned about incomplete recordings, especially when poorly performing residents were concerned. I am not familiar with the Canadian system, and I wonder if the authors had seen the actual clinical encounters or did they just evaluate the quality of the given ITER reports. Were the clinical encounters standardized or live encounters in real clinical settings? Surprisingly, the intervention was not successful and the authors discuss about the effect of underlying motivation of the participants. They gave their informed consent from the phase II onwards, but they did not volunteer to the study. The authors discuss about types of motivation – perhaps the participants did not have enough intrinsic motivation to improve. This is a great effort for faculty improvement, but it is an enormous task and I do hope that the authors keep on trying.

**Competing Interests:** No conflicts of interest were disclosed.

---
Megan Anakin
University of Otago

This review has been migrated. The reviewer awarded 4 stars out of 5

Thank you for inviting me to review this article. Like Trevor Gibbs, I found article thought-provoking. I am wondering if the authors might consider adopting a different theoretical approach to analyse the data in a future study of their intervention. I’m thinking of the work by Patricia O’Sullivan and David Irby (see: O’sullivan PS, Irby DM. Reframing research on faculty development. Acad Med. 2011. 86(4):421-8.) Perhaps taking a view about feedback in a system of assessment might also be helpful. The work of work of Liz Malloy and David Boud might provide another starting point for thinking (For example see: Boud D. Feedback: Ensuring that it leads to enhanced learning. Clin Teach. 2015. 12(1), 3-7. Boud D. Molloy E. (Eds.). Feedback in higher and professional education: understanding it and doing it well. Routledge. 2013.). I look forward to reading about the next phase of the study.

Competing Interests: No conflicts of interest were disclosed.

Trevor Gibbs
AMEE

This review has been migrated. The reviewer awarded 4 stars out of 5

Thank you for asking me to review this very interesting paper; interesting because it is brave enough to announce negative results but it creates more questions than answers. I was a little confused at first with the authors’ use of the terms assessment and evaluation- this often causes some confusion. However this was not enough to detract from the paper. We can often be blamed for assuming that faculty development is the answer for all things, it solves a major problem and by including modern methods of assessment with feedback indices a feel-good factor. This paper shows that complacency is these areas is something that we cannot afford. I feel we fail to consider internal motivation and other factors that
affect our enthusiasm to facilitate learning in our students, we expect every faculty member to be interested and passionate about teaching and put quantity before quality. I would encourage the authors to follow up this interesting paper with a more quality-based one that attempts to uncover the “feelings” of these faculty. A very important paper for all curriculum developers to read

**Competing Interests:** No conflicts of interest were disclosed.