Development of and Preliminary Validity Evidence for the EFeCT Feedback Scoring Tool

Shelley Ross, PhD
Deena Hamza, PhD
Rosslynn Zulla, PhD
Samantha Stasiuk, MD, MHPE
Darren Nichols, MD

ABSTRACT

Background Narrative feedback, like verbal feedback, is essential to learning. Regardless of form, all feedback should be of high quality. This is becoming even more important as programs incorporate narrative feedback into the constellation of evidence used for summative decision-making. Continuously improving the quality of narrative feedback requires tools for evaluating it, and time to score. A tool is needed that does not require clinical educator expertise so scoring can be delegated to others.

Objective To develop an evidence-based tool to evaluate the quality of documented feedback that could be reliably used by clinical educators and non-experts.

Methods Following a literature review to identify elements of high-quality feedback, an expert consensus panel developed the scoring tool. Messick's unified concept of construct validity guided the collection of validity evidence throughout development and piloting (2013–2020).

Results The Evaluation of Feedback Captured Tool (EFeCT) contains 5 categories considered to be essential elements of high-quality feedback. Preliminary validity evidence supports content, substantive, and consequential validity facets. Generalizability evidence supports that EFeCT scores assigned to feedback samples show consistent interrater reliability scores between raters across 5 sessions, regardless of level of medical education or clinical expertise (Session 1: n=3, ICC=0.94; Session 2: n=6, ICC=0.90; Session 3: n=5, ICC=0.91; Session 4: n=6, ICC=0.89; Session 5: n=6, ICC=0.92).

Conclusions There is preliminary validity evidence for the EFeCT as a useful tool for scoring the quality of documented feedback captured on assessment forms. Generalizability evidence indicated comparable EFeCT scores by raters regardless of level of expertise.

Introduction

Written feedback comments on assessment tools can inform both learning and assessment. Documentation of verbal feedback shared with a resident is valuable for subsequent reflection,[1,2] a key component of self-regulated learning.[3] Documented feedback on assessment forms is often part of the constellation of evidence that is considered by competence committees or other groups who make high-stakes or summative progress decisions.[4] In order for written feedback to be meaningful for both learning and decision-making, it must be of high quality.[5] Given the importance of feedback to learning and assessment, training programs strive to engage in continuous quality improvement (CQI) with teachers to improve the feedback captured on assessment forms.

DOI: http://dx.doi.org/10.4300/JGME-D-21-00602.1

Editor's Note: The online version of this article contains a PRISMA diagram showing the screening steps for a selection of the articles used to develop the Evaluation of Feedback Captured Tool (EFeCT) and the composition by group of the rating sessions using the final version of the EFeCT.

While some tools have been published to evaluate the quality of documented feedback, their practical value to training programs for CQI can be improved. For example, the Completed Clinical Evaluation Report Rating (CCERR) is intended as a tool for evaluating the quality of end-of-rotation high-stakes assessments.[6] The CCERR serves well for its intended purpose but is cumbersome to use for assessments that include brief captures of formative feedback. Two recently published tools, the Quality of Assessment of Learning (QuAL) score[7] and the Quality Improvement Instrument (QII),[8] also have value, but include assumptions about how feedback will be structured, which limits their generalizability. The QuAL score emphasizes the description of the resident performance and requires that feedback include a suggestion for improvement. In the case of the QII, there is an expectation that feedback will include at least one strength and at least one weakness.

These tools all share 2 common elements that may affect how they are used for CQI directed at improving the content of documented feedback. First, all 3 tools incorporate specific assumptions of the

structure of feedback, limiting their use to narratives that adhere to the assumed format. More significantly, while all 3 tools are well-suited for use by clinical educators, they may not lend themselves well for use by non-clinicians. This limits the value of these tools for CQI, as the task cannot be delegated to administrative or support staff.

The purpose of this project was to address the need for an evidence-based tool that can be used accurately by either clinical educators or support staff to evaluate the quality of feedback captured on any type of assessment form, regardless of structure. We describe the development of the Evaluation of Feedback Captured Tool (EFeCT) and present preliminary validity evidence.

Objectives
To develop an evidence-based tool to evaluate the quality of documented feedback that could be used equally reliably by clinical educators and non-experts.

Findings
Preliminary validity evidence supports content, substantive, and consequential validity facets for the Evaluation of Feedback Captured Tool (EFeCT).

Limitations
The preliminary validity evidence comes from one institution across multiple years; evidence from other institutions will be valuable.

Bottom Line
The EFeCT provides educators and researchers with an easy-to-use tool to facilitate scoring of the quality of written feedback, regardless of structure of the feedback, context, or level of learner.

Methods

Setting and Participants

Development of the EFeCT took place in a Canadian family medicine residency program and was completed in 2 iterative phases between 2013 and 2020. Participant details are included in the following phase descriptions. Validity evidence was collected and examined concurrently with both phases using the elements of Messick's unified concept of validity as a guide.[9,10]

Phase 1 used a narrative review methodology[11] with the identified purpose of mapping out common features of quality feedback as reported in original research. The findings from this literature review provided the foundation for our feedback quality evaluation tool. A literature search was carried out by a librarian using PubMed, MEDLINE, ERIC, CINAHL, Embase, Scopus, Web of Science, and PsycInfo databases from 1900 through 2013. Two members of the research team (S.S., S.R.) conducted primary screening of titles and abstracts. Inclusion criteria were descriptions of characteristics of good feedback and/or descriptions of best practices in sharing formative feedback as determined by the authors of the study. Exclusion criteria included the following: lack of descriptions outlined in inclusion criteria or referring to other studies for those descriptions rather than the study authors themselves stating their own descriptions; studies that focused on interpersonal aspects of the feedback relationship; and all studies where feedback was not the primary topic of the study.

In Phase 2, three members of the research team (S.S., S.R., S.A.S.) reviewed the full-text articles identified in Phase 1 and extracted a list of the characteristics of good formative feedback identified by the authors of the article. Two members of the research team (S.R., M.D.) discussed the lists to develop agreement about the essential elements of good formative feedback. This list was reduced using a consensus development panel methodology.[12,13] The panel included 2 master teachers[14] (M.D., D.N.) who engage in continuous self-improvement, empower students to be independent learners, and have a mindset of personal accountability as professional educators[15]; a medical student (S.S.); one current (S.A.S.) and one former (P.H.) residency program director; and a researcher with extensive graduate training and experience in education and assessment (S.R.). The panelists independently reviewed the list of characteristics of good feedback from Phase 1, and then met in person to discuss their individual perspectives about which elements from the list were essential for the evaluation of documented feedback. This resulted in a final consensus list of key characteristics of good documented feedback. The consensus decision of the panel was to put the items into question form to make the tool more intuitive to use, resulting in EFeCT.

Validity evidence gathering was based on Messick's unified concept of validity[9]; essentially, is there evidence to support the proposed interpretation of the score generated by the instrument? Validity evidence was collected concurrently with development and piloting of the tool for the following facets of Messick's unified concept: content, substantive, generalizability, and consequential validity.

The tool was initially piloted by 4 members of the research team (D.H., S.R., S.A.S., D.N.), who used the tool to individually score the documented feedback on a sample set of 100 formative narrative assessment forms (FieldNotes)[16] randomly selected from our residency program's online assessment portfolio. FieldNotes is a workplace-based assessment tool used as part of our residency program's programmatic assessment framework.[17] Each

TABLE 1
Consensus List of Elements of Good Formative Feedback Found in Published Peer-Reviewed Literature (1900–2015) and Reduced List Specific to Content of Documented Formative Feedback

Consensus List From the Literature	Revised Consensus List
Specific to an observation, not a global/overall statement	Based on (and references) an observation of a performance, skill, or demonstration of knowledge, not a global/overall statement
Based on (and references) an observation of a performance/skill/demonstration of knowledge	Focus is on encounter/task/skill/performance observed, not on learner personally
Timely (shared within a short time of the observation)	Feedback has enough information for learner to understand if they did well or need to improve
Feedback content is about the learner, not just teaching information with no relation to what or how well the learner did (connection between learner performance and feedback is explicit)	Feedback content is about the learner, not just teaching information with no relation to what or how well the learner did (connection between learner performance and feedback is explicit)
Primed, signaled, or labeled as feedback so that learner knows that it is a feedback conversation	Enough information is included in feedback to allow learner the opportunity to reflect on or respond to the feedback (specific to documented feedback)
Offers guidance ("constructive") whether feedback is reinforcing or correcting	
Focus is on encounter/task/skill/performance observed, not on learner personally	
Private (shared with learner one-on-one, not as part of a group)	
Learners are not compared to each other in the feedback	
Clearly stated, so that feedback can be easily understood	
Feedback is shared in a conversation, not a one-way talking "at" the learner	
Learner has opportunity to reflect on or respond to the feedback	
Adequate time available to have a feedback discussion	
Feedback has enough information for learner to understand if they did well or need to improve	
Respectful in delivery	

FieldNote includes a brief description of what was observed and a free-text documentation of the feedback that was shared with the resident following the observation. All FieldNotes were completed about family medicine residents across the 2 years of residency, came from a variety of supervisors, and included highly variable formats for the documented feedback. This was an intentional sampling approach as it closely replicates the situations where the tool is intended to be used. Interrater reliability was calculated using Ebel's intraclass correlation (ICC) formula[18] after the first 50 samples were scored. All score discrepancies were discussed. These discussions were used to refine the instructions before the team members scored the second 50 samples. Interrater reliability was calculated again. Following this piloting and refinement, 3 researchers (S.R., S.A.S., D.H.) scored 2 further sample sets of 100 FieldNotes, so that 3 full sample sets were available.

Evidence for content and substantive validity, as well as generalizability, was collected through multiple sessions with different groups of raters, using the 3 sample sets of 100 FieldNotes described above. Each session followed the same format: calibration, scoring, and debrief. Calibration entailed a research team member describing the tool and reviewing the instructions. Next, the participants each scored 5 samples of feedback (same samples for all participants). The facilitator then discussed the scores, with an emphasis on large discrepancies (2+ points) if any were found. The scores were reviewed in reference to the tool instructions, and the facilitator guided the group to consensus based on the instructions. The facilitator also answered any questions related to interpretation of the instructions. For scoring, each session participant independently scored the 100 samples of feedback in the set assigned to their group and gave the completed scores to the facilitator. In debrief, the facilitator gathered participants'

Instructions: This tool is for scoring the quality of written feedback captured on assessment forms. Each of the criteria below is identified as an essential element of high-quality feedback.

1. Scores for written feedback will be cumulative, resulting in a maximum score of 5. Criteria are not hierarchical; for example, it is possible to receive a score for Criterion F even if Criterion C is not met.
2. If comments are teaching only, with no reference to the performance of the learner, then the feedback ranks 0.
3. Comments that only log an encounter (with no information at all about learner involvement) rank 0.

Criterion	Element	Written Feedback	Score
A		No written feedback provided at all	0
B	***What*** did the learner do?	Some information about learner performance is provided, even if it is neutral ("diagnosed patient," "performed __"). – Using the assessment form etc as a teaching tool is **not** feedback about learner performance	1
C	Context: ***when***, ***who***, ***where*** (any or all)	There is a cue about the type of patient (ie, cardiac, mental health) and/or their demographics or context/symptoms (ie, gender, age, "demanding patient," "Type II diabetic") to help the learner remember the encounter. **OR** If feedback is about a procedure or general skill (charting, EMRs, etc), there is sufficient information about the procedure or general skill to help the learner remember the context of the feedback.	1
D	***How*** did the learner do?	The feedback specifically mentions if the task was well done, needs to be worked on, or is a concern.	1
E	***What*** was done well or needs improvement (task specificity)	There is feedback about a specific, tangible skill(s) to improve on or continue performing for future scenarios.	1
F	***How*** was it done well or ***how*** can it be improved?	If something was positive, the feedback outlines which specific element of the visit was positive. The learner should be able to replicate the specific skill in the future. **OR** There is information to guide the learner on how to do better in future (fix an error or improve performance).	1

FIGURE 1
The Evaluation of Feedback Captured Tool (EFeCT)

reflections about the tool. The facilitator asked 3 questions: (1) What was your opinion of the ease of use and utility of the tool? (2) Did the tool reflect the way that you think about good feedback? (3) Do you have any other comments about the tool? Participants were recruited from 4 groups: clinical educators, health professions students, non–health professions students, and support staff. Five sessions were held over 3 years, and each session included a mix of participants from the 4 groups (3–6 raters/session). The same set of 100 samples was used for the first 2 scoring sessions, the second sample set was used for the next 2 scoring sessions, and the third sample set was used for the last session. All participants in a specific session scored the same 100 samples, to allow for analysis of interrater reliability.

Further validity evidence was collected by using the tool for research and CQI at 2 residency teaching sites between 2015 and 2019. Evidence for generalizability was collected using the EFeCT to score real-world documented feedback generated as part of the programmatic assessment framework. Findings from the research conducted using the EFeCT were shared back to the preceptors at the teaching sites either as a simple score report or verbally as part of CQI directed at improving teaching and assessment, and the tool itself was shared to preceptors on request. At these same 2 sites, evidence for consequential validity was collected by looking for "washback," or whether introducing the scoring tool is associated with changes in quality of documented feedback,[10] by comparing mean EFeCT scores in the first and last years using independent samples *t* tests.

Additional evidence for substantive validity was collected by repeating the literature search process from Phase 1 after the EFeCT had been in use in our

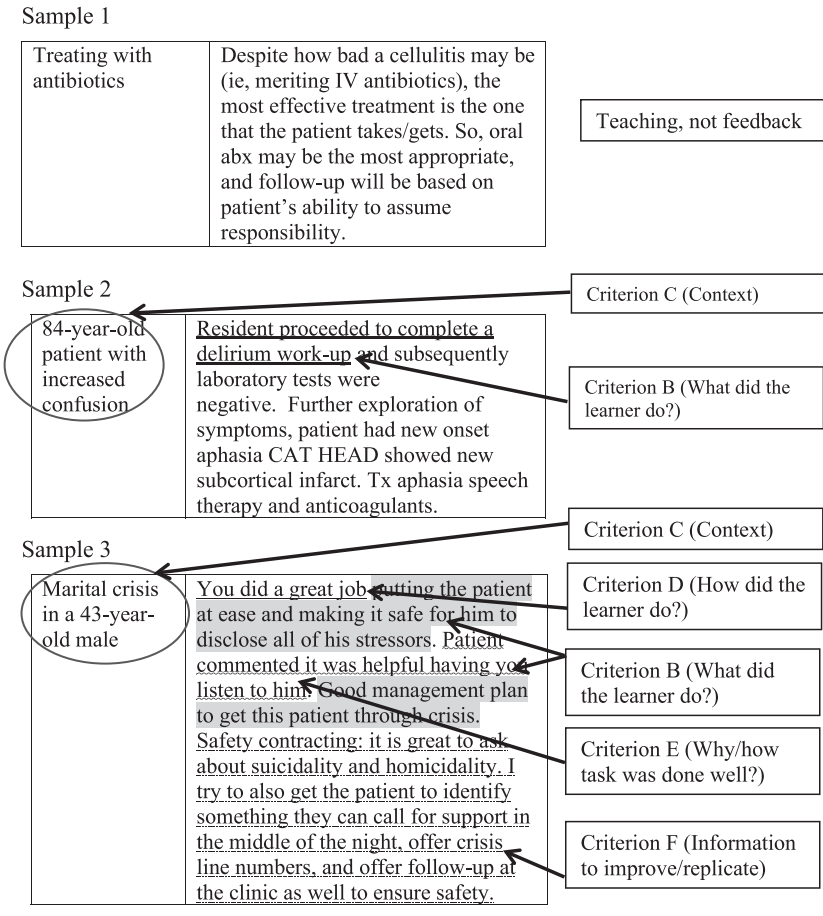

FIGURE 2
Evaluation of Feedback Captured Tool (EFeCT) Scores for Feedback Samples
Note: Sample 1: EFeCT score=0/5; Sample 2: EFeCT score=2/5; and Sample 3: EFeCT score=5/5.

program for 4 years, to ensure that the tool still reflected published literature about feedback theory.

Ethics approval was obtained from our institution's Human Research Ethics Board.

Results

The initial literature review returned 5101 records; 1307 remained after duplicates were removed. After primary screening, 104 full-text articles were retrieved for secondary screening, which resulted in a final list of 89 articles included for data extraction (see online supplementary data).

The initial list of characteristics of good feedback extracted from those articles is shown in the left column of TABLE 1. These characteristics were reviewed by the consensus development panel in Phase 2. By consensus, the panel decided to remove all elements from the original consensus list (left column, TABLE 1) that pertained to the feedback process (the actual act of sharing of feedback), rather than the documentation. Where concepts were overlapping, statements were combined. The final list of 5 key components of good feedback is shown in the right column of TABLE 1.

The final version of the EFeCT is shown in FIGURE 1. Each element of good formative feedback is phrased as a simple question. Brief descriptions of each element allow raters to decide if the element is present or absent. For each element present in an example of documented feedback, a score of 1 is given, up to a maximum score of 5. The consensus panel suggested that there should be a way to indicate which elements were present when scoring a feedback sample, so that it was clear which elements made up a final score. This resulted in a "criterion" column, with criterion A being the requirement for written feedback to be present (so that there could be a score of 0 for those instances where a sample did not have written feedback). Simple instructions clarifying how to use the tool are provided in FIGURE 1. An exemplar

TABLE 2
Evidence for Construct Validity of the EFeCT Using Messick's Unified Concept of Validity

Facet of Unified Validity	Evidence Supporting the Facet
Content (Is content relevant and representative?)	▪ Items were developed from foundational literature across disciplines. ▪ Items were refined by a consensus panel of educational experts. ▪ Items align with elements of other related tools developed and published during validation of this tool. ▪ Debriefing sessions with rater groups consistently noted that the EFeCT accurately reflected their perceptions of the elements of documentation of good feedback. ▪ Repeat of the Phase 1 literature search for articles from 2016–2019 did not result in any new components for documented feedback.
Substantive (Does the construct theory account for the content of the tool?)	▪ Consensus development panel process reviewed a deliberately expanded pool of items and selected those items that specifically represented the construct theory (TABLE 1). ▪ Debriefing sessions with rater groups indicated that raters were satisfied with the items and did not suggest removal or addition of further items. ▪ Analysis of rater responses showed consistency in how raters interpreted and applied each item in the tool.
Generalizability (Interpretation of the score generalizes across populations and settings)	▪ The samples of assessments that were used in scoring sessions included a range of formats in documenting feedback, as well as a range in the quality of the documented feedback. All samples were authentic examples of documented feedback from a health professions training program. ▪ Five rating sessions were held, and each session had raters from a broad range of backgrounds. ▪ Both in interrater reliability testing and in the debriefing sessions, there was consistency in how raters used and understood the EFeCT, as well as in how they scored the feedback samples, regardless of level of medical education expertise.
Consequential (Score interpretations may potentially influence actions, including washback, where score interpretation may influence later performance)	▪ Over a 3-year period following introduction of the EFeCT, a significant increase in the quality of documented feedback was observed from the first year to third year of use for the 2 residency program sites where the tool was used for research and quality improvement purposes (FIGURE 3).

Abbreviation: EFeCT, Evaluation of Feedback Captured Tool.
Note: Structural facet not relevant to this application as there is no weighting of items or scoring key; external facet not measured in this study.

set of sample feedback that has been scored using the EFeCT is presented in FIGURE 2.

Validity evidence is summarized in TABLE 2, where sources of evidence for each facet are described. For the piloting and refinement session, interrater reliability improved after the instructions were refined between the scoring rounds (first 50 samples: n=4, ICC=0.82; second 50 samples: n=4, ICC=0.94). Specific to the generalizability evidence, interrater reliability for the final version of the rating tool was found to be excellent for all sessions (Session 1: n=3, ICC=0.94; Session 2: n=6, ICC=0.90; Session 3: n=5, ICC=0.91; Session 4: n=6, ICC=0.89; Session 5: n=6, ICC=0.92),[14] regardless of who participated in the session (see online supplementary data) or which of the 3 sample sets was scored.

The evidence for consequential validity is shown in FIGURE 3, where the mean of EFeCT scores demonstrate a general upward trend and significant increase between the first and last years in both teaching sites.

No new information was identified related to specific components of the quality of documented/written feedback when we searched literature from 2015 to 2020. All new information about feedback that was found discussed best practices in holding feedback conversations with learners.

Discussion

The EFeCT is a rigorously developed feedback quality scoring tool that was intentionally constructed to be used by anyone—support staff or student research assistants as well as experienced clinical educators—with no compromise in the integrity or reliability of the scores produced. The collection of generalizability validity evidence for the EFeCT specifically included a diverse population of raters. While the instructions are simple (FIGURE 1), we consistently found a uniform application of the scoring tool, regardless of the experience or background of the raters. Additionally,

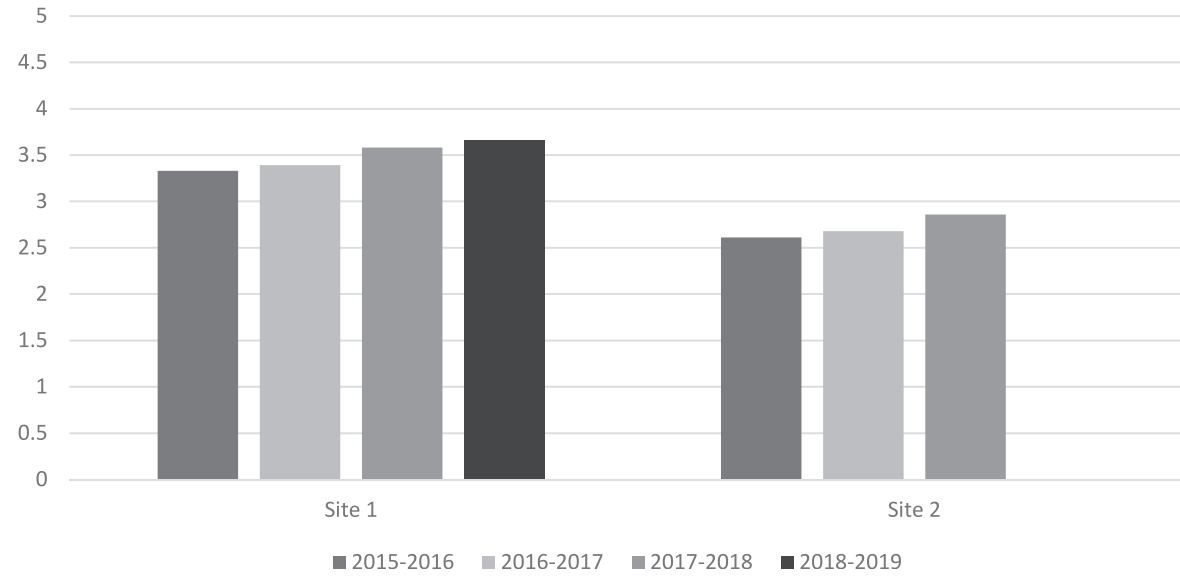

FIGURE 3
Mean EFeCT Scores for Assessments That Include Documented Feedback Across Multiple Years for 2 Major Teaching Sites

Abbreviation: EFeCT, Evaluation of Feedback Captured Tool.

Note: At both sites, mean EFeCT scores were higher in the final year (Site 1, 2018–2019: M=3.66, SD=1.55; Site 2, 2017–2018: M=2.86, SD=1.91) than in 2015–2016 (Site 1: M=3.33, SD=1.52; Site 2: M=2.61, SD=2.03). For both sites, the difference (Site 1=0.33, 95% CI 0.48- 0.18; Site 2=0.25, 95% CI 0.45-0.05) was statistically significant [Site 1: $t(1476)=4.22$, $P<.001$; Site 2: $t(1430)=2.49$, $P<.01$].

the EFeCT was designed to capture the presence of specific evidence-based elements of high-quality feedback, and to do so without requiring that the feedback be in a specific structure or format. Both of these aspects of the EFeCT differentiate it from existing tools such as the CCERR,[3] the QuAL score,[4] and the QII,[5] and make it a useful addition to these tools by offering a way to make feedback evaluation feasible and practical for programs.

Another advantage of the EFeCT is the clear and explicit language about the feedback element of "*How* was it done well or *how* can it be improved?" This element captures the importance of reinforcing a positive performance not just by saying "good job," but by articulating what it was about a task or action that went well. This helps learners to know specifically what it is that they need to keep doing, which is just as important as providing specific information on what needs to be done differently for learners who need to improve or rectify a gap. This element is also important if the feedback is being considered by a competence committee making summative decisions, as this is the type of information that gives context to ratings or scores on assessment forms.[19,20]

Interestingly, the same information is valuable for both learning and assessment. The criteria on the EFeCT are derived from published research about characteristics of feedback that positively contribute to learning. This research often addresses assessment for learning[21] and the role of formative feedback in self-regulated learning.[3,22–24] Much of this same literature has also informed assessment design in competency-based medical education (CBME) programs,[25–28] especially the incorporation of both formative and summative tools in programmatic assessment.[4,29] Summative assessment decisions in CBME are thus based on a large amount of data, including formative assessment forms containing feedback. There is clear value to including this information in summative decision-making,[5,19,20] but there is also inherent danger in doing so: the dual purposing of feedback for both formative and summative assessments may be detrimental to resident learning.[30]

Regardless of this uneasy truce between the learning and assessment uses of feedback, the fact remains that it is the quality of the feedback captured on assessment forms that determines how meaningful the feedback will be for either purpose.[5] It is therefore important for programs to have ways to monitor the quality of documented feedback beyond faculty or resident personal perceptions of good feedback, an approach that has been found to be both complex and problematic.[31,32]

The EFeCT is primarily a tool to be used to evaluate written, one-way, documented feedback captured on assessment forms. As such, it does not include measures of the process of sharing effective

feedback through verbal conversations between educators and learners, and thus does not capture the environmental, contextual, and relationship components that are key features of sharing feedback in the workplace.[28] A potential future direction for research would be to examine whether use of the EFeCT in faculty development aimed at improving the quality of captured feedback will also affect the quality of feedback conversations between teachers and learners. An additional limitation is that our validity evidence was captured at one institution. The EFeCT has been adopted at other institutions, and we are planning for formal gathering of validity evidence in the future.

Additional future research with the EFeCT is planned to compare ratings between the EFeCT, the QUaL Score, and the QII on identical sample sets of feedback. Future research could also include further exploring the generalizability of the EFeCT by using it to score feedback captured on forms in other settings, such as objective structured clinical examinations.

Conclusions

There is preliminary validity evidence for the EFeCT as a tool for scoring the quality of narrative feedback captured on assessment forms. Interrater reliability evidence showed consistent scoring by all raters who used the EFeCT, regardless of their level of expertise in medical education or experience as clinical educators.

References

1. Laughlin T, Brennan A, Brailovsky C. Effect of field notes on confidence and perceived competence: survey of faculty and residents. *Can Fam Physician*. 2012;58(6):e352–e356.
2. Quinton S, Smallbone T. Feeding forward: using feedback to promote student reflection and learning—a teaching model. *Inn Educ Teach Int*. 2010;47(1):125–135. doi:10.1080/14703290903525911
3. Nicol DJ, Macfarlane-Dick D. Formative assessment and self-regulated learning: a model and seven principles of good feedback practice. *Studies Higher Educ*. 2006;31(2):199–218. doi:10.1080/03075070600572090
4. van der Vleuten CPM, Schuwirth LWT, Driessen EW, Govearts MJB, Heeneman S. Twelve tips for programmatic assessment. *Med Teach*. 2015;37(7):641–646. doi:10.3109/0142159X.2014.973388
5. Schut S, Maggio LA, Heeneman S, van Tartwijk J, van der Vleuten C, Driessen E. Where the rubber meets the road—an integrative review of programmatic assessment in health care professions education. *Perspect Med Educ*. 2021;10(1):6–13. doi:10.1007/s40037-020-00625-w
6. Dudek NL, Marks MB, Wood TJ, Lee AC. Assessing the quality of supervisors' completed clinical evaluation reports. *Med Educ*. 2008;42(8):816–822. doi:10.1111/j.1365-2923.2008.03105.x
7. Chan TM, Sebok-Syer SS, Sampson C, Monteiro S. The quality of assessment of learning (Qual) score: validity evidence for a scoring system aimed at rating short, workplace-based comments on trainee performance. *Teach Learn Med*. 2020;32(3):319–329. doi:10.1080/10401334.2019.1708365
8. Bartlett M, Crossley J, McKinley R. Improving the quality of written feedback using written feedback. *Educ Prim Care*. 2017;28(1):16–22. doi:10.1080/14739879.2016.1217171
9. Messick S. *Validity*. ETS Research Report Series. Princeton, NJ: Educational Testing Service; 1987.
10. Messick S. Validity and washback in language testing. *Language Test*. 1996;13(3):241–256. doi:10.1177/026553229601300302
11. Ferrari R. Writing narrative style literature reviews. *Med Writing*. 2015;24(4):230–234. doi:10.1179/2047480615Z.000000000329
12. Waggoner J, Carline JD, Durning SJ. Is there a consensus on consensus methodology? Descriptions and recommendations for future consensus research. *Acad Med*. 2016;91(5):663–668. doi:10.1097/ACM.0000000000001092
13. Fink A, Kosecoff J, Chassin M, Brook RH. Consensus methods: characteristics and guidelines for use. *Am J Public Health*. 1984;74(9):979–983. doi:10.2105/ajph.74.9.979
14. Irby DM. Excellence in clinical teaching: knowledge transformation and development required. *Med Educ*. 2014;48(8):776–784. doi:10.1111/medu.12507
15. Plack MM, Goldman EF, Wesner M, Manikoth N, Haywood Y. How learning transfers: a study of how graduates of a faculty education fellowship influenced the behaviors and practices of their peers and organizations. *Acad Med*. 2015;90(3):372–378. doi:10.1097/ACM.0000000000000440
16. Donoff MG. Field notes: assisting achievement and documenting competence. *Can Fam Physician*. 2009;55(12):1260–1262.
17. Ross S, Poth C, Donoff M, et al. The Competency-Based Achievement System (CBAS): using formative feedback to teach and assess competencies with family medicine residents. *Can Fam Physician*. 2011;57(9):e323–e330.
18. Ebel RL. Estimation of the reliability of ratings. *Psychometrika*. 1951;16:407–424.
19. Ginsburg S, van der Vleuten CP, Eva KW. The hidden value of narrative comments for assessment: a

quantitative reliability analysis of qualitative data. *Acad Med*. 2017;92(11):1617–1621. doi:10.1097/ACM.0000000000001669

20. Lefebvre C, Hiestand B, Glass C, et al. Examining the effects of narrative commentary on evaluators' summative assessments of resident performance. *Eval Health Prof*. 2020;43(3):159–161. doi:10.1177/0163278718820415
21. Black P, Wiliam D. Inside the black box: raising standards through classroom assessment. *Phi Delta Kappan*. 1998;80(2):139–148. doi:10.1177/003172171009200119
22. Pereira D, Flores MA, Simão AMV, Barros A. Effectiveness and relevance of feedback in higher education: a study of undergraduate students. *Studies Educ Eval*. 2016;49:7–14. doi:10.1016/j.stueduc.2016.03.004
23. Lam R. Enacting feedback utilization from a task-specific perspective. *Curriculum J*. 2017;28(2):266–282. doi:10.1080/09585176.2016.1187185
24. ten Cate OTJ, Sargeant J. Multisource feedback for residents: how high must the stakes be? *J Grad Med Educ*. 2011;3(4):453–455. doi:10.4300/JGME-D-11-00220.1
25. Ahmed K, Miskovic D, Darzi A, Athanasiou T, Hanna GB. Observational tools for assessment of procedural skills: a systematic review. *Am J Surg*. 2011;202(4):469–480. doi:10.1016/j.amjsurg.2010.10.020
26. Harris P, Bhanji F, Topps M, et al. Evolving concepts of assessment in a competency-based world. *Med Teach*. 2017;39(6):603–608. doi:10.1080/0142159X.2017.1315071
27. Lockyer J, Carraccio C, Chan MK, et al. Core principles of assessment in competency-based medical education. *Med Teach*. 2017;39(6):609–616. doi:10.1080/0142159X.2017.1315082
28. Watling CJ, Ginsburg S. Assessment, feedback and the alchemy of learning. *Med Educ*. 2019;53(1):76–85. doi:10.1111/medu.13645
29. Schuwirth L, Ash J. Assessing tomorrow's learners: in competency-based education only a radically different holistic method of assessment will work. Six things we could forget. *Med Teach*. 2013;35(7):555–559. doi:10.3109/0142159X.2013.787140
30. Ginsburg S, Watling CJ, Schumacher DJ, Gingerich A, Hatala R. Numbers encapsulate, words elaborate: toward the best use of comments for assessment and feedback on entrustment ratings. *Acad Med*. 2021;96(suppl 7):81–86. doi:10.1097/ACM.0000000000004089
31. Boehler ML, Rogers DA, Schwing CJ, et al. An investigation of medical student reactions to feedback: a randomised controlled trial. *Med Educ*. 2006;40(8):746–749. doi:10.1111/j.1365-2929.2006.02503.x
32. van de Ridder JM, Berk FC, Stokking KM, ten Cate OTJ. Feedback providers' credibility impacts students' satisfaction with feedback and delayed performance. *Med Teach*. 2015;37(8):767–774. doi:10.3109/0142159X.2014.970617

Shelley Ross, PhD, is Professor, Department of Family Medicine, University of Alberta, Edmonton, AB, Canada; **Deena Hamza, PhD,** is Competency-Based Medical Education Evaluation Lead for Postgraduate Medical Education, University of Alberta, Edmonton, AB, Canada; **Rosslynn Zulla, PhD,** is a Specialist/Advisor, Faculty of Social Work, University of Calgary, AB, Canada; **Samantha Stasiuk, MD, MHPE,** is Clinical Assistant Professor, Department of Family Practice, University of British Columbia, BC, Canada; and **Darren Nichols, MD,** is Associate Professor, Department of Family Medicine, University of Alberta, Edmonton, AB, Canada.

Funding: Development of the Evaluation of Feedback Captured Tool (EFeCT) was assisted by a grant from the University of Alberta Faculty of Medicine & Dentistry Summer Student Medical Education Research Fund. Preparation of this manuscript was funded in part through a Department of Family Medicine Research Program graduate research assistant grant, and a grant from the Social Sciences and Humanities Research Council of Canada to the first author (Grant 435-2018-1461).

Conflict of interest: The authors declare they have no competing interests.

An earlier version of the Evaluation of Feedback Captured Tool (EFeCT), called the FFET, was presented at the Canadian Conference on Medical Education, April 20–23, 2013, Québec City, Québec, Canada.

The authors wish to thank Amy Hegstrom, Mike Donoff, Paul Humphries, Adam Kulaga, Terra Manca, and Shirley Schipper for their contributions in helping to prepare this manuscript, and Rob Woods and Andrea Gingerich for their constructive reviews of earlier drafts.

Corresponding author: Shelley Ross, PhD, University of Alberta, Edmonton, AB, Canada, sross@ualberta.ca, Twitter @S_RossUofA

Received June 8, 2021; revision received August 31, 2021; accepted November 2, 2021.