



## RESEARCH ARTICLE

## Impact on house staff evaluation scores when changing from a Dreyfus- to a Milestone-based evaluation model: one internal medicine residency program's findings

Karen A. Friedman<sup>1\*</sup>, Sandy Balwan<sup>1</sup>, Frank Cacace<sup>1</sup>, Kyle Katona<sup>1</sup>, Suzanne Sunday<sup>2</sup> and Saima Chaudhry<sup>1</sup>

<sup>1</sup>Department of Internal Medicine, Hofstra North Shore LIJ School of Medicine, Hempstead, New York, NY, USA; <sup>2</sup>Department of Psychiatry, Hofstra North Shore LIJ School of Medicine, Hempstead, New York, NY, USA

**Purpose:** As graduate medical education (GME) moves into the Next Accreditation System (NAS), programs must take a critical look at their current models of evaluation and assess how well they align with reporting outcomes. Our objective was to assess the impact on house staff evaluation scores when transitioning from a Dreyfus-based model of evaluation to a Milestone-based model of evaluation. Milestones are a key component of the NAS.

**Method:** We analyzed all end of rotation evaluations of house staff completed by faculty for academic years 2010–2011 (pre-Dreyfus model) and 2011–2012 (post-Milestone model) in one large university-based internal medicine residency training program. Main measures included change in PGY-level average score; slope, range, and separation of average scores across all six Accreditation Council for Graduate Medical Education (ACGME) competencies.

**Results:** Transitioning from a Dreyfus-based model to a Milestone-based model resulted in a larger separation in the scores between our three post-graduate year classes, a steeper progression of scores in the PGY-1 class, a wider use of the 5-point scale on our global end of rotation evaluation form, and a downward shift in the PGY-1 scores and an upward shift in the PGY-3 scores.

**Conclusions:** For faculty trained in both models of assessment, the Milestone-based model had greater discriminatory ability as evidenced by the larger separation in the scores for all the classes, in particular the PGY-1 class.

**Keywords:** *Medical Education-Graduate; Medical Education-assessment methods; Milestones; Next Accreditation System; ACGME core competencies*

\*Correspondence to: Karen A. Friedman, Department of Internal Medicine, 300 Community Drive, Department of Graduate Medical Education, 4th Floor, Manhasset, NY 11030, USA, Email: [kfriedma@nshs.edu](mailto:kfriedma@nshs.edu)

Received: 12 June 2014; Revised: 1 October 2014; Accepted: 29 October 2014; Published: 24 November 2014

Graduate medical education (GME) training programs across the country are already entrusted with cultivating and assessing the knowledge, skills, and attitudes their trainees must attain upon completion of residency via the six domains of clinical competency as outlined by the Accreditation Council for Graduate Medical Education (ACGME). In the Next Accreditation System (NAS), measurement and reporting of these outcomes will be done via educational milestones (1). Many undergraduate medical education programs have already adopted the ACGME competencies. It is expected that milestones will also reach the undergraduate level and 'contribute to a more seamless transition across

the medical-education continuum' (1). The initial milestones for entering residents will add performance-based vocabulary to conversations with medical schools about graduates' preparedness for supervised practice (2).

On a GME level, assessing resident progress in achieving these domains has been and continues to be a challenge. The spectrum of conceptual models and tools utilized by individual programs speaks for the elusive nature of assessment, especially with regard to competencies outside of medical knowledge (3, 4).

The Dreyfus Framework has been one popular conceptual model on which many evaluation systems have been based (5). It describes phases of achieving professional

expertise as novice, advanced beginner, competent, proficient, and expert. One negotiates each phase, moving from adhering to ‘context-free rules’ to developing ‘intuition’ in decision making and formulating plans. Some authors have called into question the application of the Dreyfus concept as an adequate way to understand complex implicit knowledge necessary in clinical thinking (6). Other developmental models have recently been developed, most notably the ‘Milestone model’ of resident assessment. These milestones were developed by an ACGME and American Board of Internal Medicine (ABIM) task force in 2007 (7).

Phrased in observable behavioral terms, the milestones operationalize the knowledge, skills, and attitudes a resident should demonstrate at different points in the 36 months of training across each competency (8). All training programs will eventually have to adopt the milestones in their evaluation tools to document trainee competence.

It is unclear what impact the transition to a Milestone-based model of assessment will have on the summative evaluations programs must report in the NAS. The aim of this study is to assess and report the distribution of resident performance ratings, across the 3 years of training for all core competencies, pre- and post-adoption of a Milestone model of assessment in one large internal medicine residency program.

We hypothesized that compared to a previous Dreyfus model of assessment, a Milestone model would add increased discriminative ability in assessing the six ACGME competencies for internal medicine residents in training.

### Methods

Our internal medicine training program is a large university-based program in New York with 146 trainees. House staff are evaluated at the end of every rotation by the attending they have worked with using an electronic global end of rotation form, among many other tools. The faculty in our Department of Medicine is diverse and includes a large component of subspecialty faculty, ambulatory faculty, and hospitalists.

In the academic year 2010–2011, our global rating form was based on a modified Dreyfus model of assessment and utilized a 5-point scale to evaluate trainees (1 = beginner, 2 = advanced beginner, 3 = approaching competency, 4 = competent, and 5 = advanced competent). Full-time day hospitalists were trained on this evaluation method with a 2-h faculty development workshop in June of 2009. The session included a lecture presentation defining the model, a video tutorial, and a break out session where the faculty practiced using the Dreyfus model. Sporadic and repeated reinforcement of key concepts in the Dreyfus scale and its use in trainee evaluation occurred throughout the

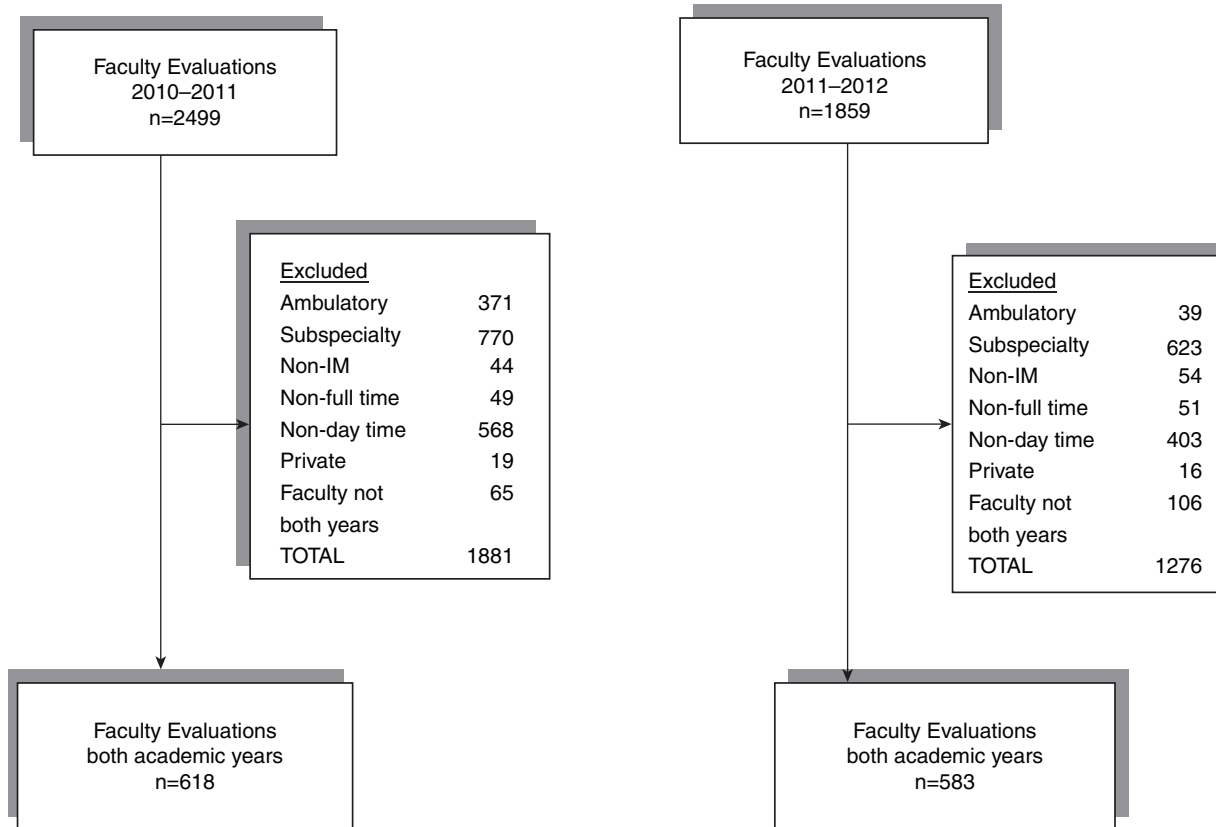


Fig. 1. Diagram showing the inclusion and exclusion criteria of faculty evaluations.

**Table 1.** Means (standard deviations) and significant pairwise comparisons from the ANOVA results for each competency score and total scores across the four evaluations (July–September, October–December, January–March, April–June) for PGY-1, PGY-2, and PGY-3 residents evaluated using the Dreyfus and Milestone models

Competency score	Evaluation	Model	PGY-1	PGY-2	PGY-3
Patient care	1 (n = 102, 25, 36)	Dreyfus	2.60 (0.86)	3.88 (0.78) <sup>a</sup>	4.39 (0.64) <sup>a</sup>
	1 (n = 123, 40, 37)	Milestone	1.76 (0.71) <sup>b</sup>	4.00 (0.55) <sup>a</sup>	4.92 (0.28) <sup>a,c</sup>
	2 (n = 100, 31, 37)	Dreyfus	3.14 (0.88) <sup>d</sup>	4.16 (0.69) <sup>a</sup>	4.35 (0.75) <sup>a</sup>
	2 (n = 95, 37, 20)	Milestone	2.41 (0.68) <sup>b,d</sup>	3.92 (0.43) <sup>a</sup>	4.95 (0.22) <sup>a,c</sup>
	3 (n = 90, 22, 23)	Dreyfus	3.39 (0.91) <sup>d</sup>	4.27 (0.46) <sup>a</sup>	4.52 (0.51) <sup>a</sup>
	3 (n = 83, 34, 15)	Milestone	2.82 (0.50) <sup>b,d,e</sup>	4.15 (0.44) <sup>a</sup>	4.80 (0.41) <sup>a</sup>
	4 (n = 103, 21, 27)	Dreyfus	3.54 (0.81) <sup>e,f</sup>	4.14 (0.57) <sup>a</sup>	4.26 (0.59) <sup>a</sup>
	4 (n = 59, 26, 14)	Milestone	3.24 (0.54) <sup>d,e,f</sup>	4.23 (0.43) <sup>a</sup>	5.00 (0.00) <sup>a</sup>
Medical knowledge	1 (n = 102, 25, 36)	Dreyfus	2.60 (0.85)	3.84 (0.75) <sup>a</sup>	4.25 (0.60) <sup>a</sup>
	1 (n = 123, 40, 37)	Milestone	1.84 (0.73) <sup>b</sup>	3.78 (0.62) <sup>a</sup>	4.76 (0.43) <sup>a,c</sup>
	2 (n = 100, 31, 37)	Dreyfus	3.10 (0.83) <sup>d</sup>	4.10 (0.70) <sup>a</sup>	4.22 (0.75) <sup>a</sup>
	2 (n = 95, 31, 20)	Milestone	2.45 (0.65) <sup>b,d</sup>	3.95 (0.33) <sup>a</sup>	4.90 (0.31) <sup>a,b,c</sup>
	3 (n = 90, 22, 23)	Dreyfus	3.31 (0.86)	4.18 (0.50) <sup>a</sup>	4.43 (0.51) <sup>a</sup>
	3 (n = 83, 34, 15)	Milestone	2.82 (0.52) <sup>b,d,e</sup>	4.09 (0.45) <sup>a</sup>	4.87 (0.35) <sup>a,c</sup>
	4 (n = 104, 21, 27)	Dreyfus	3.50 (0.72) <sup>e,f</sup>	4.10 (0.70) <sup>a</sup>	4.22 (0.64) <sup>a</sup>
	4 (n = 59, 26, 14)	Milestone	3.27 (0.49) <sup>d,e,f</sup>	4.15 (0.37) <sup>a</sup>	4.93 (0.27) <sup>a,c</sup>
Practice-based learning and improvement	1 (n = 102, 25, 36)	Dreyfus	2.71 (0.86)	4.00 (0.58) <sup>a</sup>	4.31 (0.58) <sup>a</sup>
	1 (n = 123, 40, 37)	Milestone	1.80 (0.65) <sup>b</sup>	3.90 (0.50) <sup>a</sup>	4.84 (0.37) <sup>a,c</sup>
	2 (n = 98, 30, 37)	Dreyfus	3.19 (0.82) <sup>d</sup>	4.17 (0.53) <sup>a</sup>	4.24 (0.76) <sup>a,b</sup>
	2 (n = 95, 31, 20)	Milestone	2.37 (0.65) <sup>b,d</sup>	3.92 (0.49) <sup>a</sup>	4.90 (0.31) <sup>a,c</sup>
	3 (n = 90, 22, 22)	Dreyfus	3.41 (0.97) <sup>d</sup>	4.18 (0.50) <sup>a</sup>	4.55 (0.51) <sup>a</sup>
	3 (n = 83, 34, 15)	Milestone	2.72 (0.53) <sup>b,d,f</sup>	4.09 (0.51) <sup>a</sup>	4.80 (0.41) <sup>a</sup>
	4 (n = 104, 21, 27)	Dreyfus	3.54 (0.82) <sup>d,e</sup>	4.14 (0.73) <sup>a</sup>	4.15 (0.60) <sup>a,b</sup>
	4 (n = 59, 26, 14)	Milestone	3.10 (0.36) <sup>b,d,e,f</sup>	4.19 (0.49) <sup>a</sup>	4.93 (0.27) <sup>a,c</sup>
Interpersonal and communication skills	1 (n = 102, 25, 36)	Dreyfus	3.25 (1.03)	4.28 (0.61) <sup>a</sup>	4.53 (0.56) <sup>a</sup>
	1 (n = 123, 40, 37)	Milestone	2.11 (0.82) <sup>b</sup>	4.20 (0.56) <sup>a</sup>	4.81 (0.40) <sup>a</sup>
	2 (n = 100, 30, 37)	Dreyfus	3.78 (0.95) <sup>d</sup>	4.47 (0.57) <sup>a</sup>	4.38 (0.76) <sup>a</sup>
	2 (n = 95, 37, 20)	Milestone	2.61 (0.79) <sup>b,d</sup>	3.95 (0.57) <sup>a</sup>	4.90 (0.31) <sup>a,c</sup>
	3 (n = 89, 22, 22)	Dreyfus	3.85 (1.08) <sup>d</sup>	4.32 (0.48)	4.55 (0.51) <sup>a</sup>
	3 (n = 83, 34, 15)	Milestone	3.01 (0.65) <sup>b,d,e</sup>	4.24 (0.50) <sup>a</sup>	4.87 (0.35) <sup>a</sup>
	4 (n = 103, 21, 27)	Dreyfus	3.91 (0.86) <sup>d</sup>	4.43 (0.60)	4.26 (0.59)
	4 (n = 59, 26, 14)	Milestone	3.49 (0.68) <sup>b,d,e,f</sup>	4.38 (0.50) <sup>a</sup>	4.93 (0.27) <sup>a</sup>
Systems-based practice	1 (n = 102, 25, 36)	Dreyfus	2.73 (0.94)	3.92 (0.57) <sup>a</sup>	4.36 (0.54) <sup>a,c</sup>
	1 (n = 123, 40, 37)	Milestone	1.93 (0.67) <sup>b</sup>	3.93 (0.53) <sup>a</sup>	4.65 (0.48) <sup>a,c</sup>
	2 (n = 100, 31, 37)	Dreyfus	3.18 (0.90) <sup>d</sup>	4.16 (0.58) <sup>a</sup>	4.27 (0.80) <sup>a</sup>
	2 (n = 95, 37, 20)	Milestone	2.47 (0.63) <sup>b,d</sup>	3.84 (0.44) <sup>a</sup>	4.80 (0.41) <sup>a,c</sup>
	3 (n = 89, 21, 23)	Dreyfus	3.44 (1.03) <sup>d</sup>	4.29 (0.46) <sup>a</sup>	4.57 (0.51) <sup>a</sup>
	3 (n = 83, 34, 15)	Milestone	2.75 (0.51) <sup>b,d</sup>	4.09 (0.51) <sup>a</sup>	4.73 (0.46) <sup>a</sup>
	4 (n = 104, 21, 27)	Dreyfus	3.57 (0.76) <sup>d,e</sup>	4.19 (0.75) <sup>a</sup>	4.22 (0.64) <sup>a</sup>
	4 (n = 59, 26, 14)	Milestone	3.15 (0.45) <sup>b,d,e,f</sup>	4.19 (0.40) <sup>a</sup>	4.93 (0.27) <sup>a,c</sup>
Professionalism	1 (n = 102, 25, 36)	Dreyfus	3.19 (0.54)	3.60 (0.50)	3.69 (0.47) <sup>a</sup>
	1 (n = 123, 40, 37)	Milestone	2.20 (0.90) <sup>b</sup>	4.18 (0.59) <sup>a,b</sup>	4.92 (0.28) <sup>a,b,c</sup>
	2 (n = 98, 31, 37)	Dreyfus	3.39 (0.59)	3.52 (0.57)	3.54 (0.60)
	2 (n = 95, 37, 20)	Milestone	2.75 (0.77) <sup>b,d</sup>	4.03 (0.60) <sup>a,b</sup>	5.00 (0.00) <sup>a,b,c</sup>
	3 (n = 90, 22, 23)	Dreyfus	3.36 (0.68)	3.55 (0.51)	3.70 (0.47)
	3 (n = 83, 34, 15)	Milestone	3.07 (0.66) <sup>d,e</sup>	4.26 (0.51) <sup>a,b</sup>	4.80 (0.41) <sup>a,b</sup>
	4 (n = 104, 21, 27)	Dreyfus	3.49 (0.52)	3.76 (0.44)	3.41 (0.64)
	4 (n = 59, 26, 14)	Milestone	3.49 (0.63) <sup>d,f</sup>	4.38 (0.50) <sup>a,b</sup>	4.93 (0.27) <sup>a,b</sup>
Overall	1 (n = 102, 25, 36)	Dreyfus	2.64 (0.91)	3.88 (0.73) <sup>a</sup>	4.42 (0.60) <sup>a,c</sup>
	1 (n = 123, 40, 37)	Milestone	1.88 (0.74) <sup>b</sup>	4.00 (0.51) <sup>a</sup>	4.86 (0.35) <sup>a,c</sup>

Table 1 (Continued)

Competency score	Evaluation	Model	PGY-1	PGY-2	PGY-3
	2 (n = 98, 30, 37)	Dreyfus	3.16 (0.84) <sup>d</sup>	4.19 (0.70) <sup>a</sup>	4.35 (0.75) <sup>a</sup>
	2 (n = 95, 37, 20)	Milestone	2.52 (0.63) <sup>b,d</sup>	3.92 (0.43) <sup>a</sup>	4.95 (0.22) <sup>a,c</sup>
	3 (n = 88, 20, 22)	Dreyfus	3.45 (0.96) <sup>d,e</sup>	4.25 (0.55) <sup>a</sup>	4.57 (0.51) <sup>a</sup>
	3 (n = 83, 34, 15)	Milestone	2.86 (0.57) <sup>b,d</sup>	4.18 (0.46) <sup>a</sup>	4.80 (0.41) <sup>a</sup>
	4 (n = 103, 21, 27)	Dreyfus	3.59 (0.77) <sup>b,d</sup>	4.14 (0.57) <sup>a</sup>	4.27 (0.60) <sup>a</sup>
	4 (n = 59, 26, 14)	Milestone	3.19 (0.47) <sup>b,d,e,f</sup>	4.23 (0.43) <sup>a</sup>	5.00 (0.00) <sup>a,c</sup>

<sup>a</sup>Within model, significantly different from PGY-1; <sup>b</sup>within PGY year and month of evaluation, significantly different from Dreyfus; <sup>c</sup>within model, significantly different from PGY-2; <sup>d</sup>within PGY year and model, significantly ( $p < 0.001$ ) different from evaluation 1; <sup>e</sup>from evaluation 2; <sup>f</sup>from evaluation 3.

academic year during our monthly clinical competency meetings, which the hospitalists attended.

In the academic year 2011–2012, we shifted from the Dreyfus model of trainee assessment and began using a Milestone model. Thus, our global rating form changed to incorporate the ACGME milestones, again using a 5-point scale (i.e., 1 = 0–3 months, 2 = 6 months, 3 = 12 months, 4 = 18–24 months, 5 = 30–36 months). Representative behaviors expected of trainees for each of the time frames was taken from published milestones literature and provided on the form along with a hyperlink listing all milestones as described by the ACGME-ABIM task force. Full-time day hospitalists were trained on this evaluation method with a 2-h faculty development workshop in September of 2011. The session included a lecture presentation defining the Milestone model and a break out session where faculty were able to read and interpret the milestones in groups. Sporadic and repeated reinforcement of key concepts in the Milestone scale and its use in trainee evaluation occurred throughout the academic year during our monthly clinical competency meetings, which the hospitalists attended.

In order to assess trainee progression along either developmental model (Dreyfus vs. Milestone), we analyzed all general medicine house staff evaluations completed by faculty from both academic years 2010–2011 and 2011–2012. Data pull of all faculty evaluations for both academic years included 4,358 evaluations. Exclusion criteria were applied as seen in Fig. 1. Ambulatory evaluations were excluded because the Milestone scale created for the ambulatory form was not 5 points and therefore not comparable to the prior Dreyfus model. Subspecialists, non-internal medicine faculty, non-full-time faculty, non-day time faculty, and faculty without evaluations in both academic years were excluded. When these exclusion criteria were applied, 1,201 evaluations were left for analysis (Fig. 1). The study was approved by the Hofstra North Shore-LIJ School of Medicine Institutional Review Board.

We analyzed all resident evaluations ( $N = 1,201$ ) that met our inclusion criteria. The primary outcomes were

the distribution of scores on the 5-point evaluation form in the Dreyfus and Milestone model calculated for all six competencies. Means and standard deviations were computed for the primary outcomes. A repeated-measures MANOVA (Tables 1 and 2) was conducted using a mixed model approach for each of the scores within PGY year, between models (Dreyfus vs. Milestones) and evaluations (four 3-month periods per year) entered as fixed effects and evaluator and resident entered as random effects. Interactions between the three fixed effects were also entered. Since all of our hypotheses reflected differences contained within the three-way interaction, the interaction was not removed from the model even if it was non-significant. Based on a priori hypotheses to compare PGYs at each 3-month time period for each of the two models, to compare the models for each of the PGYs at each 3-month time period, and to compare each of the 3-month time periods within a model for each PGY year, significance levels were Bonferroni corrected such that  $p = 0.001$  for all pairwise comparisons. We plotted the 3-month time periods during the 2010–2011 and 2011–2012 academic years along the x-axis and average scores for each PGY class for each competency along the y-axis.

## Results

Figures 2 and 3 display seven graphs depicting trainee skill acquisition (y-axis) over time (x-axis) in the Dreyfus and Milestone models as assessed by faculty meeting inclusion criteria. There is one graph for each competency. The average score for each PGY class at 3-month intervals across both evaluation models is plotted. These data are also depicted quantitatively in Table 1, with superscript letters indicating the following: within PGY year and model, significant differences from evaluation 1 (a), from evaluation 2 (b), and from evaluation 3 (c); within model, significant differences from PGY-1 (d) and from PGY-2 (e); and within PGY year and month of evaluation, significant difference from the Dreyfus model (f).

A larger separation of the scores between the PGY classes was seen in the Milestone model than in the Dreyfus model. There were no differences between PGY-2

Table 2. MANOVA results for each of the competency and total scores

Scores	Model	PGY year	Evaluation	Model*PGY	Model*Evaluation	PGY*Evaluation	Model*PGY*Evaluation
Patient care	(1,948) = 1.50, p = 0.22	(2,948) = 515.75, p < 0.0001	(3,948) = 22.31, p < 0.0001	(2,948) = 42.00, p < 0.0001	(3,948) = 1.88, p = 0.13	(6,948) = 18.03, p < 0.0001	(6,948) = 2.35, p = 0.03
Medical knowledge	(1,949) = 0.11, p = 0.74	(2,949) = 542.21, p < 0.0001	(3,949) = 28.35, p < 0.0001	(2,949) = 42.03, p < 0.0001	(3,949) = 2.17, p = 0.09	(6,949) = 13.49, p < 0.0001	(6,949) = 1.49, p = 0.18
Practice-based learning and improvement	(1,945) = 5.41, p = 0.02	(2,945) = 574.19, p < 0.0001	(3,945) = 19.76, p < 0.0001	(2,945) = 61.12, p < 0.0001	(3,945) = 3.19, p = 0.02	(6,945) = 13.97, p < 0.0001	(6,945) = 1.66, p = 0.13
Interpersonal and communication skills	(1,945) = 17.55, p < 0.0001	(2,945) = 290.32, p < 0.0001	(3,945) = 11.19, p < 0.0001	(2,945) = 51.75, p < 0.0001	(3,945) = 3.56, p = 0.01	(6,945) = 10.96, p < 0.0001	(6,945) = 1.81, p = 0.09
System-based practice	(1,947) = 9.11, p = 0.0026	(2,947) = 539.01, p < 0.0001	(3,947) = 22.30, p < 0.0001	(2,947) = 45.18, p < 0.0001	(3,947) = 2.95, p = 0.03	(6,947) = 10.30, p < 0.0001	(6,947) = 1.47, p = 0.19
Professionalism	(1,947) = 119.49, p < 0.0001	(2,947) = 246.54, p < 0.0001	(3,947) = 8.66, p < 0.0001	(2,947) = 135.35, p < 0.0001	(3,947) = 5.64, p = 0.0008	(6,947) = 11.42, p < 0.0001	(6,947) = 5.16, p < 0.0001
Overall	(1,942) = 1.68, p = 0.20	(2,942) = 501.56, p < 0.0001	(3,942) = 21.13, p < 0.0001	(2,942) = 40.37, p < 0.0001	(3,942) = 1.84, p = 0.14	(6,942) = 14.90, p < 0.0001	(6,942) = 1.91, p = 0.08

and PGY-3 classes with the Dreyfus model although separation from PGY-1 was seen for all competencies except professionalism. In the Milestone model, we were able to achieve a much greater separation for all of the competencies, between all classes.

A steady progressive improvement in scores can be seen for the PGY-1 class in both models for all competencies; however, the Milestone model had a steeper curve. The PGY-2s and the PGY-3s displayed no progression in either model for all competencies. Thus, changing to the Milestone model did not change the apparent progression of skills acquisition in either the PGY-2 or PGY-3 class.

The Milestone model resulted in a wider use of the range in the 5-point scale compared to the Dreyfus model. Table 1 shows a range of mean scores from 2.6 to 5 and 1.76 to 5 in the Dreyfus and Milestone models, respectively.

The Milestone model resulted in a shift of the average scores for the PGY-1 and PGY-3 class. Figures 2 and 3 show that in the Milestone model compared with the Dreyfus model, the PGY-1 class had lower scores, the PGY-3 class scored higher and the PGY-2 class stayed the same.

## Discussion

All residency programs must report Milestone achievement directly to the ACGME in the NAS. When transitioning to a Milestone-based model, programs should assess if the resultant system is an improvement over prior models of documenting trainees' developmental skills. One way to document this progression is to graph each resident's performance over time. One of the four possible conclusions can then be made: 1) the residents are progressing appropriately and the evaluation tool reflects that progression; 2) the residents are progressing appropriately, but the evaluation tool does not reflect the progression; 3) residents are not progressing appropriately and the evaluation tool captures the lack of progression; 4) residents are not progressing appropriately but the evaluation tool fails to capture the lack of progression. Programs can draw conclusions as to which of the four scenarios they feel represents their residents. If the evaluation systems are deemed to accurately represent skill acquisition, program directors can make decisions regarding remediation or advancement. If the evaluation systems are deemed invalid, they need to be changed. Analysis of our program's aggregated data indicated that in our program we were likely to be experiencing scenario number two.

Data analysis revealed four problems with how day time faculty utilized the global rating form in the Dreyfus model in our program. First, the scores for each class were not separate from each other and did not reflect the difference in clinical skills that faculty felt truly existed between classes. The PGY-2 and PGY-3 classes overlapped significantly in the Dreyfus model. In particular,



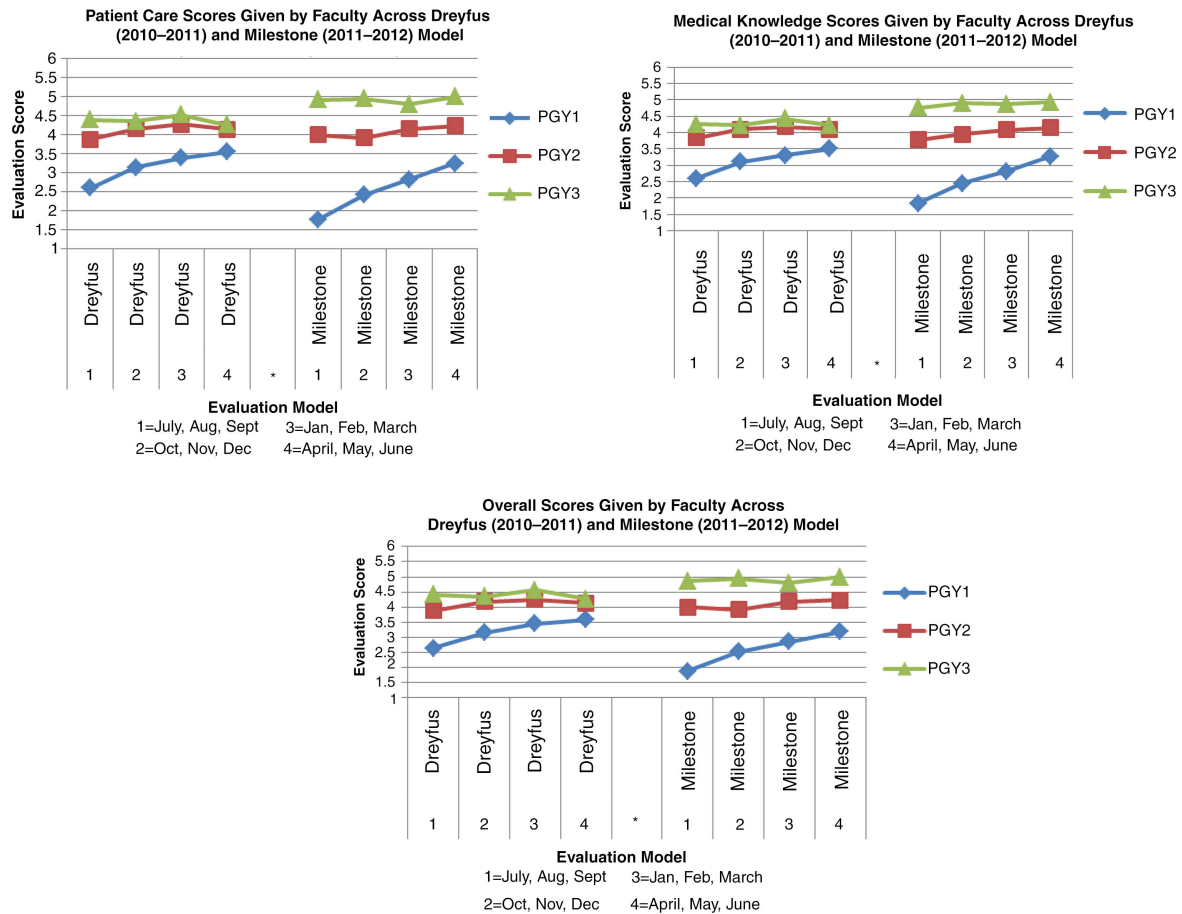


Fig. 2. House staff evaluation scores done by full-time day faculty across the competencies patient care and medical knowledge and overall score across both evaluation models.

all classes overlapped for professionalism. Assessing professionalism is hampered by varying definitions of the competency and thus is a concept that can be difficult to pin down and evaluate on single source tools (9). Shifting to the Milestone model, enhanced the discriminatory ability of the evaluator and allowed for a greater separation of the graphs as illustrated in Figs. 2 and 3. In our opinion, this scenario better mirrors the difference in clinical skills among our PGY classes and is most obvious in the competency of professionalism.

The second problem with the Dreyfus model was an upward progression in slope for the PGY-1 class, but virtually no slope in the PGY-2 and PGY-3 classes. The Milestone model created a steeper slope for the PGY-1 class. However, the PGY-2 and PGY-3 classes remained flat, albeit to a lesser degree (Figs. 2 and 3, Table 1).

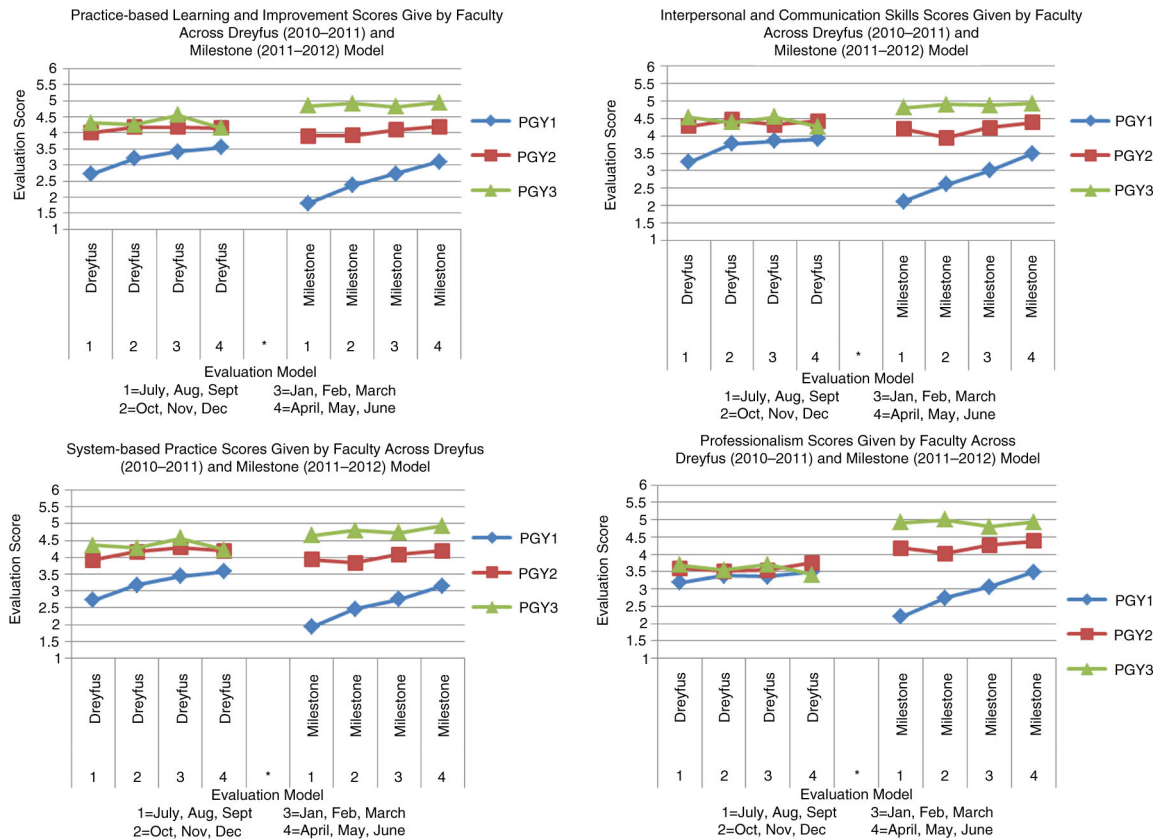
Third, analysis of the scores in the Dreyfus model showed that faculty was not using the full range of the 5-point scale. This is a well-recognized limitation of global rating forms (10). With the Milestone evaluation form, we were pleased to see a greater use of the full range of the scale. In particular, evaluators more frequently rated

interns with a score of 1 in the Milestone model (0–3 months) than in the Dreyfus model (i.e., beginner).

Finally, the average score in July for the PGY-1 class was significantly lower in the Milestone model compared with the Dreyfus model, as seen in Table 1. This result reflects the faculty’s increased willingness to use the lower end of the rating scale thereby counteracting the well-known ceiling effect, where ratings are primarily clustered in the highest category of the scale, as seen with many global rating forms (11).

It is worthwhile to note that the results we observed, regardless of model used, are a reflection of the faculty’s ability to observe and document resident behavior. It is well recognized that observation of trainees by faculty is at best suboptimal (12). Milestones are meant to help focus faculty’s observation skills.

To our knowledge, this is the first study to report on how aggregate data may change as programs transition to Milestone reporting in the NAS. By keeping a 5-point scale for both academic years, we were able to directly compare scores and graphs. We have provided two full academic years’ worth of aggregate data using a large



**Fig. 3.** House staff evaluation scores done by full-time day faculty across the competencies practice-based learning and improvement, interpersonal and communication skills, systems-based practice and professionalism across both evaluation models.

number of faculty evaluations. By critically analyzing our data, we were able to see the improvements milestones have made in distinguishing the PGY-2 and PGY-3 classes from each other over the Dreyfus model but also where improvement still needed to be made. We did not expect a lack of progression to be seen in these classes. This discovery will guide us to make changes once again. We were especially pleased to see that the Milestone model was able to separate scores even for the competency of professionalism, traditionally a difficult competency to evaluate on a rating form.

We acknowledge certain limitations of our study. It was conducted at one site with milestones only incorporated into our global end of rotation rating form. Our program has multiple other evaluation tools that were not included in this analysis. Global rating forms have limited intra-rater and interrater reliability and lack reproducibility (13, 14). They can be subject to central tendency errors, where the evaluators fail to use the entire rating scale, and the ‘halo’ effect, where evaluators’ biases about an initial impression by one component of a resident is extended to other aspects of the resident’s performance (8, 13). The last limitation of this study was our need for more robust faculty development. Faculty development was present in the pilot phases of both our models, but sparse through-

out the academic years. Faculty development is pivotal in operationalizing whatever evaluation system one chooses and has been shown to be most effective when done in a comprehensive and continuous fashion (13, 14).

Despite its limitation, the global end of rotation rating form, with its low cost, flexibility and ease of use, will likely be a popular evaluation tool for future use (8, 10). Using the tool successfully requires studying the distribution of scores achieved and reflecting on whether these scores truly reflect trainee competence. Our program aimed to achieve a greater distribution of scores using a 5-point scale with milestones as anchors. With the Milestone anchors there was a greater spread but even with this widened distribution of scores there was still room to improve. The ACGME, with its move toward reportable milestones for each trainee on a 9-point scale will undoubtedly also assess trainee progression over time graphically. We are pleased to have 2 years of experience doing the same. Our program has subsequently changed to an 8-point scale to give faculty more of a chance to differentiate behaviors between the PGY-2 and PGY-3 classes. Further change will continue as we adjust to the ACGME reportable milestones in the NAS. It will be important for the ACGME to similarly graph its nationwide aggregate trainee performance data overtime. Programs

will need to continuously self reflect upon their evaluation systems and adjust them until a 'true' accounting of trainees skill set is represented.

Graduate training programs are just beginning to learn how to incorporate milestones into their evaluation tools. Full incorporation will take time and will rely heavily on faculty buy-in and faculty development. Only through a thoughtful reflection of trainee evaluations and rating scales will programs be able to successfully document competency in the future generation of physicians.

### Conflict of interest and funding

The authors have not received any funding or benefits from industry or elsewhere to conduct this study.

### References

1. Nasca TJ, Philibert I, Brigham T, Flynn TC. The Next GME accreditation system-rationale and benefits. *N Engl J Med* 2012; 366: 1051–6.
2. Lypson ML, Frohna JG, Gruppen LD, Woolliscroft JO. Assessing residents' competencies at baseline: identifying gaps. *Acad Med* 2004; 79: 564–70.
3. Boateng B, Bass LD, Blaszek RT, Farrar HC. The development of a competency-based assessment rubric to measure resident milestones. *J Grad Med Educ* 2009; 1: 45–8.
4. Lurie SJ, Mooney CJ, Lyness JM. Measurement of the general competencies of the ACGME. A systematic review. *Acad Med* 2009; 84: 301–9.
5. Dreyfus SE, Dreyfus HL. A five stage model of the mental activities involved in directed skill acquisition. Air Force Office of Scientific Research under contract F49620-79-C-0063. Berkeley, CA: University of California; 1980.
6. Pena A. The Dreyfus model of clinical problem-solving skills acquisition: a critical perspective. *Med Educ Online* 2010; 15: 4846. doi: 10.3402/meo.v15i0.4846.
7. Green ML, Aagaard EM, Caverzagie K, Chick DA, Holmboe E, Kane G, et al. Charting the road to competence: developmental milestones for internal medicine residency training. *J Grad Med Educ* 2009; 1: 5–20.
8. Silber CB, Nasca TJ, Paster DL, Elger G, Robeson M, Velosdi JJ. Do global rating forms enable program directors to assess the ACGME competencies? *Acad Med* 2004; 79: 549–56.
9. Wilkenson TJ, Wade WB, Knock LD. A blueprint to assess professionalism: results of a systematic review. *Acad Med* 2009; 84: 551–8.
10. Gray JD. Global rating scales in residency education. *Acad Med* 1996; 71(Suppl 1): S55–63.
11. Vita S, Coplin H, Feiereisel K, Garten S, Mechaber AJ, Estrada C. Decreasing the ceiling effect in assessing meeting quality at an academic professional meeting. *Teach Learn Med* 2013; 25: 47–54.
12. Holmboe ES. Faculty and the observation of trainees' clinical skills: problems and opportunities. *Acad Med* 2004; 79: 16–22.
13. Wilkerson L, Irby D. Strategies for improving teaching practices: a comprehensive approach to faculty development. *Acad Med* 1998; 73: 387–96.
14. Steinert Y, Mann K, Centeno A, Dolmans D, Spencer J, Gelula M, et al. A systematic review of faculty development initiatives designed to improve teaching effectiveness in medical education: BEME Guide No.8. *Med Teach* 2006; 28: 497–526.