Nuance and Noise: Lessons Learned From Longitudinal Aggregated Assessment Data

Teresa M. Chan, MD, MHPE
Jonathan Sherbino, MD, MEd
Mathew Mercuri, PhD

ABSTRACT

Background Competency-based medical education requires frequent assessment to tailor learning experiences to the needs of trainees. In 2012, we implemented the McMaster Modular Assessment Program, which captures shift-based assessments of resident global performance.

Objective We described patterns (ie, trends and sources of variance) in aggregated workplace-based assessment data.

Methods Emergency medicine residents and faculty members from 3 Canadian university-affiliated, urban, tertiary care teaching hospitals participated in this study. During each shift, supervising physicians rated residents' performance using a behaviorally anchored scale that hinged on endorsements for progression. We used a multilevel regression model to examine the relationship between global rating scores and time, adjusting for data clustering by resident and rater.

Results We analyzed data from 23 second-year residents between July 2012 and June 2015, which yielded 1498 unique ratings (65 $\pm$ 18.5 per resident) from 82 raters. The model estimated an average score of 5.7 $\pm$ 0.6 at baseline, with an increase of 0.005 $\pm$ 0.01 for each additional assessment. There was significant variation among residents' starting score (y-intercept) and trajectory (slope).

Conclusions Our model suggests that residents begin at different points and progress at different rates. Meta-raters such as program directors and Clinical Competency Committee members should bear in mind that progression may take time and learning trajectories will be nuanced. Individuals involved in ratings should be aware of sources of noise in the system, including the raters themselves.

Introduction

Ensuring high-quality patient care in the face of increasing patient volumes[1] and duty hour restrictions[2,3] is increasingly challenging. These increases raise concerns about safe clinical care as residents transition to unsupervised practice. The ultimate goal of assessment in medical education is to determine when graduate trainees are ready for unsupervised practice.[4] Competency-based medical education is an outcomes-based approach to physician training.[5,6] Assessment is used to determine when residents achieve expected abilities, mapped to a staged progression of responsibility (ie, junior to senior).[6] Such programmatic assessment[7] uses multiple representative "biopsies" linked to a master blueprint, with staged criterion-based standards such as milestones.[8–13]

To date, the model for graduate medical education has been time based, where time spent on service served as a surrogate for the attainment of competence.[14] Locally, we have noted that learners tend to value individual pieces of feedback more than trends in global performance.[15] While it may be the case that individual observation encounters fit within an *assessment as learning* framework[16] and precipitate learning encounters between faculty teachers and trainees, this approach alone may not be sufficient for defensible advancement or remediation decisions.[17] If decision makers (such as program directors or Clinical Competency Committees [CCCs]) are to make defensible decisions using available data, it is incumbent on the designers of the assessment system to identify patterns of advanced and remedial performance within large assessment data sets and to identify how to combine data to determine this.[17] Understanding the nature of information acquired from longitudinal data sets is imperative for educators responsible for interpreting available trends and rendering decisions derived from programmatic assessment data systems.

This study describes the patterns arising from longitudinal aggregate assessments of performance toward global competence for intermediate-level residents (ie, postgraduate year 2 [PGY-2]).

Methods

The study environment consists of 3 publicly funded, university-affiliated teaching hospitals associated with 1 residency training program. Since 2012, this training program has used a workplace-based assessment system called the McMaster Modular

DOI: http://dx.doi.org/10.4300/JGME-D-17-00086.1

Assessment Program (McMAP).[18] Residents are asked to gather daily digital faculty assessments of their stage-specific global performance and specific sentinel clinical tasks relevant to the practice of emergency medicine. We have previously shown that the McMAP system has internal consistency[19] and is superior to traditional end-of-rotation reports.[18]

During PGY-1, residents complete a rotating off-service internship that includes a 2-block introductory rotation in emergency medicine, alongside multiple off-service rotations including general surgery, internal medicine, pediatrics, obstetrics and gynecology, orthopedics, and anesthesia. In PGY-2, residents complete ten 4-week blocks of emergency medicine, during which their performance is rated every shift using the McMAP system. This allows our program to examine the performance of our PGY-2 residents as they transition from highly heterogeneous off-service experiences into clinical rotations in emergency medicine.

In addition to a workplace-based assessment portfolio of specific emergency medicine task assessments, residents' daily global performance is rated using a global rating score (FIGURE). The global rating score is completed by supervising physicians using a behaviorally anchored, competency-based scale (the CanMEDS 2015 framework).[18,20–22] A multilevel regression model was developed to examine the relationship between the global rating score and time (ie, sequential shifts), adjusting for data clustering by resident and rater. This allowed us to attribute partition variance to the resident and the rater, while also modeling variation among residents with respect to learning trajectory and beginning point. The dependent variable was the global rating score (1 to 7) of resident performance for each shift. The independent variable was time (ie, when the shift took place chronologically). Both the y-intercept (or beginning point) and time were included as random factors in the model. The mean score for each consecutive 4-week period (ie, a single block) was calculated for each resident. Analyses were performed using Stata/SE version 13.1 (StataCorp LLC, College Station, TX).

The McMaster University/Hamilton Integrated Research Ethics Board granted this study an exemption.

Results

The study included 82 individual raters (57 faculty members and 25 senior [PGY-4 and PGY-5]) residents. Fourteen resident raters joined the faculty during the study period. The average number of years in practice postresidency was 6.4 ± 9.5.

What was known and gap
Clinical Competency Committees (CCCs) rely on work-based ratings of trainees to make decisions about competence and progress in the program.

What is new
Shift-based assessments of emergency medicine residents showed variation in their level of competence at the start of the second year and the rate in which they progressed.

Limitations
Single institution, single specialty study limits generalizability.

Bottom line
Differences among trainees and "noise" in ratings have implications for program directors and CCCs.

From July 2012 through June 2015, data were collected on 23 (of a total of 23, 100%) PGY-2 residents from 3 resident classes. This yielded 1498 unique ratings (65 ± 18.5 per resident; 18.3 ± 15.7 per rater). Data on the number of shifts assessed and mean global rating score (overall, first 4-week block, last block) for each resident are presented in TABLE 1.

Unadjusted Resident Performance Analytics

Not accounting for the effect of different raters, the mean global rating score at the beginning of the year was 5.3 ± 0.6. The average score increased for 19 of 23 residents between their first and last blocks (average mean increase of 0.32; TABLE 1). However, only 12 of 23 residents (52%) managed to achieve an average global rating score of more than 6.25 in the final block (the a priori criterion for progression to senior-resident status based on pilot data). This criterion had been defined by the program director and the CCC, and the global rating data informed competency committee proceedings and judgments.

Adjusted Resident Performance Analytics: A Proposed Model

The model estimated an average global rating score of 5.7 ± 0.6 at the start of PGY-2 (ie, y-intercept). This score was estimated to increase 0.005 ± 0.01 with each additional assessment (ie, slope). There was significant variation among residents with respect to the intercept and slope, suggesting that residents significantly differ in ability at the start of their first block and progress at different rates (ie, have a different slope and rate of achieving competence). The model showed an interaction between resident intercept and slope; as the intercept increased, the slope decreased, suggesting a ceiling effect for those with a high global rating score at the start of the year.

The analyses suggest significant variation within and between individual residents and individual

RATE THIS TASK | CIRCLE ONE THAT THAT BEST DESCRIBES LEVEL OF PROFICIENCY						
1 Needs Assistance	2	3	4	5	6	7 Ready for to be a Senior Resident
Any of the following apply to the resident: • Displays major areas of knowledge deficit (i.e. only displaying Beginner Level knowledge).* • Displays major weaknesses with functioning in the ED environment (culture, logistics, collaboration) • input, revision, intervention or attentive supervision from attending throughout shift. • Performs actions that place patients at risk.* • Has lapses in professional behaviour. * • Ineffectively or offensively communicates with patient(s) or colleague(s).* • Cannot begin to remedy knowledge gaps at the point of care. • Shows lack of insight into own limitations or knowledge gaps.*			Most of the following apply to the resident: • Integrates well within the ED environment (culture, logistics, collaboration) • Has appropriate intermediate-level knowledge of EM-evidence and basic science. • Independently and accurately examines, diagnoses and determines care plan for non-critically ill patient(s). • Performs basic procedures safely with minimal supervision. • Effectively communicates with patient and colleagues (e.g. forms effective working relationships) • Is consistently professional. • Develops a plan to begin remedying knowledge gaps, limitations, deficits in exposure.			The resident displays mostly ALL of the following: • Functioning proficiently and efficiently in the ED environment (culture, logistics, collaboration) • Displays thorough knowledge of EM-evidence and basic science, or is able to independently access this information in a timely fashion. • Able to independently and accurately examine, diagnose and determine care plan for most patients (including the critically ill). • Able to perform procedures safely with minimal supervision. • Communicates efficiently with patients and colleagues (displays empathy, and forms good rapport). • Role models exceptional professional behaviour.* • Skilled at reflective practice and insight into own limitations, knowledge gaps; Able to self-identify and plan for continued improvement.

FIGURE
The Intermediate McMAP Rating Scale
* Denotes a descriptor that would necessitate additional qualitative comments to explain the rating.

raters. The highest source of variance in the global rating score was between different residents, as denoted by the intercept. Once time and rater effects were accounted for, within-resident variation was still substantial (TABLE 2).

Discussion

The determination of competence requires the aggregation of many observations from multiple observers to make a judgment (ie, create a meta-rating). In this exemplar study, we demonstrated certain patterns in aggregated data that may be important for those using multiple data points derived from assessment programs.

After a common, time-based year of training (the internship year), individuals begin at different observed levels of competence. As expected, through frequent, criterion-based assessments of authentic performance, we observed that trainees progress at different rates. Second, we described a learning trajectory that allows systems designers to anticipate the number of shifts an "average" resident requires to transition from being an intermediate resident to a senior resident, thereby allowing educational administrators and designers to allocate resources and plan residents' rotations. Finally, we have seen confirmatory evidence that raters can introduce a fair degree of noise (ie, variance) into the system.

Previously, data used to assess performance during rotations were typically collected via retrospective surveys of single faculty members (ie, post hoc in-training assessments of performance over the entire rotation), without a systematic process to ensure direct observation of resident performance.[23,24] Systems like McMAP overcome this by contemporaneously gathering prospective data,[18] which may then be evaluated.

At the same time, large data sets introduce new problems. The program director and/or CCCs now must interpret data sets that contain hundreds of data points. Thus, a competency-based medical education decision maker is a *meta-rater*, combining data from multiple sources into a specific judgment about competence. In this discussion, we highlight key points that such meta-raters should consider when making global judgments.

Nuances of Individualized Baselines and Progression

The observed range of resident baseline global rating score (4.3 to 6) suggests that even after a full year of "common" training, our residents did not enter their second year of residency with the same level of competence. Traditional education models assume that all learners progress equally. End-of-year examinations and end-of-rotation assessments are

TABLE 1
Tabulation of Resident Assessments and Score Outcomes for Each Individual Resident

Resident	No. of Shifts Rated	Overall GRS		First Block GRS		Last Block GRS		Difference
		Mean	SD	Mean	SD	Mean	SD	First to Last Block
A	92	6.16	0.75	6.00	0.67	6.42	0.67	0.42
B	47	5.67	0.84	5.13	0.48	6.00	0.00	0.87
C	60	5.78	0.58	5.67	0.5	6.00	0.00	0.33
D	87	6.29	0.76	6.09	0.54	6.73	0.59	0.64
E	49	6.06	0.66	6.00	1	6.71	0.49	0.71
F	81	6.09	0.67	5.77	0.88	6.45	0.69	0.68
G	79	6.11	0.75	6.25	0.71	6.54	0.52	0.29
H	91	5.98	0.78	5.10	0.97	6.13	0.99	1.03
I	52	5.87	0.62	5.85	0.75	6.25	0.46	0.40
J	77	5.61	0.99	5.33	0.78	4.90	1.29	−0.43
K	63	5.96	0.65	6.08	0.49	5.83	0.41	−0.24
L	74	6.29	0.60	6.14	0.71	6.46	0.63	0.33
M	46	6.13	0.59	5.80	0.79	6.38	0.52	0.58
N	36	6.22	0.77	5.92	0.58	6.50	1.00	0.58
O	61	6.14	0.51	6.00	0.00	6.31	0.60	0.31
P	28	5.43	0.69	5.29	0.76	5.31	0.67	0.02
Q	40	5.16	0.96	4.64	0.92	5.25	0.71	0.61
R	71	6.23	0.57	6.14	0.69	6.29	0.76	0.15
S	90	6.11	0.71	5.93	0.83	5.50	0.55	−0.43
T	63	6.32	0.67	5.88	0.64	6.50	0.53	0.62
U	72	5.83	1.10	6.67	0.71	6.00	0.71	−0.67
V	59	6.10	0.69	6.11	0.6	6.17	0.41	0.06
W	80	5.76	0.70	5.20	0.63	5.60	0.52	0.40

Abbreviation: GRS, global rating score.

presumed to identify trainees who are not advancing along a standard measure of progression. This leaves the educator with only the option of holding the resident back a year or advancing him or her with the hope that the resident can catch up. Differential learning trajectories for individual trainees suggest that there is no standard number of shifts at which individuals achieve the threshold score required for advancement. Such modeling may help educators anticipate the need for additional clinical exposures with relevant educational interventions before the end of PGY-2.

Over multiple years of training, modest differences can become substantial with respect to global competence. Gradual trajectories suggest educators have time to act. If analyzed correctly, careful attention to learning trajectories may allow educators to intervene earlier in the learning process, initiating

TABLE 2
Model for Intermediate Resident Progression Within McMaster Modular Assessment Program

Clusters	Variance	95% CI	SE	Significant
Rater	0.127	0.079 to 0.204	0.03059	Yes
Within resident	0.2615	0.2352 to 0.2908	0.014152	Yes
Fixed Effects	**Coefficient**	**95% CI**	**SE**	***P* Value**
Time (slope)	0.0053	0.0036 to 0.0070	0.00086	< .001
Intercept	5.71	5.58 to 5.84	0.06647	< .001
Random Effects	**Variance**	**95% CI**	**SE**	**Significant**
Time	0.0001	0.00006 to 0.00018	0.000028	Yes
Intercept	0.3913	0.2648 to 0.5782	0.077947	Yes
Time x-intercept	−0.0048412	−0.0077 to −0.0020	0.001453	Yes

Abbreviation: CI, confidence interval.

individualized learning plans with small changes and attention to neglected areas, rather than drastic remediation plans when significant gaps are identified late in residency. Residents who begin the year at a higher score and then trend downward may warrant closer observation and feedback, and they may be provided with added challenges to create desirable amounts of difficulty.[25,26] Residents who excel may similarly be identified by these trends, permitting earlier progression toward unsupervised practice.

Noise in the System: Raters and Other Sources of Noise

Curiously, our observational data suggested that time is only a minor contributor to the score variance. This may suggest that competency-based advancement is more appropriate than automatic time-based progression. The largest sources of variance were due to individual differences between and within residents as well as the effect of raters on the system.

The variance within a resident from shift to shift is to be expected, since context and performance will vary from day to day. Raters, however, present a particular challenge to decision makers. Our longitudinal, pragmatic data set demonstrates significant rater variance, consistent with experimental studies on rater cognition and variance.[27–30] Despite our attempts to create a shared mental model via a behaviorally anchored scale, we saw evidence of interrater variability, which is consistent with previous literature.[31] Furthermore, the range of the number of assessments gathered by each resident may also pose a problem.[32]

This study has limitations. It was based in a single program and specialty: local culture and context may limit the generalizability of our findings. Moreover, the interaction between intercept and slope suggests regression to the mean for some residents' ratings over the course of the training year. Our data set is not large enough to make robust conclusions about facets that contribute to the variance in our model. Our forms may suffer from problems shared with other CanMEDS-based assessment forms, including impressions of performance from one role spilling over to affect another.[33] Our global scale was designed to combat this phenomenon by asking raters one integrated question rather than multiple questions, which has been associated with rater variance.[34] As the amount of data increases, new and novel techniques for both visualizing and analyzing data will need to be attempted. Opportunities for using machine learning computer algorithms may further enhance data visualization for decision makers.[35]

Conclusion

Aggregated ratings can show the tailored progression of learner competence and document achievement of competence. In our study, emergency medicine PGY-2 residents did not enter their second year with the same assessed abilities, and their progression toward competence over the year varied. Some of these differences could be attributed to rater variability, which did produce some noise into the system. Other nuances and trends in these data can inform rotation planning and anticipate needs for remediation or advancement.

References

1. Fraser SW, Greenhalgh T. Complexity science: coping with complexity: educating for capability. *BMJ Clin Res Ed*. 2001;323(7316):799–803.
2. Ahmed N, Devitt KS, Keshet I, et al. A systematic review of the effects of resident duty hour restrictions in surgery: impact on resident wellness, training, and patient outcomes. *Ann Surg*. 2014;259(6):1041–1053.
3. Fung CH, Lim Y, Mattke S, et al. Systematic review: the evidence that publishing patient care performance data improves quality of care. *Ann Intern Med*. 2008;148(2):111–123.
4. Jones MD Jr, Rosenberg AA, Gilhooly JT, et al. Competencies, outcomes, and controversy—linking professional activities to competencies to improve resident education and practice. *Acad Med*. 2011;86(2):161–165.
5. Frank JR, Snell LS, ten Cate O, et al. Competency-based medical education: theory to practice. *Med Teach*. 2010;32(8):638–645.
6. Iobst WF, Sherbino J, ten Cate O, et al. Competency-based medical education in postgraduate medical education. *Med Teach*. 2010;32(8):651–656.
7. Schuwirth LWT, Van der Vleuten CPM. Programmatic assessment: from assessment of learning to assessment for learning. *Med Teach*. 2011;33(6):478–485.
8. van der Vleuten CPM, Schuwirth LWT, Driessen EW, et al. A model for programmatic assessment fit for purpose. *Med Teach*. 2012;34(3):205–214.
9. Van Der Vleuten CPM, Schuwirth LWT, Driessen EW, et al. Twelve tips for programmatic assessment. *Med Teach*. 2015;37(7):641–646.
10. Dijkstra J, Van der Vleuten CPM, Schuwirth LWT. A new framework for designing programmes of assessment. *Adv Heal Sci Educ*. 2010;15(3):379–393.
11. Korte RC, Beeson MS, Russ CM, et al. The emergency medicine milestones: a validation study. *Acad Emerg Med*. 2013;20(7):730–735.
12. Nabors C, Peterson SJ, Forman L, et al. Operationalizing the internal medicine milestones—an early status report. *J Grad Med Educ*. 2013;5(1):130–137.

13. Beeson MS, Carter WA, Christopher TA, et al. Emergency medicine milestones. *J Grad Med Educ*. 2013;5(1 suppl 1):5–13.
14. Snell LS, Frank JR. Competencies, the tea bag model, and the end of time. *Med Teach*. 2010;32(8):629–630.
15. Li S, Sherbino J, Chan TM. McMaster Modular Assessment Program (McMAP) through the years: residents' experience with an evolving feedback culture over a 3-year period. *AEM Educ Train*. 2017;1(1):5–14.
16. van der Vleuten C, Sluiijsmans D, Joosten-ten Brinke D. Chapter 28: competence assessment as learner support in education. In: Mulder M, ed. *Competence-Based Vocational and Professional Education: Bridging the Worlds of Work and Education*. Dordrecht, The Netherlands: Springer; 2017:607–630.
17. Hays RB, Hamlin G, Crane L, et al. Twelve tips for increasing the defensibility of assessment decisions. *Med Teach*. 2016;37(5):433–436.
18. Chan T, Sherbino J. The McMaster Modular Assessment Program (McMAP). *Acad Med*. 2015;90(7):900–905.
19. Sebok-Syer SS, Klinger DA, Sherbino J, et al. Mixed messages or miscommunication? Investigating the relationship between assessors? Workplace-based assessment scores and written comments [published online ahead of print May 30, 2017]. *Acad Med*. doi:10.1097/ACM.0000000000001743.
20. Chan TM, Sherbino J, eds. *McMaster Modular Assessment Program: Junior Edition*. San Francisco, CA: Academic Life in Emergency Medicine; 2015.
21. Chan TM, Sherbino J, eds. *McMaster Modular Assessment Program: Intermediate Edition*. San Francisco, CA: Academic Life in Emergency Medicine; 2015.
22. Chan TM, Sherbino J, eds. *McMaster Modular Assessment Program: Senior Edition*. San Francisco, CA: Academic Life in Emergency Medicine; 2015.
23. Epstein RM. Assessment in medical education. *N Engl J Med*. 2007;356(4):387–396.
24. Pulito AR, Donnelly MB, Plymale M, et al. What do faculty observe of medical students' clinical performance? *Teach Learn Med*. 2006;18(2):99–104.
25. Brown PC, Roediger HL, McDaniel MA. *Make It Stick*. Cambridge, MA: Harvard University Press; 2014.
26. Bjork R, Bjork E. A new theory of disuse and an old theory of stimulus fluctuation. In: Healy A, Kosslyn S, Shiffrin RM, eds. *From Learning Processes to Cognitive Processes: Essays in Honor of William K. Estes*. Hillsdal, NJ: Erlbaum; 1992:35–67.
27. Govaerts MJB, Van Der Vleuten CPM, Schuwirth LWT, et al. Broadening perspectives on clinical performance assessment: rethinking the nature of in-training assessment. *Adv Heal Sci Educ*. 2007;12(2):239–260.
28. Kogan JR, Conforti L, Bernabeo E, et al. Opening the black box of clinical skills assessment via observation: a conceptual model. *Med Educ*. 2011;45(10):1048–1060.
29. Gingerich A, Kogan J, Yeates P, et al. Seeing the "black box" differently: assessor cognition from three research perspectives. *Med Educ*. 2014;48(11):1055–1068.
30. Govaerts MJB, Schuwirth LWT, van der Vleuten CPM, et al. Workplace-based assessment: effects of rater expertise. *Adv Heal Sci Educ*. 2011;16(2):151–165.
31. Sterkenburg A, Barach P, Kalkman C, et al. When do supervising physicians decide to entrust residents with unsupervised tasks? *Acad Med*. 2010;85(9):1408–1417.
32. McConnell M, Sherbino J, Chan TM. Mind the gap: the prospects of missing data. *J Grad Med Educ*. 2016;8(5):708–712.
33. Kassam A, Donnon T, Rigby I. Validity and reliability of an in-training evaluation report to measure the CanMEDS roles in emergency medicine residents. *CJEM*. 2014;16(2):144–150.
34. Sherbino J, Kulasegaram K, Worster A, et al. The reliability of encounter cards to assess the CanMEDS roles. *Adv Heal Sci Educ*. 2013;18(5):987–996.
35. Ariaeinejad A, Samavi R, Chan T, et al. A performance predictive model for emergency medicine residents. In: Proceedings from the 27th Annual International Conference on Computer Science and Software Engineering; Toronto, ON, Canada; 2017.

Teresa M. Chan, MD, MHPE, is Assistant Professor, Division of Emergency Medicine, Department of Medicine, McMaster University, Hamilton, Ontario, Canada; **Jonathan Sherbino, MD, MEd,** is Associate Professor, Division of Emergency Medicine, Department of Medicine, Assistant Dean of Health Professions Education Research, and head of the MERIT (McMaster Education Research: Innovation and Theory) unit, McMaster University; and **Mathew Mercuri, PhD,** is Assistant Professor, Division of Emergency Medicine, Department of Medicine, McMaster University; Senior Research Associate, African Centre for Epistemology and Philosophy of Science, Department of Philosophy, University of Johannesburg, Johannesburg, South Africa; and Doctoral Candidate, Institute for the History and Philosophy of Science and Technology, University of Toronto, Toronto, Ontario, Canada.

Funding: Dr. Chan holds a McMaster University Department of Medicine Internal Career Research Award for her work on this project. Drs. Chan and Sherbino have also previously received funding from the Royal College of Physicians and Surgeons of Canada for various unrelated projects.

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

Corresponding author: Teresa M. Chan, MD, MHPE, Hamilton General Hospital, McMaster Clinic, Room 255, 237 Barton Street E, Hamilton, ON L8L 2X2 Canada, 905.521.2100 ext 76207, teresa.chan@medportal.ca

Received February 2, 2017; revision received July 4, 2017; accepted August 22, 2017.