

RESEARCH ARTICLE

 OPEN ACCESS  Check for updates

Assessment methods and resource requirements for milestone reporting by an emergency medicine clinical competency committee

Nikhil Goyal ^{a,b}, Jason Folt ^a, Bradley Jaskulka ^a, Sudhir Baliga^a, Michelle Slezak^a, Lonni R. Schultz^c and Phyllis Vallee^a

^aDepartment of Emergency Medicine, Henry Ford Health System, Detroit, MI, USA; ^bDepartment of Internal Medicine, Henry Ford Health System, Detroit, MI, USA; ^cDepartment of Public Health Sciences, Henry Ford Health System, Detroit, MI, USA

ABSTRACT

Background: The Accreditation Council for Graduate Medical Education (ACGME) introduced milestones for Emergency Medicine (EM) in 2012. Clinical Competency Committees (CCC) are tasked with assessing residents on milestones and reporting them to the ACGME. Appropriate workflows for CCCs are not well defined.

Objective: Our objective was to compare different approaches to milestone assessment by a CCC, quantify resource requirements for each and to identify the most efficient workflow.

Design: Three distinct processes for rendering milestone assessments were compared:

- (1) Full milestone assessments (FMA) utilizing all available resident assessment data,
- (2) Ad-hoc milestone assessments (AMA) created by multiple expert educators using their personal assessment of resident performance,
- (3) Self-assessments (SMA) completed by residents.

FMA were selected as the theoretical gold standard. Intraclass correlation coefficients were used to analyze for agreement between different assessment methods. Kendall's coefficient was used to assess the inter-rater agreement for the AMA.

Results: All 13 second-year residents and 7 educational faculty of an urban EM Residency Program participated in the study in 2013. Substantial or better agreement between FMA and AMA was seen for 8 of the 23 total subcompetencies (PC4, PC8, PC9, PC11, MK, PROF2, ICS2, SBP2), and for 1 subcompetency (SBP1) between FMA and SMA. Multiple AMA for individual residents demonstrated substantial or better interobserver agreement in 3 subcompetencies (PC1, PC2, and PROF2). FMA took longer to complete compared to AMA (80.9 vs. 5.3 min, $p < 0.001$).

Conclusions: Using AMA to evaluate residents on the milestones takes significantly less time than FMA. However, AMA and SMA agree with FMA on only 8 and 1 subcompetencies, respectively. An estimated 23.5 h of faculty time are required each month to fulfill the requirement for semiannual reporting for a residency with 42 trainees.

ARTICLE HISTORY

Received 11 August 2018
Revised 13 October 2018
Accepted 15 October 2018

KEYWORDS




Accreditation; graduate medical education; milestones; assessment; cost; clinical competency committee

Introduction

The Accreditation Council for Graduate Medical Education (ACGME) introduced the Next Accreditation System (NAS) in 2012, which is based on a continuous accreditation model with assessment of residents along educational milestones [1]. The NAS requires that training programs establish a Clinical Competency Committee (CCC) to assess each resident's performance on the milestones and reports this data to the ACGME semiannually. Milestones are defined as 'competency-based developmental outcomes that can be demonstrated progressively by residents and fellows from the beginning of their education through graduation to the unsupervised practice of their specialties,' and amongst other things, are intended to provide 'a rich descriptive, developmental framework for CCCs' [2]. Milestones allow residency faculty to report their observations of resident performance without abstraction of

performance data into large categories such as the 6 core competencies. In Emergency Medicine (EM), there are 227 milestones arranged in 23 subcompetencies (Supplemental Table 1) and each subcompetency has 5 levels of achievement [3,4]. Unlike many training programs where residents may spend a large portion of the day away from supervising faculty, most emergency departments have residents working side by side with faculty throughout the clinical shift. This close interaction affords faculty the ability to assess many of the EM milestones on a continuous basis. As a result, it may be possible to determine milestone achievement levels using the global assessments of faculty, rather than a resource-intensive process that collects data from multiple sources.

The Henry Ford EM Residency Program has been continually accredited since 1982 and utilized robust competency-based resident assessment tools typical of a large urban EM residency (Table 1). In 2013,

CONTACT Nikhil Goyal  med@spandan.com  Henry Ford Hospital, CFP-259, 2799 West Grand Boulevard, Detroit, Michigan 48202, USA
 Supplemental data for this article can be accessed here.

© 2018 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Table 1. Assessment methods used in our training program prior to implementation of the next accreditation system.

Assessment type	Description
End of shift evaluation	Descriptive text on assessment by faculty at the end of every shift in the Emergency Department. Focus on 'things done well' and 'areas needing improvement.' This is based on direct observation of resident performance by faculty, and constitutes the majority of assessment data.
Monthly rotation evaluation	Competency-based assessment completed at the end of each clinical assignment. Includes EM and non-EM rotations and may be completed by faculty, fellows or senior residents.
Procedure Log	Resident-created log of all procedures performed during the training period
Direct observation checklist	Checklist of behaviors demonstrated by a resident during a single patient encounter, when directly observed by EM faculty. Includes a mandatory summative 'satisfactory/unsatisfactory' designation and optional free-text comments.
Monthly staff meeting resident review	Summary comments from a group discussion on each resident's performance during the monthly EM departmental faculty meeting
Quarterly REACH Dashboard [14]	Color-coded dashboard that summarizes recent clinical assessments, examinations and administrative requirements (such as USMLE Step 3 completion, duty hour logging, etc.)
Monthly education committee resident review	Detailed group discussion of overall resident performance and new resident issues by EM educational faculty
In-training examination [15]	National, standardized test for EM residents administered by the American Board of EM
Grand rounds presentation evaluation	Assessment of content and delivery of presentation at weekly EM conference. Each resident is scheduled to present annually.
QI and scholarly activity requirement	Review of resident publications, scholarly activity, QI projects or participation in patient safety activities
Simulation center sessions	Residents are evaluated in a simulated patient encounter; activities include procedure training, mock codes, breaking bad news to patients, error disclosure, cultural competency, etc.
Unsolicited feedback	Positive or negative feedback provided by patients, nurses, ancillary staff, peers, faculty, etc.

EM, Emergency Medicine; QI = Quality Improvement; USMLE, United States Medical Licensing Examination

the upcoming implementation of the NAS created an opportunity for system change. We designed this study to explore different options for the CCC process. Our overall goal was to identify the most resource-efficient way to determine milestone achievement levels for reporting to the ACGME. Specifically, we completed resident milestone assessments based on utilization of all existing data and compared it to resident self-assessment and 'ad hoc' assessment (i.e., based on faculty members' personal recollection of resident performance). Factors contributing to incongruence between assessments are explored and recommendations for CCC structure are provided.

Our primary objective was to check for agreement between comprehensive data-driven milestone assessments, faculty opinion ('ad hoc') assessments and resident self-assessment. The secondary objective was to quantify the faculty time needed to execute an effective CCC.

Materials and methods

Seven faculty members with a wide range of experience volunteered to participate in this study. The program director was not included in the group as it was felt that he had extensive preexisting knowledge on assessment data for each resident. The Henry Ford Hospital Institutional Review Board approved the study and waived the need for informed consent.

The study was conducted during the 2013–14 academic year, as our program entered the NAS and experimented with different CCC workflows. PGY1 residents in our program were excluded because they spend a majority of their time on non-EM rotations that provide limited assessment data. PGY3 residents

were also excluded because they were scheduled to graduate in a few months and formal milestone assessments were not needed. Therefore, our group focused the study on all PGY2 residents ($n = 13$) in our 3-year program (EM1-EM3). Three distinct assessments were created, each utilizing the EM Milestones document [3]:

- (1) **Full Milestone Assessment (FMA):** A single FMA was created for each resident by a single faculty member. This assessment was intended to be as detailed as possible and would be considered the theoretical 'gold standard' assessment. The faculty member was asked to utilize all available data (Table 1) to objectively determine the performance level for each of the 23 subcompetencies; they were asked to refrain from using their personal opinion on resident performance when adjudicating these milestones. Faculty also recorded the time taken to produce the FMA. A total of 13 FMA were created, 1 on each resident.
- (2) **Ad hoc Milestone Assessment (AMA):** Six residents were randomly selected for AMA. These assessments were completed from memory and the faculty were asked to not refer to the resident file or any other objective assessment data. The faculty member subjectively scored the resident's performance using only their own previous clinical interactions with the resident and their personal recall of any resident assessment data. All seven members of the faculty group completed an independent AMA on the same six residents; a faculty member who had previously completed an FMA on a particular resident did not complete an AMA on the same resident. Therefore, we

had six independently created AMA on each of the six residents. As before, each faculty member recorded the time taken to complete this exercise.

- (3) **Self-Milestone Assessment (SMA)**: Each resident assessed and scored their own performance using the EM milestones document.

The only specific training provided to the faculty and residents for completing assessments was a detailed review of the instructions provided on page *v* of the ACGME EM milestones document. These instructions delineate when to select a specific score at or between the defined 1–5 levels for each subcompetency [3].

Following completion of the milestone documents, the faculty met as a group to review the assessments and discuss how they adjudicated each subcompetency for the FMA.

Statistical analysis

Intraclass correlation coefficients (ICC), along with 95% confidence intervals, were computed to assess agreement between FMA and SMA for the subcompetency measures. These values range from 0 to 1, with high values indicating consistent responses between the faculty and residents and low values indicating inconsistent responses. Landis and Koch provide interpretation for levels of agreement using ICC [5]. They proposed values 0 as poor, 0.01 to 0.2 as slight, 0.21 to 0.4 as fair, 0.41 to 0.6 as moderate, 0.61 to 0.8 as substantial, and over 0.8 as almost perfect agreement.

To assess the inter-rater agreement of AMA among the six faculty members, Kendall's coefficients with corresponding p-values were computed. Kendall's coefficients were used instead of Kappa statistics because it takes into account the scoring of the AMA, with values from 0 to 5.

A two-sample *t*-test was used to compare the time taken for AMA vs. FMA. All analyses were done using SAS version 9.2 (SAS Institute Inc., Cary, NC, USA).

Results

The volunteer faculty group consisted of seven educational faculty, of which two were associate program directors for the residency program and the others were active members of the Education Committee. Characteristics of the faculty group are described in Table 2.

Comparison of AMA and FMA

Six AMA and one FMA were completed on each of the six residents. Table 3 contains the ICC with confidence intervals for assessing agreement between the FMA and AMA. There was almost perfect agreement for PC8 and

Table 2. Faculty ($n = 7$).

Characteristics	
Age, years, mean (range)	41 (32–56)
Sex, <i>N</i>	Female: 3; Male: 4
No. of years of experience evaluating residents, mean (range)	11 (3–27)

Table 3. Inter-rater agreement between FMA and AMA ($n = 6$ residents).

Subcompetency	ICC (95% CI)	Level of agreement
PC1	0.47 (0.07, 0.92)	Moderate
PC2	0.55 (0.12, 0.92)	Moderate
PC3	0.47 (0.07, 0.92)	Moderate
PC4	0.71 (0.27, 0.94)	Substantial
PC5	0.21 (0.00, 0.96)	Fair
PC6	0.49 (0.08, 0.92)	Moderate
PC7	0.59 (0.14, 0.93)	Moderate
PC8	0.88 (0.57, 0.98)	Almost perfect
PC9	0.77 (0.35, 0.95)	Substantial
PC10	0.58 (0.14, 0.92)	Moderate
PC11	0.91 (0.66, 0.98)	Almost perfect
PC12	0.28 (0.01, 0.94)	Fair
PC13	0.00 (NA, NA)	Poor/none
PC14	0.43 (0.05, 0.92)	Moderate
MK	0.80 (0.40, 0.96)	Substantial
PROF1	0.04 (0.00, 1.00)	Slight
PROF2	0.64 (0.19, 0.93)	Substantial
ICS1	0.46 (0.06, 0.92)	Moderate
ICS2	0.63 (0.18, 0.93)	Substantial
PBLI	0.41 (0.04, 0.92)	Moderate
SBP1	0.11 (0.00, 1.00)	Slight
SBP2	0.63 (0.18, 0.93)	Substantial
SBP3	0.02 (0.00, 1.00)	Fair

AMA, ad hoc milestone assessment; FMA, full milestone assessment; ICC, intraclass correlation coefficient; NA, no agreement

PC11 and substantial agreement for PC4, PC9, MK, PROF2, ICS2 and SBP2. Moderate agreement was observed for PC1, PC2, PC3, PC6, PC7, PC10, PC14, ICS1 and PBLI. Fair agreement was observed for PC5 and PC12, slight agreement was observed for PROF1, SBP1 and SBP3 and poor to no agreement for PC13.

AMA interobserver variation

Kendall's coefficients were computed for inter-rater agreement among the six AMA completed for each resident. There was substantial agreement for PC1, PC2 and PROF2 and moderate agreement for PC3, PC4, PC5, PC7, PC8, PC9, MK, ICS1 and ICS2. Fair agreement was observed for PC6, PC10, PC11, PBLI, SBP1, SBP2 and SBP3 and slight agreement for PC12, PC13, PC14 and PROF1. Data are summarized in Table 4.

Resident self-assessment

There were 13 residents with both SMA and FMA. Table 5 contains the ICC with confidence intervals for each of the 23 subcompetencies. There was substantial agreement for subcompetency SBP1 and moderate agreement for ICS1, ICS2, PBLI and SBP3. Fair agreement was observed for PC4, PC5, PC6, PC11, PC13 and SBP2 and slight agreement for PC3, PC8 and PC12.

Table 4. Inter-rater agreement for AMA ($n = 6$ residents).

Subcompetency	Kendall's coefficient	p -value	Level of agreement
PC1	0.63	< 0.001	Substantial
PC2	0.71	< 0.001	Substantial
PC3	0.45	0.009	Moderate
PC4	0.60	< 0.001	Moderate
PC5	0.42	0.016	Moderate
PC6	0.34	0.056	Fair
PC7	0.44	0.011	Moderate
PC8	0.60	< 0.001	Moderate
PC9	0.45	0.009	Moderate
PC10	0.24	0.212	Fair
PC11	0.30	0.096	Fair
PC12	0.19	0.365	Slight
PC13	0.17	0.441	Slight
PC14	0.26	0.170	Slight
MK	0.53	0.001	Moderate
PROF1	0.21	0.292	Slight
PROF2	0.72	< 0.001	Substantial
ICS1	0.53	0.001	Moderate
ICS2	0.43	0.012	Moderate
PBLI	0.37	0.037	Fair
SBP1	0.33	0.067	Fair
SBP2	0.24	0.209	Fair
SBP3	0.27	0.143	Fair

AMA, ad hoc milestone assessment; ICC, intraclass correlation coefficient.

Table 5. Inter-rater agreement between FMA and SMA ($n = 13$ residents).

Subcompetency	ICC (95% CI)	Level of agreement
PC1	0.00 (NA, NA)	Poor/none
PC2	0.00 (NA, NA)	Poor/none
PC3	0.18 (0.01, 0.88)	Slight
PC4	0.25 (0.02, 0.83)	Fair
PC5	0.26 (0.03, 0.83)	Fair
PC6	0.21 (0.01, 0.86)	Fair
PC7	0.00 (NA, NA)	Poor/none
PC8	0.03 (0.00, 1.00)	Slight
PC9	0.00 (NA, NA)	Poor/none
PC10	0.00 (NA, NA)	Poor/none
PC11	0.40 (0.09, 0.82)	Fair
PC12	0.13 (0.00, 0.94)	Slight
PC13	0.25 (0.02, 0.85)	Fair
PC14	0.00 (NA, NA)	Poor/none
MK	0.00 (NA, NA)	Poor/none
PROF1	0.00 (NA, NA)	Poor/none
PROF2	0.00 (NA, NA)	Poor/none
ICS1	0.52 (0.16, 0.86)	Moderate
ICS2	0.41 (0.09, 0.82)	Moderate
PBLI	0.42 (0.10, 0.82)	Moderate
SBP1	0.70 (0.38, 0.90)	Substantial
SBP2	0.28 (0.03, 0.82)	Fair
SBP3	0.42 (0.10, 0.82)	Moderate

FMA, full milestone assessment; SMA, self-assessment; ICC, intraclass correlation coefficient; NA, no agreement.

Poor or no agreement was observed for PC1, PC2, PC7, PC9, PC10, PC14, MK, PROF1 and PROF2.

Time requirement for milestone assessment

The average time to complete the AMA was 5.3 ± 2.1 min while the average time for the FMA was 80.9 ± 20.6 min ($p < 0.001$).

Discussion

While an appropriate 'gold standard' for resident milestone assessment is unclear, intuitively an assessment

performed using all available data would appear to be the most valid and reliable. Given the unique Emergency Department environment where faculty and residents work together in close proximity 24/7, it is possible that the global opinion of an experienced, educational faculty member (represented by AMA in this study) may be similar to the gold-standard assessment (as represented by FMA). AMA would be desirable given the anticipated time saved – faculty could devote more time to teaching rather than to milestone scoring and reporting. This study did demonstrate that significantly less time was needed for AMA than FMA. Eight of 23 subcompetencies had substantial or better agreement between AMA and FMA; overall 17 of the 23 had moderate agreement or better (Table 3).

AMA reflect actual observation and knowledge of resident performance by experienced educational faculty, and it was anticipated that all AMA for a given resident would be similar. However, when multiple faculty AMAs for individual residents were assessed, substantial agreement was only established for PC1, PC2 and PROF2 (Table 4). Clinical reasons for this lack of consistency include differences in the amount of clinical time faculty and residents worked together, variation in the clinical cases observed, halo effect from an exceptional clinical case or shift, negative effect from a poor patient encounter, and variations in faculty expectations – certain faculty may be more stringent and some may be more lenient while evaluating the same observation (the hawk-dove problem) [6,7]. Nonclinical factors, such as the impact of personality clashes and resident performance variations based on external, nonclinical stressors, also may have played a role. These issues as well as variations in the documentation of events into databases available for FMA completion may have added to the lack of correlation between AMA and FMA. It is also possible that each faculty interpreted the milestone language differently, suggesting that the milestones are not very objective. For example, two faculty members observing the same resident action may have come to different conclusions as to whether that action represented 'orders appropriate diagnostic studies' (PC3, Level 2, milestone 1) (Supplemental Table 1).

The use of self-assessment in resident development is itself an EM milestone, and therefore, it may have some value in resident assessment [3]. When residents performed a self-assessment by completing their own milestone document, their comparisons with FMA were worse than those found for AMA – Only 5 of the 23 subcompetencies (ICS1, ICS2, PBLI, SBP1 and SBP3) showed moderate or greater agreement between SMA and FMA (Table 5). Only one subcompetency (SBP1) showed strong agreement between SMA and FMA, and interestingly this same subcompetency showed only slight agreement when comparing AMA to FMA. Of the 8 subcompetencies with substantial or better agreement between AMA

and FMA, none showed a similar level of agreement between SMA and FMA.

These SMA results reaffirm prior research which show that resident and physician self-assessment of strengths and weaknesses do not correlate with proficiency [8]. Specifically for our study, factors that may have had additional impact were residents' unfamiliarity with the milestones framework and scoring system, resident use of peer comparison rather than an absolute reference, such as the Model of Clinical Practice of EM, and resident physicians' lack of meta-cognition to understand the skills required to perform a subcompetency at the staff physician level [9].

Thus, despite the time saving value of using AMA and SMA, the results demonstrate that at best only 8 of 23 subcompetencies seem amenable to AMA. Furthermore, only one of these eight, PROF2 meets the substantial agreement level for both inter-rater reliability and comparison to the gold standard (FMA). Therefore, the opinion of an experienced educational faculty may serve as a starting point for the assessment of this single subcompetency, but overall, we did not demonstrate a less resource-intensive method to generate valid milestone-based assessments and subcompetency scoring other than the FMA.

Our recommendation is to continue performing an objective review of all available resident assessment data for reporting to the ACGME semiannually. To help improve this process, the following will be needed: creation of better assessment tools for milestones, improving ease of access to data by CCC members and development of consensus within the CCC as to the meaning of each milestone and the metric by which it is considered achieved. Revision of the milestones language may also improve the validity of the assessment. In addition, institutional support is needed for the added faculty demands created by the CCC and milestone reporting. Faculty time commitment is substantial. This study calculated a mean time of 81 min to complete each FMA and 4 h of group meeting time to review one class of residents for one reporting cycle. With our complement of 42 residents and 7 faculty CCC members, this would equate to a total 23.5 h of faculty time required for the CCC each month. Since completing this study, our program has started developing novel milestone-based assessment tools and we are studying the resource requirements of our current CCC [10].

Limitations

Our study was limited to a single center and the sample size is small, this limiting our ability to detect certain differences. Legacy assessments (Table 1) are not standardized across all residency programs, therefore our experience utilizing these to implement the NAS may

not be generalizable to other programs. In addition, residents may have demonstrated certain milestone behaviors during clinical shifts that were not recorded in legacy assessments and subsequently were not reflected in the FMA, though the faculty evaluator would have integrated this information into the AMA.

FMA was considered the gold standard assessment in this study, but this has not been validated. While faculty were creating FMA, it is possible that their personal opinion of resident performance may have determined how they adjudicated certain milestones. Blinding was not possible in this study due to the structure of our department where all faculty work with all residents and provide written feedback on resident performance. In addition, only one FMA was completed per resident so inter-rater agreement of FMAs was not assessed. Another significant limitation is the fact that the EM milestones themselves have not been fully validated [11,12].

Our CCC experience described here predates the change in assessment processes necessary for implementation of the milestones. It is important to note multiple assessment methods are important to provide feedback to residents, even if the assessment does not directly link to any specific milestone. As EM programs develop assessment tools focused around the milestones, faculty become more facile with milestone assessment, rules for adjudication of individual milestones are defined, and milestones are adjusted to better reflect the needs of EM, we expect that the process will become easier. Specifically, the time commitment required to report data to the ACGME may decrease. With continued use of the milestones, faculty may become increasingly proficient in noticing relevant behaviors when assessing resident performance; this could potentially improve the agreement between AMA and FMA. Our goal from this study was to measure the resources required to move to the NAS from existing systems; this data may be very helpful to emergency departments looking to establish new ACGME-accredited residencies [13].

Conclusions

Using personal opinions of expert educational faculty to evaluate EM residents on the milestones takes significantly less time than using an objective summary of all assessment data. However, neither faculty opinions nor resident self-assessments agree with the objective summary on most subcompetencies. A total of 23.5 h of faculty time are required per month to fulfill the current ACGME requirement for semiannual reporting of EM milestones for a residency consisting of 42 trainees. Our study establishes that programs may need to develop new, milestone-based objective assessments to implement an effective and efficient CCC that generates valid resident assessments. Further study is required to assess

CCC resource requirements after development of novel assessment tools.

Acknowledgments

The authors thank Dr Julia Hays for her contributions as a member of the initial CCC. Ms. Stephanie Stebens from Sladen Library provided valuable assistance in proofreading the manuscript and preparing it for submission. Dr. James Yang and Dr. Lonni Schultz from the Department of Public Health Sciences provided statistical support for our study. Finally, this study would not have been possible without the resident physicians of the Henry Ford Hospital Emergency Medicine Residency Program who provide outstanding patient care and inspire the authors to become better teachers.

Financial Support

None

Disclosure of Interest

The authors report no conflict of interest. [NG] has created a software application to facilitate milestone data collection, interpretation and reporting; this product is not being commercialized.

ORCID

Nikhil Goyal  <http://orcid.org/0000-0003-1077-8003>

Jason Folt  <http://orcid.org/0000-0002-0230-6703>

Bradley Jaskulka  <http://orcid.org/0000-0001-8066-893X>

References

- [1] Nasca TJ, Philibert I, Brigham T, et al. The next GME accreditation system—rationale and benefits. *The New England Journal of Medicine*. 2012 Mar 15;366(11):1051–1056. PubMed PMID: 22356262.
- [2] Accreditation Council for Graduate Medical Education. Frequently asked questions: Milestones [Internet]. 2015 [updated 09 2015; cited 2018 Jan 24]. Available from: <http://www.acgme.org/Portals/0/MilestonesFAQ.pdf?ver=2015-11-06-115640-040>
- [3] Accreditation Council for Graduate Medical Education, American Board of Emergency Medicine. The Emergency Medicine Milestone Project [Internet]. 2018 [updated July 2015; cited 2018 Jan 24]. Available from: <https://www.acgme.org/acgmeweb/Portals/0/PDFs/Milestones/EmergencyMedicineMilestones.pdf>
- [4] Dreyfus SE, Dreyfus HL. A Five-Stage Model of the Mental Activities Involved in Directed Skill Acquisition. United States Air Force, Report number ORC 80-2. Available from: <http://www.dtic.mil/dtic/tr/fulltext/u2/a084551.pdf>. 1980.
- [5] Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977 Mar;33(1):159–174. PubMed PMID: 843571.
- [6] McManus IC, Thompson M, Mollon J. Assessment of examiner leniency and stringency ('hawk-dove effect') in the MRCP(UK) clinical examination (PACES) using multi-facet Rasch modelling. *BMC Med Educ*. 2006;6:42. PubMed PMID: 16919156; PubMed Central PMCID: PMC1569374.
- [7] Fleming PR, Manderson WG, Matthews MB, et al. Evolution of an examination: M.R.C.P. (U.K.). *Br Med J*. 1974 Apr 13;2(5910):99–107. PubMed PMID: 4596404; PubMed Central PMCID: PMC1610728.
- [8] Davis DA, Mazmanian PE, Fordis M, et al. Accuracy of physician self-assessment compared with observed measures of competence: a systematic review. *JAMA*. 2006 Sep 6;296(9):1094–1102. PubMed PMID: 16954489.
- [9] Counselman FL, Borenstein MA, Chisholm CD, et al. The 2013 Model of the Clinical Practice of Emergency Medicine. *Academic Emergency Medicine: Official Journal of the Society for Academic Emergency Medicine*. 2014 May;21(5):574–598. PubMed PMID: 24842511.
- [10] Goyal N, Vallee PA, Folt J, et al. WIRED for milestones. *J Grad Med Educ*. 2016 Jul;8(3):445–446. PubMed PMID: 27413456; PubMed Central PMCID: PMC4936871.
- [11] Peck TC, Dubosh N, Rosen C, et al. Practicing emergency physicians report performing well on most emergency medicine milestones. *The Journal of Emergency Medicine*. 2014 Oct;47(4):432–440. PubMed PMID: 25012279.
- [12] Korte RC, Beeson MS, Russ CM, et al. The emergency medicine milestones: a validation study. *Acad Emerg Med*. 2013 Jul;20(7):730–735. PubMed PMID: 23859587.
- [13] Accreditation Council for Graduate Medical Education, American Osteopathic Association, American Association of Colleges of Osteopathic Medicine. Allopathic and osteopathic medical communities commit to a single graduate medical education accreditation system [Internet]. Accreditation Council for Graduate Medical Education; 2014 [cited 2018 Jan 24]. Available from: <https://www.acgme.org/acgmeweb/portals/0/PDFs/Nasca-Community/SingleAccreditationRelease2-26.pdf>
- [14] Slezak M, Goyal N, Baliga S, et al. REACH: a novel process to collate resident performance metrics [abstract]. *Acad Emerg Med*. 2014;21(5 Suppl 1):S172.
- [15] American Board of Emergency Medicine. In-training examination overview [Internet]. 2018 [cited 2018 Jan 24]. Available from: <https://www.abem.org/public/emergency-medicine-training/in-training-examination/in-training-examination-overview>