2018 Annual Educational Conference

SES084: Milestone Evaluations: Discovery of Some Thought Provoking Rating Trends

Marty Keil, MA, PhD Candidate, Stanford University
Nancy Piro, PhD, Sr. Program Manager/Education Specialist, Stanford Health Care
Ann Dohn, MA, DIO & GME Director, Stanford Health Care
Marty Keil, MA
Nancy Piro, PhD
Ann Dohn, MA

None of the above speakers have any conflicts of interest to report.
Agenda

- **Milestones**
  - Background and Goals
  - Milestone Project Initiation

- **Milestone Score Analysis**
  - Research Questions
  - Data Collection and Analysis
  - Summary of Findings
  - What does this mean for your institution?

- **Milestone Narrative (Comments) Analysis**
  - Research Questions
  - Data Collection and Analysis
  - Summary of Findings
  - What does this mean for your institution?

- **Current Directions – Feedback Model**

- **Impact of Resident Evaluations on Faculty Evaluations of Residents**

- **Interventions**
What is a Milestone?

- “A defined observable marker of an individual’s ability along a developmental continuum”
  - Used for planning and teaching
- “Educators will use Milestones to design educational activities and teach specific abilities.”
ACGME goals for Milestone Evaluation System

- Develop measures of resident performance across core competencies that also provide valuable feedback for residents to focus their practice

- High Quality Assessment Measures
  - Reliable
  - Valid

- High Quality Feedback System
  - Specific
  - Timely
  - Constructive
What are we experiencing?

- We now have Milestones in all ACGME Programs… but
  - Our largest number of ACGME citations has been for evaluations and feedback
  - Our internal GME House Staff survey demonstrates that our trainees are dissatisfied with their evaluation feedback
  - Our AGGME Survey:

```
<table>
<thead>
<tr>
<th>Evaluation</th>
<th>% Program Compliant</th>
<th>Program Mean</th>
<th>% National Compliant</th>
<th>National Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Able to access evaluations</td>
<td>95%</td>
<td>5.0</td>
<td>99%</td>
<td>4.9</td>
</tr>
<tr>
<td>Opportunity to evaluate faculty members</td>
<td>100%</td>
<td>5.0</td>
<td>99%</td>
<td>4.9</td>
</tr>
<tr>
<td>Satisfied that evaluations of faculty are confidential</td>
<td>85%</td>
<td>4.3</td>
<td>85%</td>
<td>4.3</td>
</tr>
<tr>
<td>Opportunity to evaluate program</td>
<td>95%</td>
<td>5.0</td>
<td>99%</td>
<td>4.9</td>
</tr>
<tr>
<td>Satisfied that evaluations of program are confidential</td>
<td>87%</td>
<td>4.4</td>
<td>87%</td>
<td>4.3</td>
</tr>
<tr>
<td>Satisfied that program uses evaluations to improve</td>
<td>76%</td>
<td>4.1</td>
<td>75%</td>
<td>4.1</td>
</tr>
<tr>
<td>Satisfied with feedback after assignments</td>
<td>71%</td>
<td>3.9</td>
<td>72%</td>
<td>4.0</td>
</tr>
</tbody>
</table>
```
So what to do???

- Milestone Project Initiation
  - 2016 Annual Institutional Review (AIR)
    - Committee members recommended pursuing a collaborative relationship with the Stanford Graduate School of Education (SGSOE) to leverage their expertise to
      - analyze Milestone Evaluations Trends
      - create improvement strategies
Milestone Project Initiation and Goals

- Initial Groundwork (GME and SGSOE)
  - Defined the problem from GME’s perspective
  - Brainstormed with the School of Education
    - Began working with Faculty and PhD Candidate focusing on improving graduate education

- Project Implementation
  - Stakeholder Determination (Anesthesia, Neurology, OB/GYN, Orthopedic Surgery, Psychiatry and Radiology)
  - Expanding to include other Institutions
Milestone Score Analysis – Research Questions

- How reliable are milestone score evaluations? What factors contribute to variance within the system, including attendings, residents, specialties, and institutions?

- Do attendings differentiate between the six core competencies when evaluating milestones?

- What are the relationships between milestone scores and time, as well as milestone scores and PGY, and how do these variables interact?

- Can we identify evaluators that consistently provide high and low quality feedback?

- What evaluator characteristics contribute to high quality feedback and assessment?
Milestone Score Analysis – Data Collection

- Data extracted from our Residency Management System (RMS) Online Portal and from the Ad Hoc Resident Reporting functionality concatenating several cells to delineate PGY Levels

- Data collected from 5 ACGME Core Specialties across 2 large academic medical centers (8 programs total)

- All data de-identified to protect Resident Confidentiality

- Datasets from 2013-2017
  - 4 year period beginning with the implementation of ACGME Milestones

- Data Filtering to include only
  - End-of-rotation milestone subcompetency scores from Faculty
  - Competencies evaluated in more than 30% of rotations

- Data analysis using R -> open-source code applicable to all files
Milestone Score Analysis – Data Collection and Overview

- Summary of Data

<table>
<thead>
<tr>
<th>Program</th>
<th>Institution</th>
<th>Evaluations</th>
<th>Evaluators</th>
<th>Residents</th>
<th>Competencies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neurology</td>
<td>B</td>
<td>504</td>
<td>55</td>
<td>39</td>
<td>28</td>
</tr>
<tr>
<td>Neurology</td>
<td>A</td>
<td>1067</td>
<td>90</td>
<td>36</td>
<td>29</td>
</tr>
<tr>
<td>OBGYN</td>
<td>B</td>
<td>1851</td>
<td>58</td>
<td>57</td>
<td>28</td>
</tr>
<tr>
<td>OBGYN</td>
<td>A</td>
<td>1376</td>
<td>66</td>
<td>31</td>
<td>27</td>
</tr>
<tr>
<td>Psychiatry</td>
<td>B</td>
<td>648</td>
<td>50</td>
<td>64</td>
<td>22</td>
</tr>
<tr>
<td>Psychiatry</td>
<td>A</td>
<td>935</td>
<td>97</td>
<td>65</td>
<td>21</td>
</tr>
<tr>
<td>Anesthesia</td>
<td>A</td>
<td>2406</td>
<td>66</td>
<td>116</td>
<td>25</td>
</tr>
<tr>
<td>Radiology</td>
<td>A</td>
<td>2965</td>
<td>79</td>
<td>67</td>
<td>10</td>
</tr>
</tbody>
</table>
Milestone Score Analysis-A Classic Progression Plot

- Scores increase each year in equally distributed levels
  - Signals consistent growth in resident performance and measurement validity

**ACGME National Report (2016)**

<table>
<thead>
<tr>
<th>Lev 1</th>
<th>Lev 2</th>
<th>Lev 3</th>
<th>Lev 4</th>
<th>Lev 5</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>


Mean Milestone Scores by PGY
Delving deeper we see highly variable scores for each resident across rotations.

**“B” Neurology**

- Milestone Score Analysis – Individual Plots
Milestone Score Analysis - Reliability

- Mixed Model Results demonstrate an evaluation system with low Intraclass Correlation Coefficients (ICC), suggesting poor reliability.

- Attendings explain significantly more of the variance than residents in six programs.

Milestone. Score ~ total time + Program. + Institution + (1|Evaluator.ID.) + (1|Evaluatee.ID.)

<table>
<thead>
<tr>
<th>Institution</th>
<th>Program</th>
<th>Percent Explained by Evaluator</th>
<th>ICC Score (Reliability)</th>
</tr>
</thead>
<tbody>
<tr>
<td>B Neurology</td>
<td></td>
<td>86.3%</td>
<td>0.086</td>
</tr>
<tr>
<td>A Neurology</td>
<td></td>
<td>59.6%</td>
<td>0.266</td>
</tr>
<tr>
<td>B OBGYN</td>
<td></td>
<td>88.9%</td>
<td>0.058</td>
</tr>
<tr>
<td>A OBGYN</td>
<td></td>
<td>93.7%</td>
<td>0.043</td>
</tr>
<tr>
<td>B Psychiatry</td>
<td></td>
<td>84.2%</td>
<td>0.117</td>
</tr>
<tr>
<td>A Psychiatry</td>
<td></td>
<td>92.9%</td>
<td>0.042</td>
</tr>
<tr>
<td>A Anesthesia</td>
<td></td>
<td>28.2%</td>
<td>0.481</td>
</tr>
<tr>
<td>A Radiology</td>
<td></td>
<td>43.0%</td>
<td>0.396</td>
</tr>
</tbody>
</table>
Milestone Score Analysis - Reliability Scaled

- Milestone scores scaled by average evaluator score
- Results demonstrate that the issue extends beyond incorrect internal reference point

<table>
<thead>
<tr>
<th>Institution</th>
<th>Program</th>
<th>Percent Explained by Evaluator</th>
<th>ICC Score (Reliability)</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>Neurology</td>
<td>43.0 %</td>
<td>0.156</td>
</tr>
<tr>
<td>A</td>
<td>Neurology</td>
<td>19.3%</td>
<td>0.32</td>
</tr>
<tr>
<td>B</td>
<td>OBGYN</td>
<td>90.4%</td>
<td>0.05</td>
</tr>
<tr>
<td>A</td>
<td>OBGYN</td>
<td>73.1%</td>
<td>0.075</td>
</tr>
<tr>
<td>B</td>
<td>Psychiatry</td>
<td>23.3%</td>
<td>0.234</td>
</tr>
<tr>
<td>A</td>
<td>Psychiatry</td>
<td>29.3%</td>
<td>0.087</td>
</tr>
<tr>
<td>A</td>
<td>Anesthesia</td>
<td>44.6%</td>
<td>0.351</td>
</tr>
<tr>
<td>A</td>
<td>Radiology</td>
<td>71.9%</td>
<td>0.21</td>
</tr>
</tbody>
</table>
Evaluator means dramatically different compared to resident means for any particular PGY level (PGY1 level most variable) – Institution “A”
Milestone Score Analysis – Split Regression Analysis

- Split mixed model into PGY by days (every 365 days)
- Determine Slope and Intercept for each model result

**Mixed Model:**

Milestone Score ~ Time + Program + Institution + (1|Evaluator.ID.) + (1|Evaluatee.ID.)

**Radiology**
Milestone Score Analysis – Lag Analysis

- For all programs besides psychiatry the y-intercept increases across years while the slope is relatively flat

- This suggests widespread scoring bias focused on PGY instead of time

<table>
<thead>
<tr>
<th>Institution</th>
<th>Program</th>
<th>PGY1 - Intercept</th>
<th>PGY1 - Slope</th>
<th>PGY2 - Intercept</th>
<th>PGY2 - Slope</th>
<th>PGY3 - Intercept</th>
<th>PGY3 - Slope</th>
<th>PGY14 - Intercept</th>
<th>PGY4 - Slope</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>Neurology</td>
<td>2.97</td>
<td>0.0007</td>
<td>3.83</td>
<td>0.0002</td>
<td>3.97</td>
<td>0.0003</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>A</td>
<td>Neurology</td>
<td>2.9425</td>
<td>0.0013</td>
<td>3.8284</td>
<td>0.0012</td>
<td>3.9381</td>
<td>0.0011</td>
<td>2.616</td>
<td>0.0012</td>
</tr>
<tr>
<td>B</td>
<td>OBGYN</td>
<td>1.6841</td>
<td>0.0007</td>
<td>2.2305</td>
<td>0.0006</td>
<td>2.0359</td>
<td>0.0014</td>
<td>3.1788</td>
<td>0.0007</td>
</tr>
<tr>
<td>A</td>
<td>OBGYN</td>
<td>2.0119</td>
<td>0.0014</td>
<td>2.4949</td>
<td>0.0014</td>
<td>2.9962</td>
<td>0.0009</td>
<td>3.2742</td>
<td>0.0007</td>
</tr>
<tr>
<td>B</td>
<td>Psychiatry</td>
<td>2.6922</td>
<td>-0.0002</td>
<td>3.1067</td>
<td>-0.0008</td>
<td>-0.0425</td>
<td>0.0039</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>A</td>
<td>Psychiatry</td>
<td>3.0038</td>
<td>0.0005</td>
<td>2.9641</td>
<td>0.0008</td>
<td>2.5556</td>
<td>0.0013</td>
<td>2.613</td>
<td>0.001</td>
</tr>
<tr>
<td>A</td>
<td>Anesthesia</td>
<td>2.525</td>
<td>0.0008</td>
<td>2.6424</td>
<td>0.0009</td>
<td>3.014</td>
<td>0.0008</td>
<td>2.6803</td>
<td>0.0009</td>
</tr>
<tr>
<td>A</td>
<td>Radiology</td>
<td>1.45</td>
<td>0.0001</td>
<td>2.4151</td>
<td>0.0003</td>
<td>3.1739</td>
<td>0.0002</td>
<td>3.8521</td>
<td>0.0003</td>
</tr>
</tbody>
</table>
Feedback should be specific

Factor analysis results in one factor explaining more than 90% of variance

Factor Analysis Results “B” Neurology

<table>
<thead>
<tr>
<th>Institution</th>
<th>Program</th>
<th>Variance Explained by 1st Factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>Neurology</td>
<td>.95</td>
</tr>
<tr>
<td>A</td>
<td>Neurology</td>
<td>.90</td>
</tr>
<tr>
<td>B</td>
<td>OBGYN</td>
<td>.97</td>
</tr>
<tr>
<td>A</td>
<td>OBGYN</td>
<td>.92</td>
</tr>
<tr>
<td>B</td>
<td>Psychiatry</td>
<td>.96</td>
</tr>
<tr>
<td>A</td>
<td>Psychiatry</td>
<td>.82</td>
</tr>
</tbody>
</table>
Milestone Score Analysis – Mean Variation per Evaluation

- Low standard deviation between subcompetency scores within an evaluation
- Overall 56.6% of evaluations had the same scores for all rated subcompetencies
Milestone Score Analysis – Timing Analysis

- Feedback should be timely
- Mean evaluation delay from end-of-rotation was more than 1 month - “A”

<table>
<thead>
<tr>
<th>Institution</th>
<th>Program</th>
<th>Mean Delay</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>Neurology</td>
<td>81.739</td>
</tr>
<tr>
<td>A</td>
<td>Neurology</td>
<td>41.191</td>
</tr>
<tr>
<td>B</td>
<td>OBGYN</td>
<td>37.124</td>
</tr>
<tr>
<td>A</td>
<td>OBGYN</td>
<td>36.477</td>
</tr>
<tr>
<td>B</td>
<td>Psychiatry</td>
<td>59.316</td>
</tr>
<tr>
<td>A</td>
<td>Psychiatry</td>
<td>66.541</td>
</tr>
<tr>
<td>A</td>
<td>Anesthesia</td>
<td>41.071</td>
</tr>
<tr>
<td>A</td>
<td>Radiology</td>
<td>29.693</td>
</tr>
</tbody>
</table>
Milestone Score Analysis – Evaluator Characteristics

- **Evaluator Quality Outcomes:**
  - Delay of evaluation
  - Competency variation within evaluation
  - Mean milestone score

- **Evaluator Characteristics:**
  - Number of years teaching
  - Faculty position (Assistant, Associate, Full Professor, Director, Chair)
  - Clinical supervision hours

- **Mixed Models show no significant relationships between evaluator characteristics and evaluator quality outcomes**

- **Models do show a relationship between delay and other quality outcomes**
  - The longer delayed evaluations have significantly higher milestone scores and the significantly lower competency variation
Steady progression of scores over time
Factor analysis across Institution “A” 5 programs shows one factor loading across competencies

Standard deviation analysis shows increased variation within evaluations for certain program’s CCC scores

<table>
<thead>
<tr>
<th>Program</th>
<th>Mean SD across Evaluations</th>
<th>Percent all same</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neurology</td>
<td>0.953</td>
<td>33.82%</td>
</tr>
<tr>
<td>OBGYN</td>
<td>0.416</td>
<td>1.10%</td>
</tr>
<tr>
<td>Psychiatry</td>
<td>0.368</td>
<td>1.01%</td>
</tr>
<tr>
<td>Anesthesia</td>
<td>0.141</td>
<td>0%</td>
</tr>
<tr>
<td>Radiology</td>
<td>0.212</td>
<td>42.85%</td>
</tr>
</tbody>
</table>
How does the CCC committee aggregate rotation data?

- Mean of six scores
- Predictive model controlling for random effects
- Last score of six

Mean of six scores most predictive of CCC scores

Significant relationship between rotation scores and CCC scores when time is included in the model
Milestone Score Analysis – Summary of Findings

1. Milestone evaluations are unreliable across specialties and institutions
2. Evaluators are more predictive of milestone scores than the residents
3. Evaluators often rate milestone competencies along one dimension
4. Time and milestone score is a non-linear function (jump at PGY transition)
5. Evaluator characteristics do not appear to influence milestone scores
6. Monthly milestone evaluations directly impact CCC scores for residents
What does this mean for your institution?
Narrative Comment Analysis – Research Questions

- Do open-ended narrative comments more reliably differentiate residents when compared to milestone evaluation scores?
- How common is constructive feedback in narrative comments?
- How much variation is there among attendings and residents in the amount of constructive feedback they provide and receive?
- Do evaluation prompts increase the amount of high quality feedback?
- What relationships are there between attending characteristics and narrative feedback quality?
Data extracted from the RMSs for end-of rotation attending comments

Data from 5 specialties at 2 institutions, 10 programs total

10,044 complete evaluations divided into 21,290 sentence chunks

<table>
<thead>
<tr>
<th>Program</th>
<th>Institution</th>
<th>Comment Chunks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neurology B</td>
<td>1681</td>
<td></td>
</tr>
<tr>
<td>Neurology A</td>
<td>1039</td>
<td></td>
</tr>
<tr>
<td>OBGYN B</td>
<td>529</td>
<td></td>
</tr>
<tr>
<td>OBGYN A</td>
<td>930</td>
<td></td>
</tr>
<tr>
<td>Psychiatry B</td>
<td>2793</td>
<td></td>
</tr>
<tr>
<td>Psychiatry A</td>
<td>3864</td>
<td></td>
</tr>
<tr>
<td>Anesthesia B</td>
<td>3471</td>
<td></td>
</tr>
<tr>
<td>Anesthesia A</td>
<td>3244</td>
<td></td>
</tr>
<tr>
<td>Radiology B</td>
<td>247</td>
<td></td>
</tr>
<tr>
<td>Radiology A</td>
<td>3492</td>
<td></td>
</tr>
</tbody>
</table>
Narrative Comment Analysis – Coding Procedure

- **Valence**: Positive, Neutral, Negative/Corrective

- **Specificity**: General, Specific

**Example 1: Positive General**
I really like this Resident. She is hard working, personable, takes initiative and goes the extra mile. These qualities are rare and greatly appreciated.

**Example 2: Neutral General**
Appropriate fund of knowledge for her level

**Example 3: Negative Specific**
After the meeting, the timeliness of her notes did improve, but still needs work on content

**Example 4: Positive Specific**
He was efficient, responsible, and went the extra mile with patient to create safety discharge plans that were feasible for patients in follow-up

**Example 5: Neutral General**
Continue reading.
Narrative Comment Analysis – Evaluator and Resident Variation Plot

- 40% of OB/GYN evaluators never provided critical feedback to residents
- There is greater variation in narrative critical feedback than milestone score among residents
Narrative Comment Analysis – Summary of Data

Overall Prompts: n = 31502

Overview: n = 13392
  - Complete: N = 6458
  - Nothing: N = 6934

Weaknesses / Improvement: n = 9548
  - Complete: N = 2306
  - Nothing: N = 7242

Strengths: n = 8112
  - Complete: N = 5347
  - Nothing: N = 2765

Positive: N = 180
Constructive: N = 898
Nothing: N = 1228
Narrative Comment Analysis “A” – Evaluator Characteristics

- No OB/GYN faculty teaching over 20 years provided negative feedback
- This relationship was also apparent with faculty rank
Narrative Comment Analysis – Summary

- OB/GYN narrative comments more reliably differentiate resident abilities compared to milestone scores (may apply more broadly to all specialties)

- There is significant variation between evaluators in the amount of critical feedback they provide

- Critical and constructive feedback is rare in narrative comments, occurring less than 10% of the time

- Prompting attendings for resident areas of improvement/weaknesses does not guarantee the comment will address this request (still less that 10% of these prompts were constructive)

- There is a significant positive relationship between faculty rank and the amount of constructive feedback given to residents. This is also true for years teaching.
What does this mean for your institution?
Current Directions – Feedback Model

Evaluator → Interpretation of data → Evaluator

Data provided to evaluator → Student

Providing Feedback → Evaluator → Student adapts behavior → Student
Current Directions – Feedback Model Breakdown

- **Data Provided to Evaluator**
  - Inadequate observational data to make an accurate assessment
  - Not a valid construct: The gradation of behaviors described do not exist

- **Evaluator Data Interpretation**
  - Unable to differentiate levels of competence
  - Incorrect reference points of competency
  - Cognitive overload: Unable to evaluate many competencies at once

- **Evaluator Providing Feedback**
  - Lack of descriptive feedback: poor motivation, lacking incentive structure, time
  - Lack of critical feedback: fear retaliation, uncertain, poorly documented, hurt relationship

Currently conducting resident focus groups, attending interviews, and online surveys to explore these hypotheses
Current Directions – A call for data collaborators
Measures of Resident Teaching Quality

- Student evaluations of teacher performance on rotations
- Student learning outcomes
- Observational Data of resident teachers
  - Observation data by peers or supervisors
  - Process-based analysis through electronic data
- Example: Assessing feedback quality
  - **Milestone score analysis**: Evaluator mean score, competency variation, timing
  - **Narrative comment analysis**: Positive/Negative, Specific
Resident Scores of Faculty Teaching

- Ceiling effect in teaching scores for all residency programs
- Across 5 programs from Institution A 62% of evaluations of faculty had no variance in scores

### Constructive Feedback Rating Distribution

#### Institution A Neurology

<table>
<thead>
<tr>
<th>Constructive Feedback Rating</th>
<th># of Ratings</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
</tr>
</tbody>
</table>

### Program Percent all same

<table>
<thead>
<tr>
<th>Program</th>
<th>Percent all same</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neurology</td>
<td>56.3%</td>
</tr>
<tr>
<td>OBGYN</td>
<td>53.4%</td>
</tr>
<tr>
<td>Psychiatry</td>
<td>52.3%</td>
</tr>
<tr>
<td>Anesthesia</td>
<td>75.2%</td>
</tr>
<tr>
<td>Radiology</td>
<td>69.2%</td>
</tr>
</tbody>
</table>
Individual Teacher Ratings

- Students evaluations of attending feedback does not adequately differentiate faculty performance
- Teacher ratings do not correlate strongly with process data
Effect of Student Ratings on Teaching Behavior

- In many residency programs teaching scores contribute to promotion and bonus decisions.
- Basing teacher ratings solely on student evaluations, faculty could be incentivized to avoid giving negative, specific feedback.
- Faculty may also be incentivized not to provide low milestone scores.
- Automated process measures allow program directors to identify faculty that may need extra training on differentiating competencies and submitting feedback in a timely, informative manner.
Current Directions – A call for data collaborators
Questions?

- Marty Keil: mkeil@stanford.edu
- Nancy Piro: npiro@stanford.edu
- Ann Dohn: aohn1@stanford.edu