REVIEWS

# Gender Bias in Resident Assessment in Graduate Medical Education: Review of the Literature

*Robin Klein, MD MEHP[1], Katherine A. Julian, MD[2], Erin D. Snyder, MD[3], Jennifer Koch, MD[4], Nneka N. Ufere, MD[5], Anna Volerman, MD[6,7], Ann E. Vandenberg, PhD, MPH[1], Sarah Schaeffer, MD, MPH[8], and Kerri Palamara, MD[9]From the Gender Equity in Medicine (GEM) workgroup*

[1]Department of Medicine, Division of General Internal Medicine and Geriatrics, Emory University School of Medicine, Atlanta, GA, USA; [2]Division of General Internal Medicine, University of California, San Francisco, San Francisco, CA, USA; [3]Department of Medicine, Division of General Internal Medicine, University of Alabama Birmingham School of Medicine, Birmingham, AL, USA; [4]Department of Medicine, University of Louisville, Louisville, KY, USA; [5]Department of Medicine, Division of Gastroenterology, Massachusetts General Hospital, Boston, MA, USA; [6]Department of Medicine, University of Chicago, Chicago, IL, USA; [7]Department of Pediatrics, University of Chicago, Chicago, IL, USA; [8]Department of Medicine, Division of Hospital Medicine, University of California, San Francisco, San Francisco, CA, USA; [9]Department of Medicine, Massachusetts General Hospital, Boston, MA, USA.

***BACKGROUND:*** Competency-based medical education relies on meaningful resident assessment. Implicit gender bias represents a potential threat to the integrity of resident assessment. We sought to examine the available evidence of the potential for and impact of gender bias in resident assessment in graduate medical education.

***METHODS:*** A systematic literature review was performed to evaluate the presence and influence of gender bias on resident assessment. We searched Medline and Embase databases to capture relevant articles using a tiered strategy. Review was conducted by two independent, blinded reviewers. We included studies with primary objective of examining the impact of gender on resident assessment in graduate medical education in the USA or Canada published from 1998 to 2018.

***RESULTS:*** Nine studies examined the existence and influence of gender bias in resident assessment and data included rating scores and qualitative comments. Heterogeneity in tools, outcome measures, and methodologic approach precluded meta-analysis. Five of the nine studies reported a difference in outcomes attributed to gender including gender-based differences in traits ascribed to residents, consistency of feedback, and performance measures.

***CONCLUSION:*** Our review suggests that gender bias poses a potential threat to the integrity of resident assessment in graduate medical education. Future study is warranted to understand how gender bias manifests in resident assessment, impact on learners and approaches to mitigate this bias.

*KEY WORDS:* gender bias; implicit bias; gender; assessment; evaluation; residency training; graduate medical education; postgraduate medical education.

J Gen Intern Med 34(5):712–9
DOI: 10.1007/s11606-019-04884-0


*Published online April 16, 2019*

## BACKGROUND

As graduate medical education shifts to a competency-based medical education model, meaningful assessment becomes of critical importance.[1] The "Next Accreditation System" of the Accreditation Council for Graduate Medical Education (ACGME) relies on frequent, criterion-based, authentic assessment of residents to inform judgments about resident progress. Ensuring meaningful assessment requires surveying for threats to the integrity of resident assessment.

One concern garnering attention is unconscious or implicit gender bias. Implicit gender bias refers to the way that culturally established gender roles and beliefs impact our perceptions and actions without conscious intention.[2] Manifestations of gender bias among practicing physicians include differences in patient referral patterns, compensation, and career advancement.[3–5] Evidence suggests that gender bias impacts faculty assessment of medical student learners. Studies of Medical Student Performance Evaluations (MSPE) found gender-based differences in the traits ascribed to students, with female students more frequently described using communal traits such as compassionate, caring, or empathetic.[6, 7]

Of concern in graduate medical education is if and how gender bias impacts assessment. Valid and meaningful assessment has important implications to both training programs and resident learners. Notably, training programs utilize assessments to determine resident progress, advancement, and competency.

To explore this, we reviewed the evidence for gender bias in resident assessment within graduate medical education.

## METHODS

A comprehensive literature review was performed to capture relevant primary studies for inclusion into this review. Figure 1 details the search strategy employed using PRISMA guidelines. First, independent scoping searches were performed by two reviewers and a medical librarian to explore types of evidence, gaps in the literature, and inform our search strategy.

A comprehensive search of the Medline and Embase databases was conducted in March 2018 and September 2018. From our scoping search, we found focusing search terms on gender as opposed to assessment was more effective in capturing relevant articles. Search incorporated three main search themes Gender, Gender Bias, and Graduate Medical Education using established MeSH terms. Gender was captured using the terms 'female, 'male', and 'gender.' Gender bias was captured using the terms 'gender bias', 'sexism', and 'prejudice.' Graduate Medical Education was captured using the terms 'internship and residency', 'graduate medical education', 'postgraduate medical education', and 'residency.' Boolean operator 'and' was used to combine themes and search fields included title, abstract, and keywords. Limits used include journal articles, publications dating between 1998 and 2018, and English language. Finally, manual search of references and citations of captured articles was performed, and potentially relevant articles were included in the review.

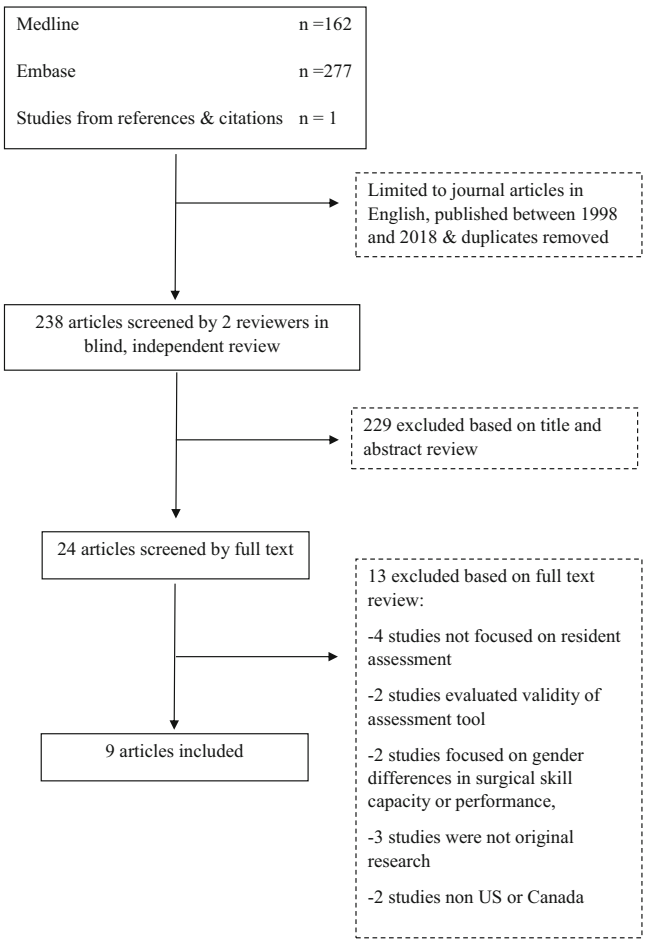


**Figure 1 Search strategy for literature review.**

A tiered review process was devised. After initial capture, two reviewers independently and blindly reviewed title and abstract to identify articles for in-depth review using the cloud-based platform Rayyan QCRI (http://rayyan.qcri.org). From this initial screen, select articles were retrieved and full text reviewed to determine inclusion. Interrater agreement was 92.0% for title and abstract review and 95.8% for full text review.

We included studies that examined the impact of gender on resident assessment in graduate medical education as primary outcome. For the purposes of this study, we defined assessment to include measures of resident performance used to inform determinations of resident competency, progress, and advancement. This included formal assessments of resident performance using structured assessment tools and qualitative comments and feedback. To this end, we excluded studies that presented gender data while assessing the validity of tests or indices and studies of gender-based differences on capacity to learn or perform tasks such as surgical procedures. We excluded articles that did not represent original research, such as reviews or commentaries, articles in which full text articles in English were not available such as conference abstract reports, and studies that occurred outside of the USA or Canada.

Data was extracted from full-text articles by one author and verified by review team. Data included participants, training setting, outcome measures, and findings. Strengths and limitations of studies were assessed individually and in aggregate. Quality characteristics of studies included sampling, assessment tools, study design, and analytic methods. Discord was settled by discussion and consensus of the team.

## RESULTS

The structured search strategy yielded nine unique studies meeting inclusion criteria.[8–16] Table 1 details extracted data and Table 2 details quality characteristics of included studies based on guidelines for qualitative and quantitative studies.[23, 24] Heterogeneity in methods and outcome measures across studies precluded meta-analysis and comparison of study quality via established indices.[24]

Data included 38,342 resident performance rating scores and 10,394 instances of narrative feedback for 1209 residents by 1287 faculty. Settings included family practice (FP),[15, 16] emergency medicine (EM),[8, 14] obstetrics and gynecology (ObGYN),[13] and internal medicine (IM) training programs.[9] Methodologies included quantitative,[8–12] qualitative,[14–16] and mixed methods approaches.[13] Eight of nine studies utilized resident assessments derived from direct observation in real-world practice[8–11, 13–16] while one study examined gender bias in a controlled setting using standardized encounters.[12] Assessment tools used included tools using Milestone framework[8, 14] standardized tool developed by various specialty groups[9, 10, 12, 13, and an institutional assessment tool.[11, 16]

**Table 1 Studies of Gender Bias in Resident Assessment in Graduate Medical Education**

| Study | Setting and participants | Design | Methods | Participants, residents, and faculty | Data | Findings |
|---|---|---|---|---|---|---|
| Resident performance metrics | | | | | | |
| Dayal A, et al. 2017[8] | 8 EM residency programs (6 academic, 2 community) | Retrospective longitudinal cohort | Examined impact of gender on resident assessment based on direct observation collected using a real-time online collection tool | 359 EM residents (122F, 237M)<br>285 faculty (91F, 194M) | 33,456 evaluations using ACMGE Milestone framework, 2013–2015 | By PGY 3, male residents rated higher than female residents in all 23 Milestones.<br>Overall rate of Milestone level attainment over time was 0.52 levels per year (95% CI = 0.49–0.53). Male residents had significantly higher rate of Milestone attainment (12.7% higher).<br>By end of training, Milestone scores were 0.15 levels higher for male than for female residents, equivalent to 3–4 months training.<br>No statistically significant difference in scores by faculty gender and resident faculty dyad. |
| Brienza RS, et al. 2004[9] | Academic IM residency program | Observational cohort | Examined influence of gender of resident faculty dyads on resident assessment | 160 IM residents (64F, 96M) on inpatient medicine rotations<br>88 faculty (18F, 70M) | 262 evaluations using ABIM assessment tool, 1997–1998 | No significant difference in performance scores attributable to gender pairs.<br>Trend to lower scores in F resident-M faculty dyad in clinical performance (7% lower, $P = 0.07$) compared to M resident-M faculty dyad. |
| Rand VE, et al. 1998[10] | Academic IM residency program | Observational cohort | Examined influence of gender on resident assessment | 132 PGY 1 and PGY 2 IM residents (47F, 85M) on inpatient medicine rotations<br>255 faculty (52F, 203M) | 974 evaluations using ABIM assessment tool, 1989–1995 | Male residents received significantly higher scores in all domains from male faculty compared to female faculty.<br>Male residents received significantly higher scores than female residents in 6 of 9 domains. |
| Thackeray EW, et al. 2012[11] | Academic IM training program, GI subspecialty | Observational cohort | Examined influence of gender on resident assessment by subspecialty faculty | 240 IM residents on GI clinical rotations<br>44 GI faculty (9F, 35M) | Evaluations using the ACGME Core Competencies, 2005–2010 | No significant difference in faculty ratings of residents attributable to gender. |
| Standardized encounters | | | | | | |
| Holmboe ES, et al. 2009[12] | 40 faculty from 16 IM residency programs | Post intervention | Examined impact of gender on faculty ratings of standardized encounters of residents performing clinical skills at varying competency | Standardized encounters, male residents depicted history taking and clinical skills, female resident depicted counseling skill<br>40 faculty (19F, 21M) | 348 ratings of taped standardized encounters using ABIM Mini-CEX, 2001–2002 | Mean ratings for female residents lower than male residents. No significant differences in ratings attributed to faculty gender. |
| Multisource feedback | | | | | | |
| Galvin SL, et al. 2015[13] | Community ObGYN residency program | Mixed methods | Examined impact of gender on assessments by nursing, including | 44 ObGYN residents (34F, 10M)<br>Nurses (100% female) | 2202 evaluations using Professional Associate | Female PGY 2 residents had significantly lower mean ratings than male |

(continued on next page)

Table 1. (continued)

| Study | Setting and participants | Design | Methods | Participants, residents, and faculty | Data | Findings |
|---|---|---|---|---|---|---|
| | | | rating scores and content analysis of qualitative comments | | Questionnaire, 2006–2014 | residents (1.5 vs 1.7, scale 0 to 2, $P = 0.001$). Female PGY 1 residents received fewer positive comments (17.3% vs 40%) and more negative agentic comments (17.3% vs 3.3%) than male residents ($P = 0.04$). |
| Qualitative comments | | | | | | |
| Mueller AS, et al. 2017[14] | Academic EM residency program | Qualitative thematic analysis | Thematic analysis of narrative comments; subset analysis examined consistency of feedback | 47 PGY3 EM residents<br>Subset: 35 residents (13F, 22M)<br>67 faculty (29F, 38M) | 1317 qualitative comments using ACMGE Milestone framework, 2013–2015 | Female residents received more discordant feedback about performance than male residents, particularly around assertiveness and receptivity to guidance. |
| Loeppky C, et al. 2017[15] | Academic FM residency program in Canada | Qualitative content analysis | Content analyses of archived real-time feedback provided to resident by faculty | 192 FP residents (104F, 88M)<br>464 faculty (188F, 276M) | 7316 instances of feedback using FieldNotes, 2012–2016 | Female faculty provided more feedback than male faculty. Female residents received more feedback than male residents. F resident-M faculty dyad had the highest proportion of communal and the lowest proportion of agentic adjectives. |
| Ringdahl EN, et al. 2004[16] | Academic FP residency program | Qualitative content analysis | Content analysis of narrative comments | 35 PGY 1 FP residents on inpatient rotations<br>44 faculty and senior residents | 1341 qualitative comments from 322 evaluations, 1996–1999 | No difference in content type or valence based on faculty gender. |

*M male, F female, FP family practice, EM emergency medicine, IM internal medicine, GI gastroenterology, ObGYN obstetrics and gynecology, PGY 1 post graduate year 1, PGY 2 post graduate year 2, PGY 3 post graduate year 3, P P value*

## RESIDENT PERFORMANCE METRICS

Four studies investigated gender bias utilizing faculty rating scores of resident performance. The largest study examined the influence of gender on resident assessment using data from 33,456 resident evaluations from eight EM training programs using an ACGME Milestone-based assessment framework.[8] Data was gathered using a real-time online collection tool and faculty elected which residents to assess and when. Dayal et al. found that while there was no difference in Milestone level ascribed to male and female residents at the start of training, by postgraduate year (PGY) 3, faculty ascribed higher levels in all 23 Milestones to male residents compared to female residents. Male residents had a significantly higher rate of Milestone attainment than female residents (12.7% higher or 0.07 Milestone levels per year) so that by the end of training, the discrepancy in Milestone level attained between male and female residents was equivalent to 3 to 4 months of additional training. There was no statistically significant difference in Milestone level by faculty gender or gender of resident faculty dyad.

Two smaller studies examined the impact of gender on resident assessment in IM training programs using an ABIM assessment tool.[9, 10] Study of 974 inpatient evaluations of IM residents found scores of male residents were significantly higher in six of nine domains compared to female residents and male faculty rated male residents significantly higher than female faculty in all nine assessed domains.[10] A later study of 262 IM resident evaluations from inpatient medicine rotations assessed the influence of faculty and resident gender pairings. While the female resident-male faculty dyad trended toward lower clinical performance scores (7% lower, $P = 0.07$), there was no significant influence attributable to gender pairing of resident and faculty.[9]

Looking at resident assessment in a subspecialty setting, analysis of 1100 evaluations of residents rotating on a digestive disease service at one institution found no significant difference in resident performance rating scores due to gender.[11] Interestingly, gender pairing was a significant factor in resident assessment of faculty.

## STANDARDIZED ENCOUNTERS

Holmboe et al. evaluated the impact of gender bias using standardized encounters as part of a larger faculty development intervention.[12] Forty IM faculty viewed and scored scripted encounters representing residents' history taking,

**Table 2 Quality Characteristics of Studies of Gender Bias in Resident Assessment in Graduate Medical Education**

| | | | | | |
|---|---|---|---|---|---|
| Quantitative studies | | | | | |
| Author | Sampling | Assessment data type | Assessment tool | Statistical analysis | Outcome level |
| Dayal A, et al. 2017[8] | Multiple institutions | Resident performance ratings by faculty | ACGME Milestone-based assessment tool, including 23 validated Milestones across 6 clinical competencies[b] | Mixed-effects linear modeling to determine association between Milestone attainment and gender | Faculty assessment of resident performance in real-world setting |
| Brienza RS, et al. 2004[9] | Single institution, single rotation type | Resident performance ratings by faculty | ABIM assessment tool[e] | Hierarchical linear modeling using M resident-M faculty dyad as reference | Faculty assessment of resident performance in real-world setting |
| Rand VE, et al. 1998[10] | Single institution, single rotation type | Resident performance ratings by faculty | ABIM assessment tool[e] | Difference in mean ratings by gender and mixed-effects linear modeling | Faculty assessment of resident performance in real-world setting |
| Thackeray EW, et al. 2012[11] | Single institution, single subspecialty rotation type | Resident performance ratings by faculty | Institutional assessment tool with 7 to 12 validated items, domains include 6 ACGME Core Competencies | Marginal effect on rating scores by gender using mixed-effects linear modeling | Faculty assessment of resident performance in real-world setting |
| Holmboe ES, et al. 2009[12] | Faculty from multiple institutions, 3% encounters not ratedz | Ratings of resident skills depicted in simulated, standardized encounters | ABIM Mini-CEX assessment tool[d] | Differences in mean ratings using regression analysis | Faculty assessment of resident performance in simulated, standardized encounter |
| Mixed methods | | | | | |
| Author | Sampling | Assessment data type | Assessment tool | Statistical analysis and qualitative approach | Outcome level |
| Galvin SL, et al. 2015[13] | Single institution | Resident performance ratings and qualitative comments by faculty | Professional Associate Questionnaire[c] | Differences in mean ratings using regression analysis<br>Thematic analysis included tiered coding approach with independent reviewers blinded to participants' gender. Outcomes include content focus, valence, and communal and agentic adjective use. | Faculty assessment of resident performance in real-world setting |
| Qualitative studies | | | | | |
| Author | Sampling | Assessment data type | Assessment tool | Qualitative approach | Outcome level |
| Mueller AS, et al. 2017[14] | Single institution | Qualitative comments by faculty | ACGME Milestone-based assessment tool, including 23 validated Milestones across 6 clinical competencies[b] | Thematic analysis included tiered coding approach with open coding to generate themes, focused coding with independent reviewers, and effort to blind reviewers to participants' gender. Outcomes include content type, valence, and consistency in feedback across faculty raters. | Faculty assessment of resident performance in real-world setting |
| Loeppky C, et al. 2017[15] | Single institution | Qualitative comments by faculty | FieldNotes included in Competency-Based Achievement System[a] | Content analysis included keyword frequency of communal and agentic adjectives. Outcomes include frequency of specific domains (sentinel habits, clinical domains, progress level) and communal and agentic adjective use | Faculty assessment of resident performance in real-world setting |
| Ringdahl EN, et al. 2004[16] | Single institution, single rotation type | Qualitative comments by faculty | Institutional assessment tool including 10 domains and comments, instrument validity not reported | Content analysis included tiered coding approach with independent reviewers. Reviewers were not blinded to participants' gender. Outcomes include content and valence. | Faculty assessment of resident performance in real-world setting |

*ABIM American Board of Internal Medicine, ABIM Mini-CEX American Board of Internal Medicine Clinical Examination Exercise, ACGME Accreditation Council for Graduate Medical Education, M male*

[a] *Canadian Competency-Based Achievement System uses FieldNotes as a tool for collecting real-time assessment and feedback on resident progress in sentinel habits (skills and habits that make a good physician) and clinical domains of the field*[17]

[b] *ACGME Milestone-based assessment tool, including 23 validated Milestones across 6 clinical competencies, including patient care and procedural skills, medical knowledge, professionalism, interpersonal and communication skills, practice-based learning and improvement, and systems-based practice*[18, 19]

[c] *Professional Associate Questionnaire developed by American College of Obstetrics and Gynecology's Council on Resident Education in Obstetrics and Gynecology Competency Task Force and includes domains of communication, compassion, reliability, integrity, responsibility, patient advocacy, and respect for patients, families, and staff*[20]

[d] *ABIM Mini-CEX includes domains of medical interview skills, physical exam skills, humanism and professionalism, clinical judgment, counseling skills, organization, and overall clinical competency*[21]

[e] *ABIM assessment tool includes domains of clinical judgment, medical knowledge, clinical skills, humanistic qualities, teaching, professionalism, medical care, and overall clinical competence*[22]

physical exam, and counseling skills at variable competency levels. In these encounters, male residents depicted history and physical exam skills and a female resident depicted the counseling skill. While ratings of the vignette featuring the female resident were lower than scores for the other vignettes, there was no statistically significant difference in ratings due to faculty gender.

## MULTISOURCE FEEDBACK

Galvin et al. examined the impact of gender on resident assessments by nursing staff collected as part of a 360° feedback system for a community-based OBGYN residency training program.[13] This mixed methods study analyzed 2202 rating scores and 420 narrative comments and found that female residents received significantly lower scores than male residents (1.5 vs 1.7, scale 0 to 2, $P = 0.001$) in the PGY2 year, which involved more opportunities to interact with nursing. Female interns received fewer positive comments and more negative agentic comments than male interns (17.3% vs 40% and 17.3% vs 3.3% respectively, $P = 0.04$).

## QUALITATIVE COMMENTS

Following the work of Dayal et al., a qualitative study examined 1317 narrative comments included in PGY3 EM resident evaluations.[14] Subgroup analysis of 35 residents with multiple evaluations found that female residents received more discordant feedback across faculty, particularly regarding autonomy, assertiveness, and receptiveness to oversight compared to male residents.

Two studies analyzed the influence of gender on resident narrative comments in FP training programs.[15, 16] Loeppky et al. analyzed the feedback faculty provide to residents in a Canadian FP training program using FieldNotes or real time, written feedback based on direct observation.[15] Analysis of 7316 FieldNote comments found that female faculty provided more feedback and female residents received more feedback than their male counterparts. The female resident-male faculty dyad had the highest proportion of communal adjectives and lowest proportion of agentic adjectives. Ringdahl et al. analyzed 1341 narrative comments from inpatient evaluations of FP interns by faculty and senior residents and found no significant difference in valence or content by faculty gender.[16]

## DISCUSSION

Our review examined the potential for and impact of gender bias on resident assessment in graduate medical education. From this, we surmise three key points. First, gender bias poses a potential threat to the integrity of resident assessment in graduate medical education. Five of nine reviewed studies reported a difference in outcomes attributed to gender including significant differences in resident performance metrics[8, 10, 13] and narrative assessment[13–15] including consistency and valence of feedback and traits ascribed to residents.

Strength of the evidence demonstrating an impact argues in favor of potential for gender bias in resident assessment. Studies that showed significant differences in resident performance metrics employed greater numbers of evaluations as data and utilized established assessment tools.[8, 10, 13] The most robust and persuasive evidence comes from the study by Dayal et al. which demonstrated a significant difference in Milestone attainment by resident gender.[8] The Milestone framework is the strongest measure of resident assessment across studies as the development and validation of the 23 EM Milestones is robust and well-defined.[18, 19] The large number of data points, multi-institutional design, and use of the Milestone framework adds to the robustness of the findings.

The complicated nature of implicit gender bias makes capturing this challenging. Studies assessing for gender bias must consider a variety of potential manifestations and employ appropriate methodologies and analytic models. Studies that found no difference in resident assessment attributed to gender were limited by fewer participants,[9, 12, 16] low proportion of female participants,[9, 11] or limited qualitative analysis.[16] Failing to detect a difference in outcomes may reflect a failure to capture gender bias rather than confirmation of no bias. In short, negative results are insufficient to rule out gender bias. Additionally, the critical importance of valid resident assessment in competency-based medical education raises the stakes on the assessment process. The argument may be made that any good evidence of implicit gender bias in resident assessment is a sufficient harbinger to warrant concern and vigilance.

The second key point surmised from our review is the complexity of how gender bias manifests in resident assessment. Manifestations and patterns of gender bias were not uniform across studies reviewed. While differences were noted by gender of resident,[8, 10, 13–15] evidence of an influence of faculty gender or gender of resident and faculty pairings was limited.[10, 15]

Gender bias is multifaceted and may arise when gender-based normative behaviors and expectations misalign with professional roles and behaviors.[25] It may emerge in specific context such as when performing professional roles as leader or manager or working with others within a team.[26–29] Providing some context, qualitative studies suggest expectations of interpersonal dynamics and issues of power may be at play. Female residents were more often assessed using communal or warmth-based descriptors and less often agentic or competency-related descriptors.[13, 15] Study of qualitative comments by Mueller et al. found that senior female residents more often received inconsistent feedback across faculty regarding traits of autonomy, assertiveness, and receptiveness to oversight.[14] Analysis of nursing narrative assessments indicated that female interns were more susceptible to bias in

assessment suggesting that female interns must contend with both gender roles and issues of power during training and may be expected to be "overly communal" to be effective.[13]

A third key point concluded from review of the existing evidence is the need for further exploration and study focused on gender bias in graduate medical education. This includes factors that underlie gender bias, impact on learners, and interventions to address gender bias in graduate medical education. While the evidence indicates that gender bias is a factor in resident assessment in graduate medical education, a key question remains wherein lies the source of the bias. Does the gender difference in resident assessment arise from an issue with the assessment tool, the learners, or the faculty evaluators? Evidence of gender bias was reported using assessment tools of variable strength including the robust EM Milestone framework[18, 19] suggesting that gender-based differences in outcomes were not a product of a particular assessment tool.

As some have suggested, the gender-based difference in outcomes may be due to difference in performance between male and female residents. Female residents may be operating under strain when their professional role requires them to act counter to gender-based normative behaviors. Study of IM residents' experiences with cardiopulmonary resuscitation found that female residents reported that the role of code leader required them to violate gender behavioral norms and experience tension related to competing expectations.[30] Conflict between professional role and gender normative behaviors may explain a disparity in performance in contexts where the clash between gender role and role as resident is heightened, such as directing care in the emergency department, in the labor and delivery unit, or on an inpatient ward team.[8, 10, 13, 14]

Gender-based difference in outcomes may be due to disparity in how faculty assess resident performance. Faculty operate within the same gender climate as their learners, and their experiences navigating this may influence their assessment of learners. A survey of clinician educators at a Swedish academic center reported that female faculty more frequently cited gender of both faculty and learner as an important factor in their teaching.[31] A faculty development intervention aimed at addressing gender bias led to a significant increase in faculty awareness of personal bias and self-efficacy to promote gender equity.[32]

Another important topic warranting further study is the impact of gender bias on learners, as evidence of long-term impact on trainees is lacking. One study looking at the impact of gender bias on learners in primary and secondary education found that early gender bias exposure influenced learners' later achievements and had implications for career and earning potential.[33] We postulate that discordant and non-specific feedback may be a lost opportunity to assess skills and deficiencies and may undercut female residents' training experience. As resident assessments are used to inform progress through training,[1] gender bias in assessment may impact advancement and duration of residency training. Assessments are sourced for programmatic letters of recommendations for employment and fellowship and bias may result in professional disadvantages in terms of professional opportunity and growth.

Lastly, interventions to address gender bias in graduate medical education are needed. These interventions should include promoting open dialog about gender and bias among trainees and faculty, educational innovations to enable residents to manage tension due to competing roles and expectations, and faculty development to mitigate bias in assessment and feedback.

Our review highlights the importance of the issue of implicit gender bias in resident assessment. Limitations of our review include heterogeneity in methodology, assessment instruments, and outcome measures utilized made direct comparison between studies difficult. Relatively small sample size and low proportion of female participants limited results in some studies. Studies examined gender bias within different environments and specialties and continuity of contact between learner and evaluator is an unknown yet potentially relevant variable. Studies employed different assessment tools with variable evidence of tool validity. Our review did not address the intersection of gender bias and racial bias and this topic warrants dedicated study. Lastly, reviewed studies relied on a gender binary construct which does not account for those who identify on the gender continuum.

## CONCLUSION

Review of the evidence indicates that gender bias poses a potential threat to the integrity of resident assessment in graduate medical education. Despite noted differences, the majority of reviewed studies found gender-based difference in outcomes including significant differences in resident performance metrics and disparity in narrative comments. Manifestations of gender bias in resident assessment are complex and challenging to study. Given the importance of resident assessment in competency-based medical education, future study is needed to better understand how gender bias manifests in assessment, impact on learners, and interventions to address gender bias in graduate medical education.

***Corresponding Author:*** *Robin Klein, MD MEHP; Department of Medicine, Division of General Internal Medicine and Geriatrics, Emory University School of Medicine, Atlanta, GA, USA (e-mail: rklein3@emory.edu).*

***Compliance with Ethical Standards:***

***Conflict of Interest:*** *The authors declare that they do not have a conflict of interest.*

## REFERENCES


1. **Holmboe ES**, **Sherbino J**, **Long DM**, **Swing SR**, **Frank JR**. The role of assessment in competency-based medical education. Med Teach 2010;32(8):676–82.
2. **Risberg G**, **Johansson EE**, **Hamberg K** (2009) A theoretical model for analyzing gender bias in medicine. Int J Equity Health 8:28.
3. **Jena AB**, **Olenski AR**, **Blumenthal DM**. Sex differences in physician salary in US public medical schools. JAMA Intern Med 2016;176(9):1294–1304.
4. **Wehner MR**, **Nead KT**, **Linos K**, **Linos E**. Plenty of moustaches but not enough women: cross sectional study of medical leaders. BMJ 2015:16;351:h6311.
5. **Sarsons H.** Interpreting signals in the labor market: evidence from medical referrals. Job Market Paper. 2017 Nov 28.
6. **Axelson RD**, **Solow CM**, **Ferguson KJ**, **Cohen MB**. Assessing implicit gender bias in medical student performance evaluations. Eval Health Prof 2010;33(3):365–85.
7. **Ross DA**, **Boatright D**, **Nunez-Smith M**, **Jordan A**, **Chekroud A**, **Moore EZ**. Differences in words used to describe racial and gender groups in Medical Student Performance Evaluations. PLoS One 2017;12(8):e0181659.
8. **Dayal A**, **O'Connor DM**, **Qadri U**, **Arora VM**. Comparison of Male vs Female Resident Milestone Evaluations by Faculty During Emergency Medicine Residency Training. JAMA Intern Med 2017;177(5):651–657.
9. **Brienza RS**, **Huot S**, **Holmboe ES**. Influence of gender on the evaluation of internal medicine residents. J Women's Health 2004;13(1):77–83.
10. **Rand VE**, **Hudes ES**, **Browner WS**, **Wachter RM**, **Avins AL**. Effect of evaluator and resident gender on the American Board of Internal Medicine evaluation scores. J Gen Intern Med 1998;13(10):670–4.
11. **Thackeray EW**, **Halvorsen AJ**, **Ficalora RD**, **Engstler GJ**, **McDonald FS**, **Oxentenko AS**. The effects of gender and age on evaluation of trainees and faculty in gastroenterology. Am J Gastroenterol 2012;107(11):1610–4.
12. **Holmboe ES**, **Huot SJ**, **Brienza RS**, **Hawkins RE**. The association of faculty and residents' gender on faculty evaluations of internal medicine residents in 16 residencies. Acad Med 2009;84(3):381–4.
13. **Galvin SL**, **Parlier AB**, **Martino E**, **Scott KR**, **Buys E**. Gender Bias in Nurse Evaluations of Residents in Obstetrics and Gynecology. Obstet Gynecol 2015;126 Suppl 4:7S–12S.
14. **Mueller AS**, **Jenkins TM**, **Osborne M**, **Dayal A**, **O'Connor DM**, **Arora VM**. Gender Differences in Attending Physicians' Feedback to Residents: A Qualitative Analysis. J Grad Med Educ 2017;9(5):577–585.
15. **Loeppky C**, **Babenko O**, **Ross S**. Examining gender bias in the feedback shared with family medicine residents. Educ Prim Care 2017;28(6):319–324.
16. **Ringdahl EN**, **Delzell JE**, **Kruse RL**. Evaluation of interns by senior residents and faculty: is there any difference? Med Educ 2004;38(6):646–51.
17. **Ross S**, **Poth CN**, **Donoff M**, **Humphries P**, **Steiner I**, **Schipper S**, **Janke F**, **Nichols, D**. Competency-Based Achievement System: Using formative feedback to teach and assess family medicine residents' skills. Can Fam Physician 2011; 57(9): e323–e330.
18. **Beeson MS**, **Holmboe ES**, **Korte RC**, **et al.** Initial validity analysis of the emergency medicine milestones. Acad Emerg Med 2015;22(7):838–844.
19. **Beeson MS**, **Carter WA**, **Christopher TA**, **Heidt JW**, **Jones JH**, **Meyer LE**, **Promes SB**, **Rodgers KG**, **Shayne PH**, **Swing SR**, **Wagner MJ**. The development of the emergency medicine milestones. Acad Emerg Med 2013;20(7):724–9.
20. American College of Obstetricians and Gynecologists. Professional associate questionnaire. Washington (DC): ACOG; 2003. Available at: https://www.acog.org/About-ACOG/ACOG-Departments/CREOG/CREOG-Search/CREOGCOMPENTENCY-PRESENTATIONS. Retrieved April 1, 2018.
21. **Holmboe ES**, **Huot SJ**, **Chung J**, **Norcini JJ**, **Hawkins RE**. Construct validity of the mini-clinical evaluation exercise (miniCEX) Acad Med 2003;78:826–30.
22. **Thompson WG**, **Lipkin M**, **Gilbert DA**, **Guzzo RA**, **Roberson L**. Evaluating evaluation: Assessment of the American Board of Internal Medicine resident evaluation form. J Gen Intern Med 1990;5:214.
23. **Côté L**, **Turgeon J**. Appraising qualitative research articles in medicine and medical education. Med Teach 2009; 27(1): 71–75.
24. **Cook DA**, **Reed DA**. Appraising the Quality of Medical Education Research Methods: The Medical Education Research Study Quality Instrument and the Newcastle–Ottawa Scale-Education. Acad Med 2015; 90(8); 1067–1076.
25. **Eagly AH**, **Karau SJ**. Role congruity theory of prejudice toward female leaders. Psychol Rev 2002; 109:573–598.
26. **Heilman ME**, **Haynes MC**. No credit where credit is due: attributional rationalization of women's success in male-female teams. J Appl Psychol 2005; 90(5):905–16.
27. **Heilman ME**, **Wallen AS**, **Fuchs D**, **Tamkins MM**. Penalties for success: reactions to women who succeed at male gender-typed tasks. J Appl Psychol 2004;89(3):416–27.
28. **Lyness KS**, **Heilman ME**. When fit is fundamental: Performance evaluations and promotions of upper-level female and male managers. J Appl Psychol 2006: 91: 777–785.
29. **Heilman ME**, **Okimoto TG**. Why are women penalized for success at male tasks? the implied communality deficit. J Appl Psychol 2007;92(1):81–92.
30. **Kolehmainen C**, **Brennan M**, **Filut A**, **Isaac C**, **Carnes M**.Afraid of being "witchy with a 'b'": a qualitative study of how gender influences residents' experiences leading cardiopulmonary resuscitation. Acad Med 2014;89(9):1276–81.
31. **Risberg G**, **Hamberg K**, **Johansson EE**. Gender awareness among physicians–The effect of specialty and gender. A study of teachers at a Swedish medical school. BMC Med Educ 2003;3, 8.
32. **Carnes M**, **Devine PG**, **Baier L**, **Byars-Winston A**, **Fine E**, **Ford CE**, **Forsher P**, **Isaac C**, **Kaatz A**, **Magua W**, **Palta M**, **Sheridan J**. Effect of an Intervention to Break the Gender Bias Habit for Faculty at One Institution: A Cluster Randomized, Controlled Trial Acad Med 2015; 90(2): 221–230.
33. **Lavy V, Sand E.** On the origins of gender human capital gaps: Short and long term consequences of teachers' stereotypical biases. National Bureau of Economic Research; 2015 Jan 30. https://doi.org/10.3386/w20909.