Direct observation of clinical skills feedback scale

Halman, Samantha; Dudek, Nancy; Wood, Timothy; Pugh, Debra; Touchie, Claire; McAleer, Sean

Published in:
Teaching and Learning in Medicine

DOI:
10.1080/10401334.2016.1186552

Publication date:
2016

Document Version
Peer reviewed version

VALIDATION

Formatted: Indent: First line: 1.27 cm

Direct Observation of Clinical Skills Feedback Scale: Development and Validity Evidence

Samantha Halman,[a] Nancy Dudek,[a] Timothy Wood,[b] Debra Pugh,[a] Claire Touchie,[c] Sean McAleer,[d] and Susan Humphrey-Murto. [a]

[a] Department of Medicine, University of Ottawa, Ottawa, Ontario, Canada. [b] The Academy of Innovation in Medical Education, University of Ottawa, Ottawa, Ontario, Canada. [c] The Medical Council of Canada, Ottawa, Ontario, Canada. [d] The Centre for Medical Education, University of Dundee, Dundee, Scotland, UK.

Correspondence: Dr. Samantha Halman, MD, FRCPC, MMED, Email: shalman@toh.on.ca -The Ottawa Hospital, 501 Smyth Road, Box 209, Ottawa, Ontario, K1H 8 L6.

ABSTRACT

Construct: This article describes the development and validity evidence behind a new rating scale to assess feedback quality in the clinical workplace.

Background: Competency based medical education has mandated a shift to learner-centeredness, authentic observation, and frequent formative assessments with a focus on the delivery of effective feedback. Because feedback has been shown to be of variable quality and effectiveness, an assessment of feedback quality in the workplace is important to ensure we are providing trainees with optimal learning opportunities. The purpose of this project was to develop a rating scale for the quality of verbal feedback in the workplace (the Direct Observation of Clinical Skills Feedback Scale (DOCS-FBS)) and to gather validity evidence for its use.

Approach: Two panels of experts (local and national) took part in a nominal group technique to identify features of high quality feedback. Through multiple iterations and review, nine features were developed into the DOCS – FBS. Four rater types (residents n=21, medical students n=8, faculty n=12 and educators n=12) used the DOCS – FBS to rate videotaped feedback encounters of variable quality. The psychometric properties of the scale were determined using a generalizability analysis. Participants also completed a survey to gather data on a 5 point Likert scale to inform the ease of use, clarity, knowledge acquisition, and acceptability of the scale.

Results: Mean video ratings ranged from 1.38 to 2.96 out of 3 and followed the intended pattern suggesting that the tool allowed raters to distinguish between examples of higher and lower quality feedback. There were no significant differences between rater type (range = 2.36 to 2.49), suggesting that all groups of raters used the tool in the same way. The generalizability coefficients for the scale ranged

from 0.97 to 0.99. Item-total correlations were all above 0.80, suggesting some redundancy in items. Participants found the scale easy to use (mean=4.31/5), clear (mean=4.23/5) and most would recommend its use (mean=4.15/5). Use of DOCS-FBS was acceptable to both trainees (mean=4.34/5) and supervisors (mean=4.22/5).

Conclusions: The DOCS - FBS can reliably differentiate between feedback encounters of higher and lower quality. The scale has been shown to have excellent internal consistency. We foresee the DOCS – FBS being used as a means to provide objective evidence that faculty development efforts aimed at improving feedback skills can yield results through formal assessment of feedback quality.

BACKGROUND:

In the last decade, competency-based medical education (CBME) has taken a central role in medical training.[1] In this framework, there is a shift from traditional time-based curricula to learner-centeredness, observation of clinical skills in the workplace and frequent formative assessments with a particular focus on the delivery of effective feedback. Evidence shows that consistent systematic feedback delivered by a credible source can positively impact clinical performance and may be the most powerful influence in helping learners progress. [2-5] Simply put, feedback is the cornerstone of effective clinical teaching.[6-7]

Trainees identify feedback as an important skill their teachers should have[8] and as a means to gain expertise.[9] Although faculty realize the importance of doing it well,[10] there is a large body of evidence that demonstrates discrepancies in perceived feedback quantity and quality.[11-16] Despite the recognized importance of feedback from both teachers and learners, physicians often receive little formal training in the provision of effective feedback.[6,17] Physicians report being uncomfortable about giving feedback[18] and feel underprepared to do so.[19] It has been suggested that faculty members' skills in providing effective feedback may be enhanced through faculty development programs.[20] With competing interests and limited resources, cost-effectiveness for any faculty development program is warranted.[21] Interventions aimed at improving feedback need to be formally assessed with means to objectively measure feedback effectiveness.

Attempts have been made to develop tools and procedures to measure feedback. Unfortunately there have been issues with these attempts. For example, effectiveness cannot be reliably estimated using trainee satisfaction alone.[22] In addition, although best practice recommendations on feedback exist, they are either not empirically derived or are not linked directly to assessment of feedback quality. Also, although some tools for collecting and rating feedback have been developed, they have been criticized

for being too lengthy, having poor acceptability, and for having insufficient validity evidence to support widespread use.[23-25] Further, none have been developed specifically within the workplace-based assessment context.

The purpose of this study was to develop a simple rating scale to assess the quality of verbal feedback provided in the workplace and to gather validity evidence for its use. It was named the Direct Observation of Clinical Skills Feedback Scale (DOCS – FBS). Modern validity theory[26-28] was used as a framework to gather validity evidence for scores produced by the DOCS – FBS. To our knowledge, the DOCS – FBS is the first feedback rating scale designed for actual workplace observation.

METHODS

The study was completed in three phases. Phase 1 involved defining the features of high-quality feedback in the clinical environment. Phase 2 comprised of the development of a scale, and phase 3 the gathering of further validity evidence for the scale.

Phase 1. Features of high-quality feedback

We relied on consensus methodology with two panels of experts (local and national) to determine features of high-quality feedback in the workplace. Eight local physicians with an interest in medical education were selected using purposeful sampling ensuring representation from different training programs. All practiced within the local area (Canada) and had affiliations to the local University. Their field practice included internal medicine, family medicine, orthopedic surgery, and general surgery. All were involved in both undergraduate and postgraduate trainee supervision. As shown in Figure 1, participants took part in a consensus method known as a modified nominal group technique.[29] The session was facilitated by a skilled moderator and expert in the topic of feedback and clinical evaluation. To ensure participants had a clear understanding of the construct of interest, they were provided with a definition

of feedback specific to the medical education context: 'specific information about the comparison between a trainee's observed performance and a standard, given with the intent of improving the trainee's performance.'[30 (p.193)] Participants were asked to individually record features they felt were representative of highly effective feedback in the workplace. The moderator then asked each participant, in turn, to contribute one feature of highly effective feedback until no new features were generated (saturation). When two features were similar, they were discussed and grouped when appropriate. Features were then transcribed electronically and distributed to each participant. In this first iteration, participants were asked to anonymously rank each of the features generated by the group. Each of the individual rankings was tabulated and presented to the group. The rankings were discussed in the group setting. For each round consensus was defined as majority agreement and was achieved for all items. Specifically, consensus was not forced and the number of rounds was not determined a priori.

For the national panel, participants were recruited from a convenience sample of a group of medical educators from across the country gathered for a meeting at the Medical Council of Canada. Eight participants with a range of clinical experience (surgical and medical disciplines) and representing four provinces agreed to take part in the study. All clinicians were involved in both undergraduate and postgraduate trainee supervision to various degrees. Two participants held PhDs and conducted research in the field of medical education but were not active in clinical care. A similar process to that described for the local panel was repeated. The same moderator facilitated both sessions. After the initial item generation was completed, national participants were also provided with the list of feedback items generated by the local group. Two anonymous iterations were required to reach consensus. This panel was also asked to consider whether the identified features could be used in a rating scale but the details of the scale format were not discussed.

Phase 2. Scale development

As with any new rating scale, the number of points on each item and the anchors needed to be decided. Increasing the number of points on a scale will increase its sensitivity[31] but practically speaking, the relevance of each point can be minimized and if there are too many and small differences on the scale, may be difficult to interpret.[32] When discussing anchoring, it was felt that raters, especially junior ones, may not have had sufficient experience to discriminate between anchors like 'satisfactory' and 'excellent'. The emotional difficulty for a novice rater to rate a faculty member as 'unsatisfactory' was also considered. Further, the aim of the scale is not to identify those faculty who excel at giving feedback, but rather to identify those faculty who are performing consistently below competency~~,~~ so that they can be offered targeted training to improve their skills. We also wanted behavioral anchors for the scale because they have been shown to increase clarity[32] and inter-rater reliability.[33] The behavioral anchors were developed by the principal investigator and amended to reflect group suggestions. The items were ultimately mapped to a three-point scale with 1=not done, 2=attempted but room for improvement, and 3=successfully done.

One month after the face-to-face meeting, the completed scale was reviewed by the national panel participants to ensure that it was clear and representative of their discussions. The scale was distributed via email and participants were invited to submit their responses and suggestions electronically to the principal investigator. Finally, the scale was distributed to four postgraduate trainees in varied disciplines to gather comments regarding the ease of use and item clarity. These trainees were selected purposefully after manifesting interest in the study outcomes after it was presented at the local resident research day.

Phase 3. Gathering further Validity Evidence

In this phase we were interested in looking at the psychometric properties of the tool to gather evidence for internal structure. To do this, participants were recruited to use the DOCS – FBS to rate the quality of verbal feedback provided in six videotaped encounters which were purposefully selected to ensure at least two were examples of good quality feedback and two of lower quality. The videos were previously recorded for other purposes ~~and selected by t~~and then the principal investigator selected vidios after a search of publicly available electronic materials. All were fictional recreations of verbal feedback by a medical professional to a trainee. ~~The contexts were~~The investigator purposefully selected contexts to be broadly generalizable i.e. common general medical problems such as asthma/hypertension or education-focused e.g. resident peer teaching skills. ~~Permission to use the videos was granted by their original owners~~The original video owners granted permission to use them. Three expert raters from our group reviewed the videos and determined feedback quality to ensure some variability. ~~Consensus was confirmed by comparing~~ We compared the global DOCS – FBS scores assigned to each of the videos after group review to confirm consensus.

~~The~~ We determined the required participant sample size ~~was determined~~ by having study investigators and a research assistant individually rate each of six videotaped feedback encounters using the DOCS - FBS. Assuming a power to detect a significant difference of .80 and p=0.05, ~~it was~~we calculated that nine (9) participants per group would be required to detect a difference of 0.15 in mean video scores. ~~This value was chosen~~ We chose this value as it represents 5% of the individual item score. ~~Power calculations to detect differences between groups of raters were not performed a priori~~ as Since the focus was first and foremost on the assessment of feedback quality~~.~~, we did not perform power calculations to detect differences between groups of raters a priori.

Participants:

Investigators recruited Participants ~~were recruited~~ from four groups: residents, medical students, faculty, and faculty with direct involvement in educational activities (medical educators). Our aim was to see if different rater types produced similar ratings on the scale. Participants were recruited via email. All participants had affiliations to our University. A larger group of residents was initially recruited to account for potential dropouts but subsequent recruitment aimed only for a 30% increase in the calculated sample size (12 participants). ~~The videos were directly shown to the participants~~Participants viewed videos in groups ranging from 1 – 12 participants at a time, based on their availability. All participants viewed the videos in the same order.

Data Analysis:

~~Mean~~ We calculated mean video scores for each video ~~were calculated~~ by averaging over the ratings for raters and items. To determine the effect of feedback quality (individual video) and rater type, we calculatd a mean rating scale score ~~was determined~~ by taking the average of the item ratings that each rater assigned for each video. Comparing t~~T~~he mean ratings on the videos ~~were compared to~~ determined~~determine~~ if there were differences between the videos and if so, which ones differed. This analysis ~~was used to determine~~determined whether the DOCS – FBS was able to reliably differentiate between feedback encounters of high and low quality. A second purpose was to determine if there were differences in the ratings assigned by each rater type. Investigators conducted ~~A~~ a 6 x 4 ANOVA ~~was conducted~~ on the mean ratings with video (1 to 6) and rater type (residents, medical student, faculty and educator) treated as between subject variables. ~~The video effect was explored in more detail by conducting~~ Performing a post-hoc t-test (Bonferroni) helped to expolore the video effect in more detail. ~~Inter~~Then we calculated inter-item correlations between rater types ~~were calculated~~.

~~Reliability~~ In order to assess the reliability of the DOCS – FBS ~~was assessed using~~we conducted a Generalizability analyses. Use of this model allowed identification of the variables (i.e., videos, raters or items) that contributed the most and least to the overall variability in the scores. Using G_String,[34] a Windows interface for the urGENOVA program,[35] a repeated measures ANOVA with raters and items treated as within subject factors yielded variance components (VCs) and the percent of the variance attributed to each effect. We conducted ~~S~~separate analyses ~~were conducted~~ for each rater type. These VCs were then used to calculate the reliability (or generalizability (g) coefficient) of the instrument taking variability due to raters and items into account at the same time. The formula used to generate these coefficients is available in APPENDIX A. Using data from the G-Study, a decision study (D-study) provided reliability estimates if a different number of raters was used.

~~An item analysis was also conducted to identify~~Conducting an item analysis identified if there were any poorly performing items. We generated item scores by averaging over ratings from all of the raters for each video ~~To~~ to do this item analysis~~, item scores were generated by averaging over ratings from all of the raters for each video~~. ~~Inter item correlations were calculated~~ We calculated inter item correlations to gain a better appreciation of individual items and to see if a pattern between items could be identified.

Questionnaire:

Participants completed a six item questionnaire after the video rating session. The study team ~~T~~developed the questionnaire ~~was developed by the study team~~ to obtain demographic data and inform the utility and acceptability of the DOCS – FBS. ~~The questionnaire was~~Four residents pilot-tested ~~with four residents~~it prior to its first administration. We calculated ~~Means~~ means and standard deviations ~~were calculated~~ for each questionnaire item. To compare means between the four rater types we performed a~~A~~ one-way between subjects ANOVA ~~was used to compare means between the four rater types~~.

RESULTS:

Phase 1. Features of high-quality feedback

The local expert panel generated a list of 17 prioritized features of highly effective feedback. Items were reviewed and revised by the national panel. For example, the item 'face-to-face' identified by the local group was eliminated by the national group as it was felt to be inherent to the scale given its context. The resulting list of 12 items after all iterations is presented in Table 1.

Phase 2. Scale development

The research team reviewed the list of items generated by the national panel and eliminated items which were redundant given the purpose and context of the rating scale. For example, given that the tool is meant to be used immediately after face-to-face feedback of a direct observation, the items 'timely' and 'clearly identified as feedback' were eliminated. After review, three of the eight members of the national panel offered minor revisions on wording which were implemented. Two behavioral examples were slightly amended to reflect the comments of postgraduate trainees who piloted the scale. Instructions to the rater were added directly onto the scale.

As shown in Figure 2, the final version of the DOCS – FBS consists of 9 items, individually rated on a three point rating scale.

Phase 3. Gathering further Validity Evidence

Participants were recruited from four groups: residents (n=21), medical students (n=8), faculty (n=12) and medical educators (n=12). Although we aimed to recruit 12 medical students, only 8 completed

the study in the allotted recruitment period. All students were in clerkship. A variety of training programs (Internal Medicine, General Surgery, Radiology, Emergency Medicine, Family Medicine and Anaesthesiology) were represented for the resident group. The majority of faculty participants were from a tertiary care centre and represented a range of specialties similar to that of the resident group.

To address whether our scale can differentiate between encounters of higher and lower quality feedback, we first sought to demonstrate that there were differences in scores attributed to the different videos. Mean scores ranged from 1.38 to 2.96 out of a maximum possible score of 3.00. There was a significant main effect of video ($F_{(5,294)}$ = 211.49, p <.001, partial eta square = .78). This demonstrates that there was variability in the feedback encounters, associated with the videos. A post-hoc t-test (Bonferroni) was conducted to explore this main effect in more detail. As shown in Table 2, Video 2 had a lower mean rating than all other videos ($p < 0.001$) and Video 4, had the second lowest rating ($p < 0.001$), suggesting that the quality of the feedback provided was of lower quality. The mean rating for Video 5 was higher than ratings for all other videos ($p < 0.05$) except for video 3 ($p=0.98$), suggesting that the quality of the feedback was of higher quality. The mean ratings for the other videos fell in between. These means fell within the expected pattern based on our assessment of feedback quality in the videos.

The mean scores for rater types were quite similar (range 2.36 to 2.49) and analyses confirmed that there was no main effect of rater type ($F_{(3,294)}$ = 2.34, p=.07, partial eta square = .02). There was no interaction between rater type and video ($F_{(15,294)}$ = .89, p=.58, partial eta square = .04). To explore the ratings assigned by the four different rater types in more detail, the inter-item correlation between sets of raters were calculated. The correlations between ratings assigned by the raters were all high ($r > .97$ for all comparisons). The intra-class correlation which captures the correlation between all four rater types simultaneously was 0.995. This pattern of results combined with no significant differences between

rater types indicates that all four rater types produce similar scores while using the DOCS – FBS, suggesting that the tool is generalizable and can be used by different groups of raters.

The results of the G-study are provided in Table 3. As expected, videos (v, the object of measurement) accounted for the majority of variance in scores (49 – 57%). The small percentage of variance attributed to raters (r, 1 – 4%) and vr (6 – 9%) suggests little difference between raters. Similarly, the low percentage of variance attributed to items (i, 3 – 5%), vi (6- 9%) and ri (0 – 2%) suggests there is little difference between ratings on the items themselves.

These variance components were used to generate a g coefficient for each rater type. The generalizability coefficients for the scale ranged from 0.97 to 0.99. A D-study was then conducted to determine the reliability of the overall scale (across raters and items) if the number of raters was varied. With three raters, the scale reliability would be over 0.90 consistently.

All item-total correlations were above 0.80. All average inter-item correlations were high (> 0.80; range 0.63 – 1.00) suggesting redundancy in items.

The descriptive statistics for the questionnaire results are presented in Table 4. For visual clarity, questionnaire items have been shortened in the table. For most questionnaire items, participants tended to agree (4) or strongly agree (5) with the statements provided.

The only item whose rating was significantly impacted by rater type was the feedback knowledge acquisition item. Medical educators appeared to gain less knowledge by using the DOCS – FBS than other rater types.

DISCUSSION:

In the era of CBME, feedback and direct observation in the workplace are crucial elements to trainee assessment. As such, the ability to provide feedback is a skill which should be fostered in all faculty

members.[36] Objective measures are necessary to identify competence in feedback provision. This study has used sound consensus methodology with two panels of experts to identify the features of highly effective feedback in the workplace. Alone, these features are similar to commonly cited best practice recommendations on how to best deliver feedback. Developed into a simple rating scale, they allow an assessment of the quality of the feedback provided to a trainee. To our knowledge, the DOCS – FBS is the first feedback quality rating scale designed for direct clinical observation in the workplace.

Modern validity defines five sources of validity evidence: content, response process, internal consistency, consequences and relations to other variables. Our data suggests that the DOCS – FBS scores are supported by strong validity evidence.

Content evidence:

In tool development, content evidence is often presented as a detailed description of the steps taken to ensure the items represent the intended construct.[37] The definition we chose to outline our construct was one developed after a review of more than 100 papers within medical education[30] and thus was contextually relevant. The rigorous steps applied to the development of items through consensus methodology (including member validation) and the qualification of item writers, all of whom were experts in medical education with experience in providing feedback, all speak to content evidence.[38-39] After the initial pilot test, raters mentioned some difficulty distinguishing between scores of 2 (attempted but room for improvement) and 3 (successfully done). To enhance the discrimination of these two points, we anchored the items with behavioral examples for many items. The process of behaviourally-anchoring scales has been shown to increase inter-rater reliability.[33]

Response process:

The decision to report scores on all nine items rather than a total score or a single global rating despite item redundancy was made based on the purpose of the scale. The scale is designed to provide feedback to supervisors and we felt that using the nine items identified by the content experts would facilitate specific areas for improvement that would not be captured with just a single rating. In assessment, particularly in performance assessment, a demonstration that raters have received proper training is also important to response process evidence.[40] It would limit the feasibility of wide spread implementation of the DOCS – FBS in the clinical environment if every trainee received extensive rater training on how to complete the scale. In an effort to provide raters with some training, simple and clear instructions are included directly on the DOCS – FBS. Our data, supported by questionnaire answers, shows that the rater instructions on the scale and the behavioral anchors appeared to be sufficient to allow raters to complete the ratings.

Internal consistency:

Reliability for the DOCS – FBS was assessed using generalizability theory. The DOCS – FBS was able to reliably differentiate between videotaped feedback encounters of higher and lower quality. Internal structure evidence for the DOCS – FBS is supported by the scale's very high reliability (g-coefficient 0.97 – 0.99). Given these high reliability estimates, we suspected that there may be some item redundancy which was confirmed through an item analysis. This is not entirely surprising given the few items on the scale and the intended homogeneity to measure a unified construct. Similar to the study performed by Reiter *et al*. critiqued for its high inter-item correlation,[23] it may be that a shorter questionnaire or a global rating would be sufficient. That being said, our scale has a slightly lower average inter-item correlation across all raters with less than half the items of the Reiter *et al.* scale.[23] Items were not eliminated given the formative intent of the DOCS – FBS. Further, raters did not feel the time to complete the DOCS – FBS was excessive and there was subjective knowledge acquisition in most raters (although perhaps less in those

with more experience; *i.e.* the educators). It is unlikely this would be replicated if raters were only asked to complete a global rating of feedback quality.

In a resource-depleted medical education system, it is important to consider the optimal number of raters for a study. The results of our decision study did show that with three ratings, the overall scale reliability would consistently be over 0.90. Although performance assessments generally require 10 to 12 observations,[41] we would encourage ratings from three or more iterations of the DOCS – FBS to be considered valuable.

Consequences:

Although consequential evidence for the DOCS – FBS at this phase is limited, it was deemed to be acceptable to both trainees and supervisors that this scale be used to rate feedback quality. It is certainly possible that acceptability will be lower when applied broadly given that all participants in our study had volunteered knowing the underlying purpose. Consequential evidence will need to be revisited when the DOCS – FBS is implemented in the clinical environment.

The subjective knowledge acquisition component can also be viewed as a component of consequential evidence although it is hard to measure this objectively. Of note, it was not unexpected that medical educators may have more baseline knowledge surrounding feedback than other rater types and has such, had lower agreement that the use of the DOCS – FBS enhanced their knowledge.

Relations to other variables

Evidence relating to this last source of validity evidence was not collected during these phases of the research program but will be addressed in future directions.

Future directions

This study is limited to the Canadian context although we suspect it can be applied more broadly. Videotaped encounters allow for less variety than actual clinical encounters. Although we know the DOCS – FBS discriminates well between good and bad encounters, only its use in the clinical environment will allow an evaluation of how well it can discriminate between low, average and high quality feedback.

This program of research has predominantly focused on the role of the teacher in the feedback process. This has been criticised as an over-simplification of feedback in recent literature.[42-44] It is true that feedback has been traditionally conceptualized as a unidirectional delivery of content from a supervisor to a trainee, with little attention paid to the trainees themselves or to the nature of the relationship between the two. Certain items on the DOCS –FBS are careful to include an opportunity for the recipient to reflect and verify understanding (Items 1 and 9) but this may not be sufficient. In her review of videotaped formal feedback encounters, Molloy[45] describes our current practices in self-assessment solicitation as ritualistic or tokenistic rather than a true invitation for dialogue. The current wording of the DOCS – FBS does not allow a distinction of the degree of emphasis put on student participation during the feedback exchange. It would be interesting to have a third party observer rate the quality of student participation during feedback exchanges although we suspect results similar to those described by Molloy will emerge. Behavioural anchors on future iterations of the DOCS – FBS may be modified to better capture these nuances based on its performance in the clinical environment. It will be important when planning faculty development programs surrounding feedback to highlight the importance of encouraging true reflection rather than using the illusion of reflection as a stepping stone to a supervisor-driven monologue.

Other factors may also influence feedback from the learner's perspective. Given that source credibility can be influence by perceived clinical competence,[18] it would be interesting to study whether a feedback message with a pre-established high DOCS - FBS score would score similarly if provided by a

peer versus an experienced clinician. Relationships between the feedback provider and recipient have also been shown to influence credibility.[46-47] In light of this, would ratings be similar if provided by a supervisor with a pre-existing relationship with a trainee compared to one with whom they had a singular experience? Telio *et al*. have proposed the 'educational alliance' as a new conceptualization of feedback.[48] This framework, aligned with the therapeutic alliance used clinically, calls for a reorientation of the discussions of feedback from a focus on effective delivery and learner acceptance to a mutual negotiation within a supportive environment and educational relationship. While this framework intuitively makes sense, it may prove difficult to create faculty development opportunities around concepts such as the building of educational relationships and even harder to objectively measure these. While these concepts should certainly not be abandoned, linking them with objective measures that clinicians at large may better understand might be the best strategy to reinforce their importance. This could be explored as a means to collect evidence for relationship to other variables for the DOCS – FBS.

While we agree that the focus on feedback content and delivery in this study may be reductionist, these must occur *at least* with a minimum level of competency.[43] It is with this frame of mind that the current research focuses primarily on the assessment of feedback quality by the feedback provider as a first step to effective feedback.

CONCLUSION:

In summary, we have developed the DOCS – FBS and have shown, through various sources of validity evidence, that it is a strong tool for the assessment of feedback quality in the clinical environment. The ratings from this tool will allow performance review and can be the impetus to implement change where it is most needed. Although many questions still remain, including the weight of other factors (e.g. pre-existing relationships, source credibility, student self-reflection), ratings from this tool provide a first step in assessing a supervisor's ability to provide quality feedback to a trainee. It may also prove to be an

important tool in guiding faculty development opportunities and enhancing the feedback culture through an emphasis on its importance as CBME challenges us to provide it effectively and consistently.

REFERENCES:

1. Frank JR, Snell LS, Cate OT, Holmboe ES, Carraccio C, Swing SR, *et al.* Competency-based medical education: theory to practice. *Medical Teacher* 2010;32;8:638-45.

2. Hodder RV, Rivington RN, Calcutt LE, Hart IR. The effectiveness of immediate feedback during the ~~OSCE~~Objective Structured Clinical Exam. *Medical Education* 1989;23;2:184-8.

3. Wood BP. Feedback: A key feature of medical training. *Radiology* 2000;215;1:17-9.

4. Veloski J, Boex JR, Grasberger MJ, Evans A, Wolfson DB. Systematic review of the literature on assessment, feedback and physicians' clinical performance: BEME Guide No 7. *Medical Teacher* 2006;28:117-28.

5. Hattie J, Timperley H. The power of feedback. *Review of Educational Research* 2007;77;1:81-112.

6. Cantillon P, Sargeant J. Giving feedback in clinical settings. *The British Medical Journal* (*BMJ)* 2008 10;337:1961.

7. Hesketh EA, Laidlaw JM. Developing the teaching instinct: 1: feedback. *Medical Teacher* 2002;24:245-248.

8. Wolverton S, Bosworth M. A survey of resident perceptions of effective teaching behaviors. *Family Medicine* 1985;17;3:106-8.

9. Mann K, van der Vleuten C, Eva K, Armson H, Chesluk B, Dornan T, *et al.* Tensions in informed self-assessment: how the desire for feedback and reticence to collect and use it can conflict. *Academic Medicine* 2011;86;9:1120-7.

10. Kogan JR, Conforti LN, Bernabeo EC. Durning SJ, Hauer KE, Holboe ES. Faculty staff perceptions of feedback to residents after direct observation of clinical skills. *Medical Education* 2012;46:201 – 215.

11. Isaacson JH, Posk LK, Litaker DG, Halperin AK. Resident perceptions of the evaluation process. *Journal of General Internal Medicine* 1995;10;suppl:89.

12. Gil DH, Heins M, Jones PB. Perceptions of Medical-School faculty members and students on clinical clerkship feedback. *The Journal of Medical Education* 1984;59;11:856-64.

13. De SK, Henke PK, Ailawadi G, Dimick JB, Colletti LM. Attending, house officer, and medical student perceptions about teaching in the third-year medical school general surgery clerkship. *Journal of the American College of Surgeons* 2004;199;6:932-42.

14. Sender-Liberman A, Liberman M, Steinert Y, McLeod P, Meterissian S. Surgery residents and attending surgeons have different perceptions of feedback. *Medical Teacher* 2005;27;5:470-2.

15. Jensen AR, Wright AS, Kim S, Horvath KD, Calhoun KE. Educational feedback in the operating room: a gap between resident and faculty perceptions. *American Journal of Surgery* 2012;204;2:248-55.

16. Yarris LM, Linden JA, Gene Hern H, Lefebvre C, Nestler DM, Fu R, et al. Attending and resident satisfaction with feedback in the emergency department. *Academic emergency medicine* 2009;16;12:S76-S81.

17. Brunkner H, Altkorn DL, Cook S, Quinn MT, McNabb WL. Giving effective feedback to medical students: a workshop for faculty and house staff. *Medical Teacher* 1999;21;2:161-5.

18. Watling CJ, Lingard L. Toward meaningful evaluation of medical trainees: the influence of participants' perceptions of the process. *Advances in Health Sciences Education: Theory and Practice* 2010;17:183-94.

19. Hewson MG, Little ML. Giving feedback in medical education. *Journal of General Internal Medicine* 1998;13;2:111-6.

20. Sachdeva AK. Use of effective feedback to facilitate adult learning. *Journal of cancer education* 1996;11;2:106 – 18.

21. Meyer K. An Analysis of the Research on Faculty Development for Online Teaching and Identification of New Directions. *Journal of Asynchronous Learning Network* 2014;17;4:1 - 20.

22. Boehler ML, Rogers DA, Schwind CJ, Mayforth R, Quin J, Williams RG, et al. An investigation of medical student reactions to feedback: a randomised controlled trial. *Medical Education* 2006;40;8:746-9.

23. Reiter HI, Rosenfeld J, Nandagopal K, Eva KW. Do clinical clerks provide candidates with adequate formative assessment during Objective Structured Clinical Examinations? *Advances in Health Sciences Education: Theory and Practice* 2004;9;3:189-99.

24. Miller A, Archer J. Impact of workplace based assessment on doctors' education and performance: a systematic review. *The British Medical Journal* (*BMJ)* 2010;341: p. c5064.

25. May W, Fisher D, Souder D. Development of an instrument to measure the quality of standardized/simulated patient verbal feedback. *Medical Education Development* 2012;2;3:9 – 12.

26. Messick S. Validity. In RL Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-104). New York: NY: American council on education & Macmillan, 1989

27. AERA , APA , NCME. Standards for Educational and Psychological Testing. 1999 Available at: http://www.apa.org/science/programs/testing/standards.aspx. Accessed July 2, 2014.

28. Downing SM. Validity: on the meaningful interpretation of assessment data. *Medical Education* 2003;37:830-7.

29. Jones J, Hunter D. Consensus methods for medical and health services research. *The British Medical Journal (BMJ)* 1995;311;7001:376-80.

30. Van de Ridder JM, Stokking KM, McGaghie WC, ten Cate OT. What is feedback in clinical education? *Medical Education* 2008;42;2:189-97.

31. Cummins RA, Gullone E. *Why we should not use 5-point Likert scales: the case for subjective quality of life measurement.* Proceedings of the Second International Conference on Quality of Life in Cities; 2000; Signapore: National Universityof Singapore.

32. Krosnick JA, Pressure S. Question and Questionnaire Design. In: PV Marsden, JD Wright (Eds.) *Handbook of Survey Research* (2nd ed., pp.263-313). Bingley, UK: Emerald Publishing Group Limited, 2010.

33. Bernardin HJ, Smith PC. A clarification of some issues regarding the development and use of behaviorally anchored rating scales (BARS). *Journal of applied psychology* 1981;66:458 – 63.

34. PERD. The Program for Educational Research and Development. Available at: http://fhsperd.mcmaster.ca/g_string/index.html. Accessed June 2, 2014.

35. Brennan R. GENOVA suite programs. 2003. Available at: http://www.uiowa.edu/~casma/computer_programs.htm. Accessed June 2, 2014.

36. Richardson BK. Feedback. *Academic Emergency Medicine* 2004;11;12:189-99.

37. Haynes SN, Richard DC, Kubany ES. Content validity in psychological assessment: a functional approach to concepts and methods. *Psychological Assessment* 1995;7:238-47.

38. Downing SM, Haladyna TM. Validity. In SM Downing, R Yudkowsky (Eds.), *Assessment in health professions education*. (p.21-55) New York, NY: Routledge, 2009.

39. Creswell JW. *Research Design Qualitative and Quantitative Approaches.* 1st edition. California, USA: Sage, 1994.

40. Cook DA, Beckman TJ. Current concepts in validity and reliability for psychometric instruments: theory and application. *The American Journal of Medicine* 2006;119;166:e7-16.

41. van der Vleuten CPM, Swanson DB. Assessment of clinical skills with standardized patients: state of the art. *Teaching and learning in medicine* 1990;2:58-76.

42. Shute VJ. Focus on formative feedback. *Review of Educational Research* 2008;78:153 - 89.

43. Bing-You RG, Trowbridge RL. Why medical educators may be failing at feedback. *The Journal of the American Medical Association* (*JAMA)* 2009;302;12:1330-1.

44. Boud D, Malloy E. Rethinking models of feedback for learning: the challenge of design. *Assessment & Evaluation in Higher Education* 2012;38;6:1-15.

45. Molloy E. Time to pause: giving and receiving feedback in clinical education. In C Delany, E Molloy (Eds.) *Clinical Education in the health professions* (1st ed, pp. 128-146). Sydney, Australia: Churchill Livingston, 2009.

46. Sargeant J, Armson H, Chesluk B, Dornan T, Eva K, Holmboe E, *et al*. The processes and dimensions of informed self-assessment: a conceptual model. *Academic Medicine* 2010;85:1212-20.

47. Eva KW, Armson H, Holmboe E, Lockyer J, Loney E, Mann K, *et al.* Factors influencing responsiveness to feedback: on the interplay between fear, confidence, and reasoning processes. *Advances in Health Sciences Education: Theory and Practice* 2012;17:15-26.

48. Telio S, Ajjawi R, Regehr G. The "educational alliance" as a framework for reconceptualizing feedback in medical education. *Academic Medicine* 2015;90;5:609-14

Figure 1: Schematic representation of the modified nominal group technique used in panel discussions

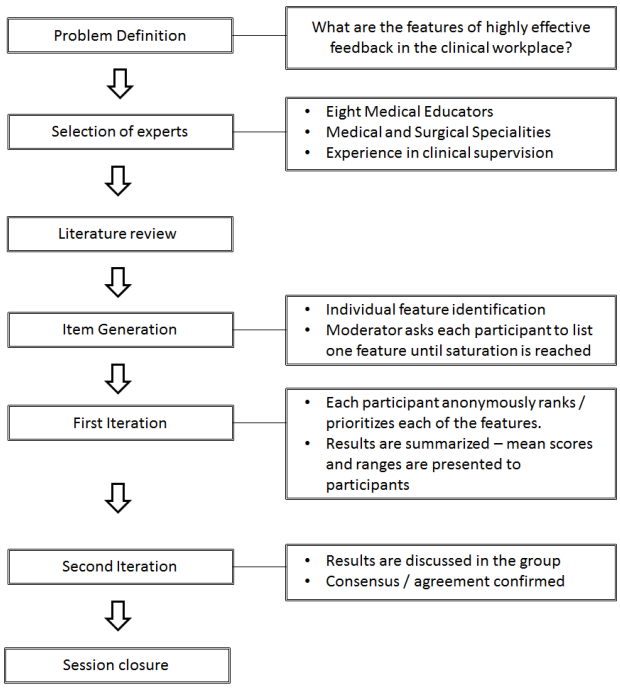

Figure 2: The DOCS – FBS

DOCS – FBS

You have just witnessed or participated in an observed clinical encounter followed by feedback. Please rate the quality of feedback provided.

Instructions to rater:

Please rate the aspects listed below using this rating scheme. Examples are provided where appropriate.

1 – Not done **2 – Attempted but room for improvement** **3 – Successfully done**

1. Offers the learner an opportunity to reflect before feedback is provided		
- 1 -	**- 2 -**	**- 3 -**
Trainee not given opportunity to reflect on performance.	Trainee asked about performance but not given opportunity to reflect. E.g. 'On a scale on 1 to 10, how do you think you did?'	Trainee allowed to reflect on performance and to discuss. E.g. 'You are right, your technique wasn't perfect. What could you have done to make it better?'
2. Feedback was provided in a respectful manner		
- 1 -	**- 2 -**	**- 3 -**
Threatening, judgmental or belittling tone.	Non-threatening tone but perhaps judgmental or provided in inappropriate environment.	Non-threatening or judgmental, preceptor adapts to trainee reactions, appropriate non-verbal language and culturally sensitive.
3. Appropriate communication style		
- 1 -	**- 2 -**	**- 3 -**
Preceptor delivers message in manner that is obviously not well understood by trainee.	Preceptor generally uses appropriate communication style but some elements lacking.	Preceptor involves trainee in conversation and adapts communication style as required.
4. Feedback focused on a specific behavior		
- 1 -	**- 2 -**	**- 3 -**
No specific behavior was identified, only general statements provided. E.g. 'You did great!'	A modifiable behavior was identified but no or limited feedback was provided. E.g. 'You should position yourself differently to auscultate'	Preceptor identifies a specific behavior and bases feedback around this. E.g. 'When auscultating for aortic regurgitation, have the patient lean forward and exhale.'
5. Feedback was constructive		
- 1 -	**- 2 -**	**- 3 -**
No suggestions geared toward identified behavior. E.g. 'Your technique was awful.'	Concise issue raised but limited suggestions provided to trainee. E.g. 'You looked very uncomfortable examining that knee'.	Concise issues identified and trainee provided with information to close a gap in knowledge. E.g. 'Your exam of the knee was very appropriate. You may be more comfortable if you position yourself this way.'
6. Ends with an action plan with goal to modify or reinforce an observed behavior		
- 1 -	**- 2 -**	**- 3 -**
Feedback terminated with no plans for follow-up or reevaluation. E.g. 'Great job!'	Broad action plan is suggested but not specific to behavior or encounter. E.g. 'Read more around your cases'	Clear plan to modify or reinforce a behavior. E.g. 'Read this article on spleen examination and I will watch you examine the next patient for splenomegaly.'
7. Limited to a manageable number of points (generally 2 – 3).		
- 1 -	**- 2 -**	**- 3 -**
No points or too many identified.	Attempted to limit to manageable number of points but room for improvement.	Limited to a manageable number of points that were appropriate for training level.
8. Appropriate time allotted to give feedback		
- 1 -	**- 2 -**	**- 3 -**
Feedback rushed or too lengthy.	Appropriate amount of time set aside but certain issues rushed or belabored.	All issues addressed with appropriate time, opportunity to address pertinent points raised.
9. Preceptor verifies understanding of feedback		
- 1 -	**- 2 -**	**- 3 -**
No verification of understanding of points raised during feedback.	Preceptor verifies understanding but does not provide adequate clarification as needed. E.g. 'Clear? Great!'	Preceptor verifies understanding and offers adequate clarification as needed. E.g. 'You say that was clear, can you summarize it for me?'

Table 1: Features of highly effective feedback in the workplace identified via consensus methodology

1.	Focused on a specific behavior
2.	Provided in a timely fashion
3.	Provided in a respectful manner
4.	Constructive
5.	Appropriate for the trainee (e.g. training level, culture)
6.	Provided in language that can be understood
7.	Appropriate and sufficient time allotted to give feedback
8.	Both the trainee and observer reflect on the encounter
9.	Limited to a manageable number of points that reflect learning objectives
10.	Clearly identified as feedback
11.	Non-normative
12.	Ends with action plan with goal to modify or reinforce an observed behavior

Table 2: Interaction between video and rater types.

	Medical Students n = 8	**Residents** n = 21	**Faculty** n = 12	**Educators** n = 12	**Total N = 53**
Video 1	2.90 (0.11)	2.72 (0.24)	2.54 (0.30)	2.75 (0.16)	2.71 (0.25)
Video 2	1.60 (0.31)	1.50 (0.23)	1.38 (0.16)	1.57 (0.23)	1.50 (0.23)
Video 3	2.86 (0.19)	2.80 (0.19)	2.86 (0.14)	2.77 (0.20)	2.82 (0.18)
Video 4	2.03 (0.56)	1.88 (0.36)	1.81 (0.41)	1.78 (0.32)	1.86 (0.40)
Video 5	2.96 (0.06)	2.92 (0.09)	2.94 (0.09)	2.87 (0.17)	2.92 (0.15)
Video 6	2.61 (0.36)	2.71 (0.24)	2.65 (0.36)	2.62 (0.45)	2.66 (0.35)

Data presented as means (standard deviation) with a maximum possible score of 3.00.

Table 3: Results of the generalizability analyses

Effect	**Medical Students**		**Residents**		**Faculty**		**Educators**	
	VC	%	VC	%	VC	%	VC	%
v	0.30	49	0.33	53	0.38	57	0.31	52
r	0.02	4	0.01	1	0.02	3	0.01	2
i	0.01	2	0.02	3	0.01	2	0.03	5
vr	0.06	9	0.04	7	0.04	6	0.04	7
vi	0.04	6	0.06	9	0.05	7	0.04	6
ri	0.00	0	0.001	2	0.01	1	0.01	1
vri	0.18	23	0.16	25	0.16	24	0.15	26

v = video, r = rater, i = item, VC = variance component, % = % variance

Table 4: Results of the questionnaire

Questionnaire item	Medical students n=8	Residents n=21	Faculty n=12	Educators n=12
Ease of use	4.25 (0.46)	4.35 (0.49)	4.25 (0.45)	4.33 (0.49)
Descriptive anchors	4.13 (0.36)	4.24 (0.62)	4.46 (0.52)	4.08 (0.79)
Acceptable Time	3.88 (0.64)	4.10 (0.83)	4.00 (1.04)	4.08 (0.79)
Recommended use	4.00 (0.53)	4.24 (0.54)	4.25 (0.62)	4.00 (0.60)
Acceptability (supervisors)	N/A	N/A	4.25 (0.87)	4.18 (0.40)
Acceptability (trainees)	4.13 (0.64)	4.43 (0.51)	N/A	N/A
Feedback knowledge	4.25 (0.46)	4.43 (0.60)	4.50 (0.52)	3.42 (1.44)

Data presented as means (standard deviation) with a maximal possible score of 5.00.

Appendix A: Formula used to generate reliability coefficients

$$\text{g-coefficient}_{\text{scale}} = \frac{\sigma_v^2 + \frac{\sigma_{vi}^2}{ni}}{\sigma_v^2 + \frac{\sigma_{vi}^2}{ni} + \frac{\sigma_{vr}^2}{nr} + \frac{\sigma_{vri}^2}{ninr}}$$

Where:

σ^2 = variance associated with videos (v), items (i) or raters (r).