Canadian Journal of Emergency Medicine (2023) 25:453–455
https://doi.org/10.1007/s43678-023-00489-z

EDITORIAL

Vigilance and validity: the necessity of assessment system surveillance to ensure equity in emergency medicine

Teresa Chan[1,2,3,4,5,7] · Simiao Li-Sauerwine[6] · Sandra Monteiro[1,2,3] · Quang Ngo[3,7,8]

Published online: 2 June 2023

Dayal and colleagues found persistent differential attainment of emergency medicine (EM) milestones between male and female residents amongst a large cohort of training programs, suggestive of significant bias [1]. Since then, many groups have looked to their own contexts to seek out the presence or absence of these gender gaps [2]. To this end, we would like to applaud Ingratta et al. in this issue of CJEM for contributing to this important work, given the emphasis many institutions are placing on this topic and measures to try and counteract these biases [3]. This team sought to examine whether gender differences existed in the quality of their workplace-based assessments (WBAs). This work is important as part of a process of continuous quality improvement of a program of assessment as issues pertaining to equity, diversity and inclusion (EDI) require high-quality data as an input.

Right tools for the right purpose

In their study, Ingratta and colleagues evaluated the hypothesis that gender of either trainee or faculty member would influence the quality of narrative feedback for WBAs or the variability of O-EDShOT scores. The authors proposed that the presence of bias could be inferred if there was a systematic difference in quality ratings (as per a tool called the Completed Clinical Evaluation Report Rating, or CCERR) or numeric scores for one gender of residents, or by one gender of faculty, or an interaction between faculty and resident genders—for example, if men faculty consistently rate women trainees lower than men trainees.

Using the CCERR, which has validity evidence in evaluating the quality of end-of-shift WBAs, they found no difference in the quality of feedback between male and female faculty being given to male and female residents, as evidenced by aggregate CCERR scores for the individual O-EDShOT tools. This finding is encouraging, suggesting that men and women EM faculty at the University of Ottawa are providing feedback at similar levels of quality as measured by the CCERR. However, we should note that the detection of bias requires tools that look for bias and as the authors have already acknowledged, the CCERR was not designed with this purpose in mind. None of the nine items in the CCERR seek to draw judgements or conclusions about the presence or absence of gendered language or numerical bias, which have been shown to exist within EM WBAs [1, 2, 4].

✉ Teresa Chan
teresa.chan@medportal.ca

Simiao Li-Sauerwine
simiao.li@gmail.com

Sandra Monteiro
monteisd@mcmaster.ca

Quang Ngo
qngo@mcmaster.ca

[1] Department of Medicine, McMaster University, Hamilton, ON, Canada

[2] Department of Health Research Methodologies, Evidence, and Impact (HEI), McMaster University, Hamilton, ON, Canada

[3] McMaster Education Research, Innovation, and Theory (MERIT) Unit, McMaster University, Hamilton, ON, Canada

[4] Continuing Professional Development, Faculty of Health Sciences, McMaster University, Hamilton, ON, Canada

[5] The Chinese University of Hong Kong, Shenzhen, China

[6] Emergency Medicine, The Ohio State University, Columbus, USA

[7] DeGroote School of Medicine, McMaster University, Hamilton, ON, Canada

[8] Department of Pediatrics, Faculty of Health Sciences, McMaster University, Hamilton, ON, Canada

Similarly, other tools that have been designed to measure quality of feedback such as the QuAL score and EFFecT score also fail to build in bias detection [5, 6].

Given prior work showing that women are more likely than men to receive feedback that is not actionable, it is curious that the authors did not explore if each of the nine elements were addressed differently by men and women faculty. For example, were there variations in how well faculty explained examples of weakness? It may be that this exploration was deemed inappropriate as there were no between group differences. But it is also possible that despite having similar total scores, the individual elements of the CCERR were influenced by bias.

Striving for high-quality feedback and avoidance of bias in any form is imperative. While we all should strive to the level of rigor that Ingratta et al. has achieved in their context, we echo the authors' acknowledgment regarding the limited generalizability of the results and recommend caution when designing similar studies—those based on assumptions of correlations between unrelated measures. Specifically, we advise researchers to think about the specific characteristics of bias they might expect to find in their data, rather than examine overall trends in mean scores.

We worry that the CCERR may not be sufficiently sensitive to detect the types of bias that contribute to gaps in the feedback provided to trainees or passed on to their competency committees. The literature reflects that not all levels of a program of assessment will contain bias [7]; thus, our search for bias must be systematic and inclusive of all the links in the chain of evidence we create about our trainees.

Offramps and detours from the best laid plan: lessons for educators and leaders

A recent scoping review showed that unstructured workplace-based assessments contain more gender bias than procedure, simulation and competency committee deliberations [2]. The hypothesis is that perhaps more structured assessments scaffold raters towards better decisions. The study by Ingratta and colleagues suggests that perhaps tools like O-EDShOT may provide more structure and, therefore, help to overcome gender biases.

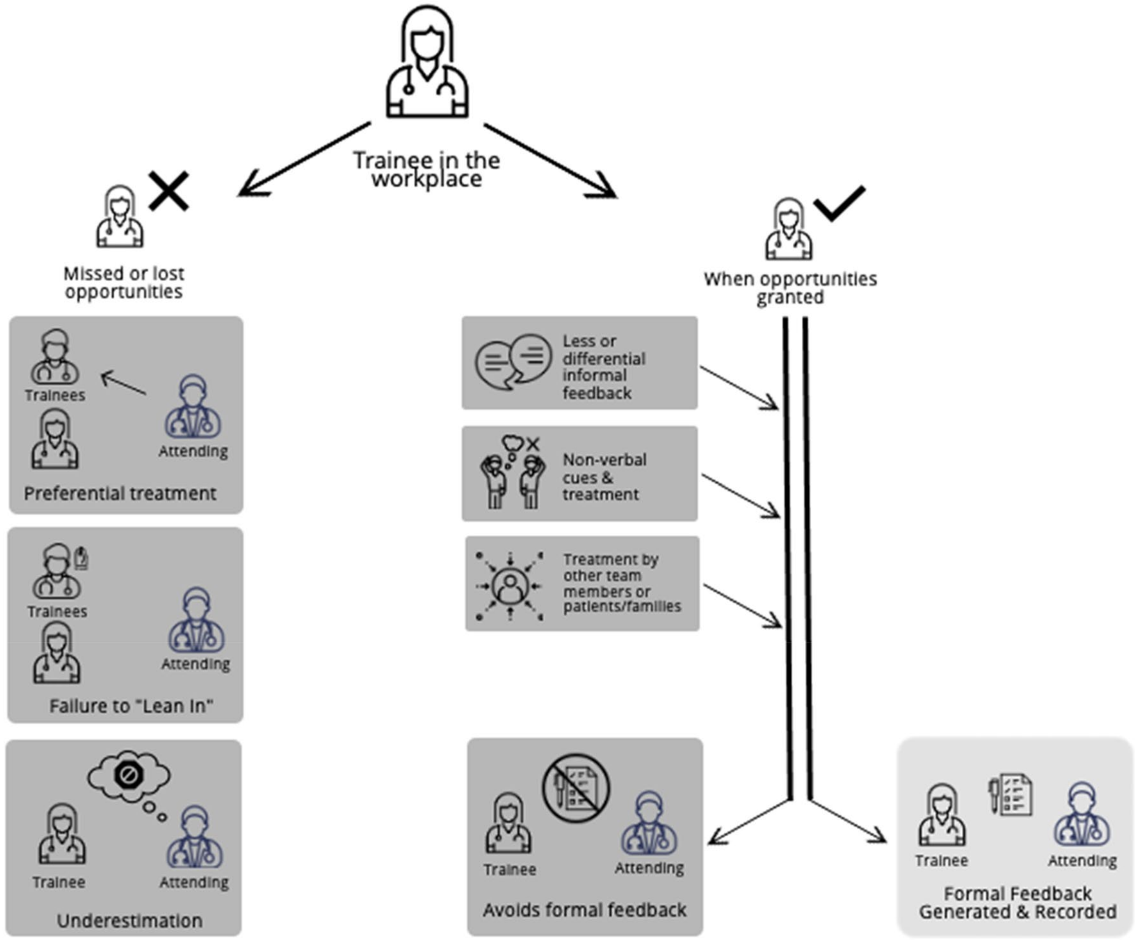
Trainee in the workplace
Missed or lost opportunities
Trainees
Attending
Preferential treatment
Trainees
Attending
Failure to "Lean In"
Trainee
Attending
Underestimation
When opportunities granted
Less or differential informal feedback
Non-verbal cues & treatment
Treatment by other team members or patients/families
Trainee
Attending
Avoids formal feedback
Trainee
Attending
Formal Feedback Generated & Recorded

Fig. 1 Diagram depicting possible off-ramps that would preclude a trainee from garnering formal written assessments and feedback

However, the origins of bias within an assessment can be insidious and exist anywhere along the chain of events leading up to the triggering of the assessment, rather than just within the assessment itself. The devil is in the details, and in this case, so too are the origins of bias. As with most retrospective studies, the study by Ingratta and colleagues is limited by the data that have been captured. Much of the bias within workplace-based assessment systems may lie in the minutes leading up to the data being entered into the system. All such retrospective studies have notable losses in data even before beginning. Leaning into the example of gender-related bias and how it might intersect with bias within the system, Fig. 1 depicts the "off-ramps" within the system where high-quality data may just never be written down or captured, therefore, precluding its inclusion in this particular type of study. This response process problem represents one of the criticisms of CBME, highlighting the gap between how an assessment tool is used and their intended use; in this case, how trainees and faculty might interact with the assessment tool to gather data. When the system is poorly designed, there can be data loss in the system due to poor user experience and response process errors. While the paper by Igratta et al. starts to examine the terminal part of the larger process (i.e., the quality of the comments that make it into an end-of-shift assessment tool), mapping out the assessment process including any off-ramps allows us to better understand how gender (or other differences) may result in divergent assessment experiences.

Studies like Ingratta et al.'s represent surveillance of programs of assessment, and should target data quality, data loss, differential opportunities, differential scoring amongst other aspects to generate feedback within the system for continuous quality improvement of local processes and faculty development. Not only measuring the quality of those assessments for both bias in the data that we have but also identifying the data that are missing is crucial for equity. We must all engage in the hard work of surveilling our assessment systems.

Author contributions QN, SM, SLS, and TMC all contributed substantially to the authorship of this work and will stand by its contents post publication.

Funding None for this study.

Declarations

Conflict of interest Dr. Ngo receives stipends from McMaster University for his educational roles as program director and chair of assessment of the DeGroote School of Medicine. Dr. Li-Sauerwine is the founder of the Academic Life in Emergency Medicine (ALiEM) Education Research Lab and Incubator, and serves as a Core Mentor for the ALiEM Faculty Incubator, for which she receives an honoraria. She serves on the Grant Advisory Board and Delphi Team for an American Medical Association Reimaging Resident Grant, for which she receives a stipend. She discloses that she has received speaker fees from academic institutions (University of Utah Health). Dr. Chan reports a honoraria from McMaster University for her education research work with the McMaster Education Research, Innovation, and Theory (MERIT) group and administrative stipend for her role of associate dean via the McMaster Faculty of Health Sciences Office of Continuing Professional Development. She discloses that she has received various unrelated research grants, teaching honoraria, and speakership fees from academic institutions (Baylor University/Texas Children's Hospital, Catholic University of Korea, Taiwan Veteran's General Hospital, Prince of Songkla University, Harvard Medical School, International Association of Medical Sciences Educators, Ontario College of Family Physicians, Northern Ontario School of Medicine, University of British Columbia, University of Northern British Columbia, Holland Bloorview). Dr. Chan has received non-relevant grants from PSI Foundation, physician organizations (Association of American Medical Colleges, Canadian Association of Emergency Physicians, Society of Academic Emergency Medicine, the Royal College of Physicians and Surgeons of Canada, Medical Council of Canada), and governmental sources (Government of Ontario, Virtual Learning Strategy eCampus Ontario program). Dr. Monteiro is a Faculty of Health Sciences MERIT Scientist and Associate Professor and receives salary support from the Department of Medicine. She has a second appointment to the Centre for Simulation Based Learning and receives salary support as the Director of Simulation Scholarship. In a prior secondment to Touchstone Institute, (2014–2021), she served as the Director of Research and Analysis, and received salary support for contributing to the development of several competency based assessments for various regulated healthcare professions in Canada. Dr. Monteiro holds a contract with the Medical Council of Canada to evaluate the test security and fairness of online examinations in a high stakes licensing context.

Ethical approval N/A.

References

1. Dayal A, O'Connor DM, Qadri U, Arora VM. Comparison of male vs female resident milestone evaluations by faculty during emergency medicine residency training. JAMA Intern Med. 2017;177:651–7.
2. Menchetti I, Eagles D, Ghanem D, Leppard J, Fournier K, Cheung WJ. Gender differences in emergency medicine resident assessment: a scoping review. AEM Educ Train. 2022;6: e10808.
3. Ingratta J, Dudek N, Lacroix L, Cortel-LeBlanc M, McConnell M, Cheung WJ. Exploring gender influences in the quality of workplace-based assessments. Can J Emerg Med. 2023.
4. Brewer A, Osborne M, Mueller AS, O'Connor DM, Dayal A, Arora VM. Who gets the benefit of the doubt? Performance evaluations, medical errors, and the production of gender inequality in emergency medical education. Am Sociol Rev. 2020;85:247–70.
5. Ross S, Hamza D, Zulla R, Stasiuk S, Nichols D. Development of and preliminary validity evidence for the EFeCT feedback scoring tool. J Grad Med Educ. 2022;14:71–9.
6. Chan TM, Sebok-Syer SS, Sampson C, Monteiro S. Reliability and validity evidence for the quality of assessment for learning (QuAL) score. Acad Emerg Med. 2018;25:S83.
7. Klein R, Julian KA, Snyder ED, Koch J, Ufere NN, Volerman A, et al. Gender bias in resident assessment in graduate medical education: review of the literature. J Gen Intern Med. 2019;34:712–9.