Gender Group Differences in Milestone Ratings: Exploring Differences in Ratings by Individuals and Clinical Competency Committees

Stanley J. Hamstra, PhD
Kenji Yamazaki, PhD
Eric Holmboe, MD

Accreditation Council for Graduate Medical Education
Chicago, Illinois

shamstra@acgme.org
Relevant Disclosures

• Paid Employee of ACGME
Program of Research

Comparison of Male and Female Resident Milestone Assessments During Emergency Medicine Residency Training: A National Study.

Santen SA¹, Yamazaki K, Holmboe ES, Yarris LM, Hamstra SJ.
Comparison of Male vs Female Resident Milestone Evaluations by Faculty During Emergency Medicine Residency Training

Arjun Dayal, BS; Daniel M. O’Connor, BA; Usama Qadri, BA; Vineet M. Arora, MD, MAPP

IMPORTANCE Although implicit bias in medical training has long been suspected, it has been difficult to study using objective measures, and the influence of sex and gender in the evaluation of medical trainees is unknown. The emergency medicine (EM) milestones provide a standardized framework for longitudinal resident assessment, allowing for analysis of resident performance across all years and programs at a scope and level of detail never previously possible.

OBJECTIVE To compare faculty-observed training milestone attainment of male vs female residency training.
Dayal et al. 2017

- 359 residents from 8 EM programs (AY2013-15)
- Individual Faculty ratings
- Used Milestones Reporting Form on a mobile app as assessment tool
- Examined full Milestone integer levels
- At end of each year: PGY-1, PGY-2, PGY-3
Dayal et al. (2017): Results

**After one year of residency training:**
- Women rated higher than men on 15 (of 23) subcompetencies

**At graduation:**
- Men rated higher than women on all 23 subcompetencies
  - Δ’s 0.15 points on the 5-level Milestone rating scales
  - ≈ 3 to 4 months of additional training
Our Approach

• We looked for the same effect as Dayal et al. in our national database

• 10,082 residents from 842 programs in 4 specialties (EM, Peds, IM, DR)
  • (AY2013-15)

  • Used Milestone ratings reported to ACGME by CCCs
    • *(NB: ACGME does not see individual faculty ratings or EPAs)*

  • Examined half-Milestone levels (finest level available)

  • every 6 months
The Role of the Clinical Competency Committee (CCC)

- Review/synthesize **multiple performance assessments** for every resident over the previous 6 months

- Consensus judgment of Milestone achievements
  → reported to the ACGME using **Milestones Reporting Form**
## INTERNAL MEDICINE MILESTONES

### ACGME Report Worksheet

1. Gathers and synthesizes essential and accurate information to define each patient's clinical problem(s) (PC1)

<table>
<thead>
<tr>
<th>Critical Deficiencies</th>
<th>Inconsistently able to acquire accurate historical information in an organized fashion</th>
<th>Consistently acquires accurate and relevant histories from patients</th>
<th>Accrues accurate histories from patients in an efficient, prioritized, and hypothesis-driven fashion</th>
<th>Obtains relevant historical subtleties, including sensitive information that informs the differential diagnosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Does not collect accurate historical data</td>
<td>Inconsistently able to acquire accurate historical information in an organized fashion</td>
<td>Consistently acquires accurate and relevant histories from patients</td>
<td>Accrues accurate histories from patients in an efficient, prioritized, and hypothesis-driven fashion</td>
<td>Obtains relevant historical subtleties, including sensitive information that informs the differential diagnosis</td>
</tr>
<tr>
<td>Does not use physical exam to confirm history</td>
<td>Does not perform an appropriately thorough physical exam or misses key physical exam findings</td>
<td>Seats and obtains data from secondary sources when needed</td>
<td>Performs accurate physical exams that are targeted to the patient's complaints</td>
<td>Identiﬁes subtle or unusual physical exam ﬁndings</td>
</tr>
<tr>
<td>Relies exclusively on documentation of others to generate own database or differential diagnosis</td>
<td>Does not seek or is overly reliant on secondary data</td>
<td>Synthesizes data to generate a prioritized differential diagnosis and problem list</td>
<td>Synthesizes data to generate a prioritized differential diagnosis and problem list</td>
<td>Efficiently utilizes all sources of secondary data to inform differential diagnosis</td>
</tr>
<tr>
<td>Fails to recognize patient's central clinical problems</td>
<td>Inconsistently recognizes patients' central clinical problem or develops limited differential diagnoses</td>
<td>Efffectively uses history and physical examination skills to minimize the need for further diagnostic testing</td>
<td>Performs accurate physical exams that are targeted to the patient's complaints</td>
<td>Role models and teaches the effective use of history and physical examination skills to minimize the need for further diagnostic testing</td>
</tr>
<tr>
<td>Fails to recognize potentially life threatening problems</td>
<td>Uses collected data to deﬁne a patient's central clinical problem(s)</td>
<td>Effectively uses history and physical examination skills to minimize the need for further diagnostic testing</td>
<td>Synthesizes data to generate a prioritized differential diagnosis and problem list</td>
<td>Role models and teaches the effective use of history and physical examination skills to minimize the need for further diagnostic testing</td>
</tr>
</tbody>
</table>

### Consistently performs accurate and appropriately thorough physical exams

**Comments:**
Our Focus

- differences between direct observation of individual faculty and committee ratings
- extension from 8 programs to national database
- extension to other specialties beyond EM
Analytic Method

• Multilevel spline regression model
• Repeated measures within residents
• Residents were nested within program
Residents’ Milestones trajectories over time by gender

M=4.11
F =4.05
(Δ 0.06, n.s.*)
Results – EM

• Applied the same procedure to the other 21 subcompetencies

  • no significant differences at the beginning of training

  • no significant differences at time of graduation
Results – IM, DR, Peds

- Applied the same procedure to IM, DR, and Peds:
  - no significant differences at the **beginning** of training
Statistically-Significant Results by Specialty

At time of graduation

<table>
<thead>
<tr>
<th>Specialty</th>
<th>No. of Residents</th>
<th>Total # Sub-comp</th>
<th>Women rated higher than Men</th>
<th>Men rated higher than Women</th>
</tr>
</thead>
<tbody>
<tr>
<td>EM</td>
<td>1,340</td>
<td>23</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>IM</td>
<td>7,062</td>
<td>22</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>DR</td>
<td>846</td>
<td>12</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Peds*</td>
<td>2,634</td>
<td>14</td>
<td>6</td>
<td>0</td>
</tr>
</tbody>
</table>

*FM shows similar results in preliminary analyses
Conclusions

• We looked for the same effect as Dayal et al.

• We didn’t find it in EM

• We did find it in IM and DR,
  ...and reversed effect in Peds

• Why?
Take-Home Message

The Power of the CCC?

• committee deliberations may adjust for individual biases
  ▪ (see “CCC Guidebook” for recommended CCC practices to address biases)

• multiple assessments may spread “construct-irrelevant variance” across raters

This process/culture may be specialty-specific

Need qualitative studies of CCC processes to uncover how biases may be handled
Questions?

1) Gender Bias remains an understudied issue
   - Individual rater bias
   - CCC processes – Hauer et al. 2016; Conforti et al. 2018; etc.

2) Use of Reporting Forms as assessment tools

3) Faculty Development:
   - transfer the CCC deliberations to individual faculty

4) Unclear if “statistically significant” differences are “educationally-significant”
   - 0.5 is the minimum difference on the Reporting Form

5) What Might Explain the Different Results by Specialty?

shamstra@acgme.org
Example – Internal Medicine

Statistically significant gender differences at the time of graduation

Note: all p-values are Bonferroni-corrected by the number of eligible subcompetencies (22)
Dayal et al. (2017) Results
Further Reading

