Post MD Education
UNIVERSITY OF TORONTO

Best Practices in Evaluation and Assessment (BPEA)

Assessment Fatigue

Catherine Moravac BSc, MSc
Acting Lead, Learner Education Support
Post MD Education, Faculty of Medicine, University of Toronto

Ian Brasg MD
Vice-President, Medical Education, Canadian Federation of Medical Students
Director CaRMS
Resident Physician in Adult Infectious Diseases, University of Toronto

Karl Iglar MD, CCFP
Director, Postgraduate Medical Education, St. Michael's Hospital
Associate Professor, Department of Family and Community Medicine, University of Toronto

UT CBME / CBD

Best Practices in Evaluation and Assessment (BPEA)
Assessment Fatigue

Catherine Moravac, Ian Brasg, and Karl Iglar

1. Executive Summary

The transition to competency-based medical education may involve a greater variety of assessment instruments and an increase in resident assessments, among other changes. Concerns have been raised by stakeholders that this will exacerbate endemic problems with "assessment fatigue" for various evaluators, including learners, patients, attending physicians, and allied health practitioners. In this paper a review of the academic and grey literature was performed using MeSH subject headings and proximity searches for phrases. Limited studies of relevance were returned.

The few studies and works that met face validity demonstrated that learners, as well as supervising assessors, have the highest response rates to short surveys spaced far apart, and that response rates decrease with each additional iteration, or as the academic year progresses. No relevant studies were found that addressed the fatigue that patients experience from evaluating learners. This review concludes with a discussion of the importance of providing dedicated protected time to faculty and learners for assessment while taking care to maximize impact by balancing qualitative and quantitative elements and limiting the length and frequency of administration of assessment measures. A tool is provided that will guide educators in their decisions about future assessment strategies by highlighting the potential risk for assessment fatigue. This is accomplished through consideration of the following parameters: time, frequency, content, and context.

2. Background

Over the next decade, as competency-based medical education (CBME) is developed and implemented, there will be many challenges with respect to the development, piloting, refinement, and implementation of assessment measures that are both efficient and effective. Consideration needs to be given to the feasibility of implementing assessment measures with respect to the number of items/activities, length of time for completion, frequency of use, demands on faculty time, burden on residents, and the varying contexts in which assessments will be conducted.

According to a report prepared by Dr. Terry Colbourne of Resident Doctors of Canada [1], ideal CBME models will include multiple well-developed and validated assessment tools, involving multiple evaluators. Assessment tools may include 360-degree evaluations, Objective

Structured Clinical Examinations (OSCEs), written examinations, self-assessments, direct observation, provision of formative feedback, and other strategies that require significant amounts of time and human labour.

There is growing concern about the potential impact of assessment fatigue, not only on assessors but also on residents and, in some cases, patients. A review of the existing literature was conducted in an effort to better understand this phenomenon.

3. Methodology

The goal of the literature review was to explore the topic of assessment/evaluation fatigue with respect to

i. learners,
ii. inter/intraprofessional assessors, and
iii. patients (either as assessors or as they are impacted by resident assessments conducted by others).

Routine searches for journal articles, book chapters or reports, using key words such as: "evaluation and medical residents and evaluator fatigue" did not provide meaningful results.

The following terms were entered into MeSH on Demand (2016): assessment of medical residents; evaluation of medical residents; assessor fatigue and medical residents; evaluator fatigue and medical residents; interprofessional assessors and medical residents; intraprofessional assessors and medical residents; interprofessional evaluators and medical residents; intraprofessional evaluators and medical residents; and evaluation burnout. An additional search with the following MeSH headings and Boolean terms failed to yield results: [evaluation studies OR surveys and questionnaires OR self-assessment] AND [students, medical OR internship and residency OR health occupations] AND [fatigue OR burnout, professional OR apathy]. The National Library of Medicine's MeSH on Demand made the following recommendations for MeSH headings: i) Fatigue, and ii) Physicians. The subcategories of these MeSH tree structures were similarly unhelpful. Fatigue often referred to patient symptoms of fatigue.

A Google Scholar search conducted with the key words "evaluator fatigue and medical residents" produced a series of articles that used the word fatigue; however, assessor's fatigue was not the primary focus of any of the articles that were subsequently reviewed.

Finally, the grey literature was assessed, using Scopus and ProQuest and conducting proximity searches for phrases relating to formal MeSH subject headings.

4. Results and Discussion

Although "evaluator/evaluation fatigue" and "assessor/assessment fatigue" are commonly used and understood terms, they are not recognized subject headings. Consequently, our literature searches were mostly unsuccessful. Despite this, in those instances where the targeted terms did appear in academic journal articles, we summarized the context in which they were used.

4.1. Learner Fatigue (Resident/Assessor Fatigue)

Some articles referenced learner fatigue within a singular evaluation-based event rather than with respect to the longitudinal fatigue that may be experienced when residents are assessed frequently over an extended period of time.

A study by Gamboa-Salcedo[2] for example, referenced the fatigue experienced by both professors and students after an OSCE that lasted two hours and twenty minutes. A Chicago-based study[3] that sought to assess the feasibility of implementing an evaluation of an online peer hand-off tool referenced the negative impact of resident fatigue post-call on the quality of hand-offs. The authors found that the tool was also too lengthy. Since internal consistency was high, they were ultimately able to reduce the number of survey items to 5. A 2009 study conducted at McGill University Faculty of Medicine[4] comparing simulation-based multiple mini interviews (MMI) with traditional unstructured interviews for medical student applicants found that both students and evaluators favoured the MMIs. They noted, however, that there was a need to improve the process, with attention being paid to evaluator fatigue, evaluator preparation, and other operational issues.

Other articles located were likewise not focused directly on the topic of interest. Porter et al.[5] found a relationship between inter-survey time and response rate, demonstrating that frequency is proportional to non-response. An additional study by Meterko et al.[6] used a survey with progressive response waves to demonstrate that similar results could be achieved with a lower response rate compared to a higher one, an important finding when fatigue is unavoidable.

4.2. Interprofessional / Intraprofessional Assessors

In 2009–2010 a group of physician researchers at New York Medical College attempted to develop milestones for an internal medicine residency program. They subsequently surveyed faculty, residents, and clinical competency committee members about the acceptability of the milestones. Faculty reported that the high number of milestones per form/tool was problematic. Comments included:

> The use of too many milestones per form may lead to evaluator fatigue, loss of evaluator interest or other problems which might hinder effective evaluation and feedback.[7]

An article by Tsue, Dugan, and Burkey[8] provided an excellent overview of the issues to be considered when assessing surgical competency. After outlining the ideal elements, the authors conclude that, in reality, no ideal assessment exists that fulfils all the required elements. Of

interest, however, was the suggestion that assessments be combined with learning activities for higher efficiency and that learners could be involved with the development and implementation of the assessment process. While evaluator fatigue was not directly mentioned in this article, it was implied in the discussion of feasibility issues.

Kendal and Baker[9] commented on the development of assessment tools for plastic surgery competencies in 2015. This work involved the development of milestones and entrustable professional activities. The authors concluded: "What we failed to realize was how onerous this data collection would be for our faculty. In the end, we failed our residents, as we did not have a tool that could be used by our faculty in a timely and effective manner." The authors then posed the question: "Recognizing the breadth of our specialty and the need to evaluate milestones and EPAs, how do you prevent or avoid evaluator fatigue, given the number of data points required?"

In 2012–2013 researchers from Canada and the U.S.[10] collaborated on the development of assessment tools for plastic surgery competencies, initially focusing on two complex procedures: breast augmentation (17 steps) and facelift (24 steps). In this study the authors stated:

> From a practical standpoint, if all of these steps were evaluated each time, assessment of technical competency would be an onerous process. … It may be more feasible to use this tool for more formal entrustable decisions, such as acknowledging a resident successfully passing the threshold allowing for decreased supervision.[10]

Farim et al.[11] from McMaster University in Hamilton developed a tool to provide formative feedback to residents on grand rounds presentations and piloted it in an academic surgical department. The tool contained 13 Likert-scale questions, accompanied by two open-ended questions, covering all seven of the CanMEDS competencies. Evaluation fatigue was identified as a major barrier to evaluators' willingness to provide effective feedback. One staff participant stated:

> Don't make the [evaluation] tool too cumbersome or too vague that people don't take it seriously or don't actually want to do it.[11]

Similar outcomes were found by another group of North American researchers attempting to evaluate how residents performed paediatric emergency airway simulation.[12] Both faculty availability and reviewer fatigue were found to be significant barriers to implementation. Chen et al[13] reported on their attempts to assess consistency of OSCE raters evaluating orthopaedic residents. They acknowledged that OSCEs require extensive resources and time, and have been known to cause patient and examiner fatigue.

Observer fatigue was also briefly noted in a study involving the development and piloting of a knowledge and competency test for birth attendants working in Benin, Ecuador, Jamaica, and Rwanda.[14] Similarly, researchers testing the validity and utility of a 360-degree evaluation tool for an American residency program in dermatology found that they had to shorten the

assessment due to evaluator fatigue.[15] Ultimately, only one Likert-scale response was used for each of the six competencies ([1] medical knowledge, [2] patient care, [3] practice based learning and improvement, [4] interpersonal and communication skills, [5] professionalism, and [6] teamwork (systems-based practice).

In a study published by researchers at the University of Arizona, 47 students were evaluated 547 times by 46 residents and attendings in end-of-shift emergency department evaluations.[16] Five-point Likert scales were used to evaluate each of six categories: energy/interest, fund of knowledge, judgement/problem solving ability, clinical skills, personal effectiveness, and systems-based practice. While the results indicated that the timing of the evaluation did not affect scoring, the authors commented that:

> In addition, evaluations are more delayed later in the academic year. This is possibly due to fatigue of the evaluators, and could confound the association between timing and evaluation score.[16]

An American group reported on the development of two new OSCEs for physical medicine and rehabilitation residents, along with an examination of inter-rater reliability.[17] Outcomes indicated that inter-rater grading of history-taking was good; however, inter-rater reliability of small-joint and large-joint examination was problematic due to examiners' inability to have full visibility, evaluators' fatigue, and confusing evaluation scoring descriptions.

Finally, a study by Higgins, et al.[18] reported on the implementation of a 360-degree feedback tool used to evaluate the Accreditation Council for Graduate Medical Education (ACGME) general competencies in a cardiothoracic surgery residency program. The authors noted that the survey was limited to 45 items in an effort to reduce evaluator fatigue. Even so, there were 12 to 15 evaluators assigned to each resident.

Wang et al.[19] examined the effect of fatigue on the accuracy of standardized patients' checklist recording in a pilot project evaluating the clinical competence of foreign medical graduates. Standardized patients (SPs) interacted with as many as 21 examinees over a 12-hour period. The encounters were video-recorded for later review. The authors concluded that fatigue did not appear to systematically affect the accuracy of SPs' checklist recording.

4.3. Patient Assessors

There is no published literature that we are aware of which directly addresses patient evaluator fatigue.

A 2002 paper exploring the considerations that need to be taken into account when developing assessment methods concerning ACGME general competencies noted that previous studies have indicated that between 20 and 50 patients would be required to provide a stable rating of residents' humanistic qualities. The author concludes by stating:

> OSCEs and SP exams seem to be the best methods for conducting assessments for high stakes decisions. [20]

Two years later, Wood et al.[21] reached a different conclusion. Their research consisted of a pilot study of a 360-degree evaluation of radiology residents, pertaining to professionalism and interpersonal/communication skills, involving faculty and patient assessors as well as resident self-assessments. The research team concluded that this assessment tool was reliable and valid for the evaluation of resident competencies in those two domains. They noted, however, that future research would be needed to determine how frequently the tool should be used, how to overcome administrative challenges, and how to reduce the burden experienced by faculty. The majority of residents in this study valued the receipt of patient feedback and indicated that they would incorporate new insights into future practice. Thomas et al.[22] also found that residents valued the perceptions of patients concerning humanistic qualities over and above assessments of their clinical skills.

Brittlebank et al.[23] explored the use of eight types of assessments for psychiatry residents, including patient satisfaction surveys. They found that peer assessments and patient satisfaction measures had lower levels of reliability than the other measures used. The authors estimated that five patient satisfaction surveys would be required for each resident to reach a "sufficiently precise" evaluation, and that further research would be needed to effectively incorporate the patient perspective into resident assessment.

Tamblyn et al.[24] also explored the feasibility of using patient satisfaction ratings to evaluate internal medicine residents. They found that patient satisfaction ratings provided valuable information about a resident's ability to establish effective physician-patient relationships; however, like Swing et al.[20] the authors concluded that the high number of ratings that would be required to obtain a reliable assessment of resident skills precluded the inclusion of patient ratings in future assessment programs.

A 2012 study conducted in Australia to explore the congruence between staff psychiatrists' and psychiatric patients' evaluations of learners found poor correlation between these two assessor groups when using the same evaluation questions.[25] The authors concluded:

> Although our findings of discrepant judgements between patients and examiners are congruent with other findings using SPs in OSCEs, our findings of congruence in more technical aspects of assessment raise further questions as to the appropriateness of excluding patients from contributing to formative assessments of trainees …[25]

5. Discussion and Conclusions

Our understanding of assessment fatigue is that in situations where assessors are burdened by lengthy assessment tools and/or high volumes of required assessments, the burdens cumulatively cause a form of "burnout." Burnout results in resistance or refusal to complete

evaluations, or rushed and poorly completed evaluations, any of which will defeat the objective of providing meaningful assessment feedback to learners. The proportion of faculty/clinician time that has been allocated to teaching and evaluation can directly affect the risk of burnout; therefore, this should be kept in mind when developing assessment tools and their proposed frequency of use.[26] Medical faculty stress and burnout also directly impact on the level of willingness to implement curricular change.[6] As CBME is being developed, a preventive approach to the experience of stress and burnout among faculty is warranted.

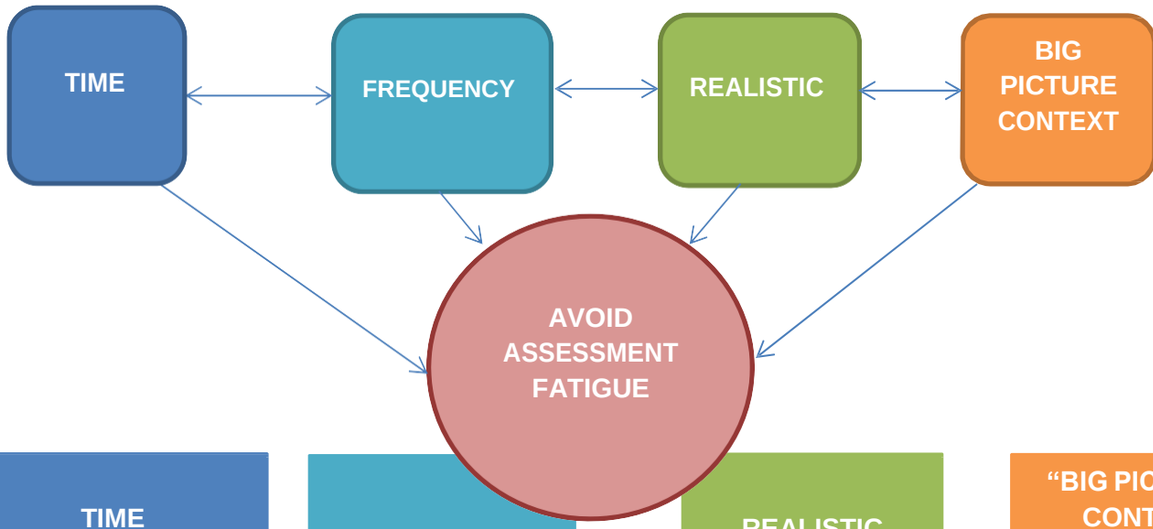
With respect to ITERs, going forward it will be important to limit the number of items, carefully consider the frequency of use, and also consider the administrative burden of managing the data points. Use of current technologies will help with efficiencies.

For Multiple Source Feedback (MSF) forms, a limit of approximately 12 items may be appropriate to avoid assessor fatigue. Consideration should be given to the number of assessors required and the frequency of implementation. Thought could be given to the number of MSF forms and other evaluations an assessor may be required to complete over the course of an academic year. Likewise, consideration should be given to the time required to provide verbal feedback to the learner in a timely fashion.

The development of OSCEs should occur with evaluator fatigue in mind, along with several other important variables. The number of performance items ideally should be pared down to a reasonable number of essential competencies. Consideration should be given to the number of residents an evaluator is expected to assess within a pre-established period of time. Cognitive fatigue could impact the quality of scoring on residents evaluated toward end of the session compared to those evaluated at the beginning of the session. The potential for standardized patient fatigue should also guide decisions about the scheduling of successive scenarios. Planned provision of routine and regular in vivo formative feedback to residents may allow for a decrease in the number of summative evaluations.

Currently there is no gold standard with respect to the balance of quantitative and qualitative items within assessment/evaluation tools. Short, quantitative measures are more likely to be completed accurately: however, learners do benefit from qualitative feedback. It may be appropriate to limit the number of qualitative items on a mixed evaluation to perhaps two. Receipt of routine informal formative feedback during clinical experiences may reduce the need for summative qualitative feedback. Further research is needed to guide these decisions. In summary, when designing new assessment tools, consideration should be given to time, frequency of use, the “fit,” and how realistic the assessment plan is, in the context of the “big picture” (Figure 1).

Figure 1



TIME

How long will it take to complete the assessment tool?

How many items are included in the assessment (can this # be reduced?)

How many other assessments does this evaluator have to complete during the academic year?

Does the evaluator have sufficient protected time for this activity?

How long will it take to upload/record the data points?

Is there a technological tool available to decrease administrative burden?

Has time been factored in to provide timely verbal or written feedback to the resident and a person assigned to do this?

Has thought been given to the potential for evaluator/SP fatigue?

Has the "big picture" context been considered?

FREQUENCY

How often does this assessment tool need to be administered? (Can it be less often?)

Does this assessment tool need to be implemented during each block or could another strategy be used?

Has the "big picture" context been considered?

REALISTIC

Is the evaluation plan realistic?

Has the math covering what will be required to successfully implement the plan been done?

Have the data points been reduced down to only the essential competencies that need to be evaluated?

Are the data points specific, measurable, and realistic in the "big picture" context?

Are sufficient administrative supports in place to fully implement the evaluation plan?

"BIG PICTURE" CONTEXT

Has consideration been given to the cumulative evaluation plan over the entire learning/training period? (Would this allow for less frequent summative evaluations?)

Has consideration been given to the amount of dedicated time evaluators have for this task and how many other evaluations they are expected to cumulatively complete on X number of learners?

Have other creative approaches to evaluation been considered within the global learning context?

Is the evaluation plan supported by available literature in best practices?

6. References

1. Colbourne T. Mitigating Risks during Implementation of Competency-Based Medical Education in Canadian Residency Programs. *International Conference on Residency Education (ICRE)* 2015; http://residentdoctors.ca/wp-content/uploads/2015/10/Oct-24-2015-Colbourne-RDoC-Mitigating-Risks-of-CBME.pdf.
2. Gamboa-Salcedo T, Martinez-Viniegra N, Pena-Alonso YR, Pacheco-Rios A, Garcia-Duran R, Sanchez-Medina J. Objective structured clinical examination as an instrument for evaluation of clinical competence in pediatrics. A pilot study. *Boston Medical Hospital Infant Mex.* 2011;68(3):169-176.
3. Arora VM, Greenstein EA, Woodruff JN, Staisiunas PG, Farnan JM. Implementing peer evaluations of hand-offs: associations with experience and workload. *Journal of Hospital Medicine.* 2013;8(3):132-136.
4. Razack S, Faremo S, Drolet F, Snell L, Wiseman J, Pickering J. Multiple mini interviews versus traditional interviews: stakeholder acceptability comparison. *Medical Education.* 2009;43:993-1000.
5. Porter S, Whitcomb M, Weitzer W. Multiple Surveys of Students and Survey Fatigue. *New Directions for Institutional Research.* 2004;121:63-73.
6. Meterko M, Restuccia JD, Stolzmann K, et al. Response rates, nonresponse bias, and data quality: Results from a national survey of senior healthcare leaders. *Public Opinion Quarterly.* 2015;79(1):130-144.
7. Nabors C, Peterson SJ, Forman L, et al. Operationalizing the Internal Medicine Milestones–An Early Status Report. *Journal of Graduate Medical Education.* 2013;5(1):130-137.
8. Tsue TT, Dugan JW, Burkey B. Assessment of surgical competency. *Otolaryngologic Clinics of North America.* 2007;40:1237-1259.
9. Kenkel JM, Baker SB. Commentary on: The development of assessment tools for plastic surgery competencies. *Aesthetic Surgery Journal.* 2015;35(5):618-620.
10. Courtea BC, Knox AD, Vassiliou MC, Warren RJG, M S. The development of assessment tools for plastic surgery competencies. *Aesthetic Surgery Journal.* 2015;35(5):611-617.
11. Fahim C, Bhandari M, Yang I, Sonnadara R. Development and early piloting of a CanMEDS competency-based feedback tool for surgical grand rounds. *Journal of Surgical Education.* 2016;73(3):409-415.
12. House JB, Dooley-Hash S, Kowalenko T, et al. Prospective comparison of live evaluation and video review in the evaluation of operator performance in a pediatric emergency airway simulation. *Journal of Graduate Medical Education.* 2012;4(3):312-316.
13. Chen ACY, Lee MS, Chen WJ, Lee ST. Assessment in orthopedic training – an analysis of rating consistency by using an objective structured examination video. *Journal of Surgical Education.* 2013;70(2):189-192.

14. Harvey SA, Ayabaca P, Bucagu M, et al. Skilled birth attendant competence: an initial assessment in four countries, and implications for the safe motherhood movement. *International Journal of Gynecology and Obstetrics.* 2004;87:203-210.
15. Wu S, Jenkins SN, Hoffman ER, Morrel DS. Validity and use of the 360-degree evaluation tools in a dermatology residency program. *Clinical Research Dermatology Open Access.* 2014;1(1):1-4.
16. Hiller KM, Waterbrook A, Waters K. Timing of emergency medicine student evaluation does not affect scoring. *The Journal of Emergency Medicine.* 2016;50(2):302-307.
17. Jain SS, DeLisa JA, Eyles MY, Nader S, Kirshblum S, Smith A. Further experience in the development of an objective structured clinical examination for physical medicine and rehabilitation residents. *American Journal of Physical Medicine and Rehabilitation.* 1998;77(4):306-310.
18. Higgins RS, Bridges J, Burke JM, O'Donnell MA, Cohen NM, Wilkes SB. Implementing the ACGME general competencies in a cardiothoracic surgery residency program using 360-degree feedback. *Annals of Thoracic Surgery.* 2004;77(12):7.
19. Wang Y, Stillman P, Stunick AI, Ben-David MF, Williams RG. The effect of fatigue on the accuracy of standardized patients' checklist recording. *Teaching and Learning in Medicine.* 1996;8(3):148-151.
20. Swing SR. Assessing the ACGME General Competencies: General considerations and assessment methods. *Academic Emergency Medicine.* 2002;9(11):1278-1288.
21. Wood J, Colllins J, Burnside ES, et al. Patient, faculty, and self-assessment of radiology resident performance: a 360-degree method of measuring professionalism and interpersonal/communication skills. *Academic Radiology.* 2004;11(8):931-939.
22. Thomas PA, Hellmann DB. Residents' attitudes towards patient feedback. *Academic Medicine.* 1999;74:90-91.
23. Brittlebank A, Archer J, Longson D, Malik A, Bhugra DK. Workplace-based assessments in psychiatry: evaluation of a whole assessment system. *Academic Psychiatry.* 2013;37(5):301-307.
24. Tamblyn R, Benaroya S, Snell L, McLeod P, Schnarch B, Abrahamowicz M. The feasibility and value of using patient satisfaction ratings to evaluate internal medicine residents. *Journal of General Internal Medicine.* 1994;9(3):146-152.
25. Mellsop GW, MacDonald J, El Badri S, Menkes D. Patients' appraisal of psychiatric trainee interview skills. *Academic Psychiatry.* 2012;36(5):374-379.
26. Lackritz JR. Exploring burnout among university faculty: incidence, performance, and demographic issues. *Teaching and Teacher Educatioin.* 2004;20:713-729.

7. Appendix 1: Annotated Bibliography

Arvandi, Z., E. Amirhossein, N. Zarghi, S. M. Alavinia, M. Shirazi and S. V. Parikh (2016). "Linking medical faculty stress/burnout to willingness to implement medical school curriculum change: a preliminary investigation." Journal of Evaluation in Clinical Practice 22: 86-92.

This team measured stages of change as well as faculty stress/burnout. They found that occupational burnout directly reduced readiness to change. The authors emphasize the importance of assessing and managing occupational burnout among faculty prior to the implementation of academic reform in medical schools.

Nabors, C., S. J. Peterson, L. Forman, G. W. Stallings, A. Mumtaz, S. Sule, T. Shah, W. Aronow, L. Delorenzo, D. Chandy, S. G. Lehrman, W. H. Frishman and E. Holmboe (2013). "Operationalizing the Internal Medicine Milestones–An Early Status Report." Journal of Graduate Medical Education 5(1): 130-137.

This article reports on a milestone-based evaluation process conducted in 2010 with general medicine and pulmonary-critical care rotations at New York Medical College. Outcomes examined acceptability of the process to faculty members, residents, and clinical competency committee members via survey method. While there was agreement across all three user groups about the advantages of utilizing milestones and their propensity to bring fairness and uniformity to the evaluation process, faculty found the use of too many milestones problematic and expressed concerns about the potential for evaluator fatigue. Another significant challenge was negotiating and settling on an acceptable balance of inclusion and exclusion of milestones. The authors also note that the implementation of this process may not be as practical in other settings without the benefit of strong faculty support.

Fahim, C., M. Bhandari, I. Yang and R. Sonnadara (2016). "Development and early piloting of a CanMEDS competency-based feedback tool for surgical grand rounds." Journal of Surgical Education 73(3): 409-415.

This article describes users' reactions to a formative feedback tool for grand rounds that was piloted in the Department of Surgery at McMaster University. Of note, in the results section, concerns were raised about evaluation fatigue.