Effect of Evaluator and Resident Gender on the American Board of Internal Medicine Evaluation Scores

Victoria E. Rand, MD, Esther S. Hudes, PhD, MPH, Warren S. Browner, MD, MPH, Robert M. Wachter, MD, Andrew L. Avins, MD, MPH

OBJECTIVE: **To examine the effects of resident and attending physician gender on the evaluation of residents in an internal medicine training program.**

DESIGN: **Cross-sectional study.**

SETTING: **Large urban academic internal medicine residency program.**

PARTICIPANTS: **During their first 2 years of training, 132 residents (85 men, 47 women) received a total of 974 evaluations from 255 attending physicians (203 men, 52 women) from 1989 to 1995.**

MEASUREMENTS: **The primary measurements were the numerical portions of the American Board of Internal Medicine evaluation form. Separate analyses were performed for each of the nine evaluation dimensions graded on a scale of 1 to 9. The primary outcome was the difference in the average scores received by each resident from male versus female attending physicians.**

RESULTS: **Compared with female trainees, male residents received significantly higher scores from male attending physicians than from female attending physicians in six of the nine dimensions: clinical judgment, history, procedures, relationships, medical care, and overall. Similar trends, not reaching conventional levels of statistical significance, were observed in the other three categories: medical knowledge, physical exam, and attitude. These differences ranged from 0.24 to 0.60 points, and were primarily due to higher grading of male residents by male attending physicians than by female attending physicians.**

CONCLUSIONS: **In one academic training program, we found a significant interaction in the grading process between the gender of internal medicine residents and the gender of their attending evaluators. This study raises the possibility that subtle aspects of gender bias may exist in medical training programs.**

KEY WORDS: **interprofessional relations; social environment; teaching; internship; residency.**
J GEN INTERN MED 1998;13:670–674.

Received from the Division of General Internal Medicine (VER), Department of Epidemiology and Biostatistics (ESH, WSB, RMW, ALA), and Department of Internal Medicine (RMW), University of California, San Francisco, and the General Internal Medicine Section, Veterans Affairs Medical Center, San Francisco, Calif. (WSB, ALA).

Address correspondence and reprint requests to Dr. Rand: Division of General Internal Medicine, University of California, San Francisco, 400 Parnassus Ave., Box 0320, San Francisco, CA 94143.

Concern has been raised about the possibility of gender bias in medical training programs.[1] One 1993 survey identified problems of sexual harassment and inappropriate experiences of both male and female residents during their training.[2] If such problems extend to the evaluation process of trainees, then residents of both genders may be adversely affected for reasons unrelated to their medical competence and clinical performance.

The American Board of Internal Medicine (ABIM) Resident Evaluation Form is a commonly used tool to provide residents with feedback regarding their strengths and weaknesses. It contains questions for judging trainees on nine dimensions of clinical proficiency and professional conduct, each using a 9-point scale. Previous studies have examined the ability of the evaluation form to reliably measure the quality of residents' performance.[3,4]

In this study, we examine the potential effect of resident and attending physician gender on the evaluation process. We performed a cross-sectional study of the standard ABIM evaluations of internal medicine residents in their first 2 years of training.

METHODS

Subjects and Measurements

The study subjects were medical residents enrolled in a 3-year internal medicine residency at a major urban multihospital training program. Residents from both categorical and primary care programs were included. The months of observation included all 1-month general medicine ward rotations at the three separate teaching hospitals; excluded were intensive care unit, coronary care unit, emergency department, and outpatient rotations. Ward months were studied because these rotations allowed for the most consistent contact between residents and attending physicians. Residents were studied during the first 2 years of their training because they spend only a few months on general medicine wards during their third year.

Residents who did not have on file at least one evaluation from both a male and female attending physician were excluded. To maintain confidentiality, all subjects were identified only by randomly assigned code numbers, and all procedures were approved by the local institutional review board. Measurements included the gender of the resident and the attending physician, the hospital site where the evaluation took place, the training year of the resident, the academic rank and type of attending physician (generalist vs subspecialist), and the numerical score on each of the ABIM dimensions. Each item was measured on a 1-to-9

scale, with 9 assigned the best score. The dimensions evaluated were clinical judgment, medical knowledge, history taking, physical exam, procedures, relationships, medical care, attitude, and the overall score.

Statistical Analysis

The primary outcome measure was the mean difference in scores each resident received from male and female attending physicians. To derive this measure, evaluation scores from male and female evaluators were averaged separately for each resident. The mean score given each resident by female attending physicians was then subtracted from that resident's mean score given on the same items by male attending physicians. A positive value indicated that, on average, a resident received higher scores from male than from female attending physicians. To compare results for male and female residents, the difference score for female residents was subtracted from that for the male residents; this "difference of the difference" score would be 0 if male and female residents were rated similarly or were differentially rated in the same manner by male and female attending physicians; values other than 0 indicate the presence of a gender interaction in the resident grading. In these analyses, each resident served as his or her own control.

A one-sample Student's *t* test was used to test the null hypothesis that there was no difference in the mean scores that residents received from male and female attending physicians; these analyses were performed separately for male and female residents. Two-sample *t* tests and Wilcoxon Rank Sum Tests were used to test the null hypothesis that the difference between the mean scores given by male and female attending physicians was the same for both male and female residents. Results were similar for both analyses, so only the results of the *t* test are presented.

Because the number of evaluations submitted for each resident varied, we also performed a series of weighted analyses. We used several different weighting schemes, including the total number of evaluations for each resident, inverse-variance weights, and the number of evaluations each resident received from female attending physicians (to examine the effect of variability in contact with female attending physicians). These analyses did not differ substantially from the unweighted analyses, so only the latter are presented.

As a confirmation of the results, we also analyzed the data with a mixed-effects generalized linear model, treating residents as clusters. In these analyses, the outcome was the individual scores given to each resident by each attending evaluator. The predictor variables in each model included the gender of the resident, the gender of the attending evaluator (both treated as fixed effects), and their interaction, the test of which was the primary analysis of interest. The individual resident and attending physician identifiers were entered as random effects. The analysis was carried out twice, once including all evaluations, and again including only the first evaluation of each unique resident-evaluator pair, if multiples existed. The model assumes that, conditional on the individual resident, evaluations are statistically independent; this assumption may be violated if there are multiple resident-evaluator pairs. The results from the two analyses were similar, so the model including all evaluations is presented.

All analyses were repeated for each of the nine items on the evaluation form, with no adjustment for multiple comparisons.[5] All analyses were performed with Stata, version 5.0 (Stata Corp., College Station, Tex., 1997), and SAS, version 6.10 (SAS Institute, Cary, N.C., 1991).

RESULTS

The study sample consisted of 85 male residents and 47 female residents who were evaluated by 203 male attending physicians and 52 female attending physicians (Table 1). Of the 219 attending physicians on whom specialty information was available, 79 (36%) were classified

Table 1. Characteristics of Male and Female Internal Medicine Residents

Characteristic	Male Residents	Female Residents
Number of residents	85	47
Mean number of evaluations (range)		
Male attending evaluators	5.7 (1–15)	6.0 (1–15)
Female attending evaluators	1.8 (1–5)	2.0 (1–4)
Mean evaluation item score (SD)		
Clinical judgment	7.96 (.92)	7.85 (1.0)
Medical knowledge	7.88 (.94)	7.59 (1.0)
History taking	7.99 (.94)	7.94 (.94)
Physical exam	7.87 (.93)	7.88 (.94)
Procedures	8.04 (.92)	7.86 (.97)
Relationships	8.05 (1.06)	8.28 (.94)
Medical care	7.96 (.97)	7.89 (1.02)
Attitude	8.20 (.96)	8.29 (.94)
Overall	8.04 (.90)	7.93 (.95)

as generalists; of the 200 attending physicians for whom academic rank was known (and did not change during the study period), 102 (51%) were at the assistant professor or instructor level. There were 610 evaluations available for male residents and 364 evaluations available for female residents. Residents received a mean (±SD) of 7.4 (±2.8) medicine ward evaluations, and attending evaluators provided a mean of 3.8 (±3.2) total evaluations during the study period. The mean number of evaluations by male and female attending physicians of male and female residents is shown in Table 1. Almost all scores given to residents were in the range of 6 to 9. For example, in the overall dimension, 94% of scores were between 7 and 9.

Male trainees received significantly higher scores from male attending physicians than from female attending physicians (Table 2). These differences ranged from 0.26 to 0.48 points and were significantly different for all dimensions. For the overall dimension, the global difference score was 0.32. For female residents, there was a trend toward higher evaluation scores by female than by male attending physicians in eight of the nine dimensions, though these differences were relatively small and none reached statistical significance.

Compared with female residents, male residents were graded significantly higher by male attending physicians relative to female attending physicians in six dimensions; trends in the same direction were observed in the other dimensions (Table 3). For example, in the overall dimension, this "difference of the difference" score was 0.43, indicating that the difference in scores received from male and female evaluators was almost half a point higher for male residents than for female residents.

We performed several subgroup analyses examining differences between training of attending physician (general internist vs subspecialist), and the academic rank of the evaluator (assistant vs associate or full professor). There were no substantial differences from the main analyses found in the mean gender-difference scores in either of these subgroup analyses.

Results using the mixed linear model were generally similar to those of the other analyses, though the *p* values tended to be higher and, in some cases, the differences were no longer statistically significant (Table 3). For example, the *p* values for the differences between male and female residents increased from .02 to .07 for clinical judgment and from .01 to .06 for relationships. The *p* value for the difference score for the overall dimension increased from .01 to .03, retaining statistical significance.

DISCUSSION

Our results suggest that significant differences exist in the way in which residents are evaluated depending on the gender of both the trainee and evaluator. Such differences indicate an influence of factors in the evaluation process that are unrelated to the competence of the trainee. In these analyses, each resident served as his/her own control, thereby minimizing bias due to any potential superiority in performance of either resident–gender group.

The factor most responsible for the gender-related difference in grading scores was a difference in the way that male residents were evaluated by male and female physicians. It is impossible to know whether this difference was due to higher grades being given to male residents by male attending physicians or to lower scores being given by female attending physicians, or both. Female residents tended to be graded similarly by male and female attending physicians.

Scores on the 9-point grading scale were not distributed over the entire range of possible scores. More than 95% of all scores given were at least 6, indicating that most evaluators used a de facto 4-point scale when evaluating resident trainees. In this context, small differences in scoring take on much greater importance; a difference of

Table 2. Within-Gender Analysis of Average Difference in Scores Residents Received from Male Attending Evaluators and Female Attending Evaluators*

	Male Residents		Female Residents	
Evaluation Item	Difference Score	95% Confidence Interval	Difference Score	95% Confidence Interval
Clinical judgment	0.39†	(0.19, 0.59)	−0.02	(−0.03, 0.25)
Medical knowledge	0.26‡	(0.04, 0.48)	−0.05	(−0.36, 0.26)
History taking	0.28‡	(0.07, 0.49)	−0.09	(−0.35, 0.18)
Physical exam	0.29‡	(0.07, 0.52)	−0.07	(−0.36, 0.23)
Procedures	0.48‡	(0.17, 0.78)	−0.12	(−0.48, 0.25)
Relationships	0.41†	(0.18, 0.65)	−0.08	(−0.29, 0.13)
Medical care	0.26§	(0.04, 0.48)	−0.12	(−0.41, 0.17)
Attitude	0.34†	(0.14, 0.54)	0.10	(−0.13, 0.32)
Overall	0.32‡	(0.12, 0.51)	−0.11	(−0.39, 0.16)

**Positive scores indicate that, on average, residents received higher scores from male attending physicians than from female attending physicians.*

†$p < .001$.

‡$p < .01$.

§$p < .05$.

Table 3. Between-Gender Analysis of the Difference of the Difference Scores for Male and Female Residents

Evaluation Item	Difference of Difference Score	95% Confidence Interval	*p* Value*	*p* Value†
Clinical judgment	0.41	(0.08, 0.74)	.02	.07
Medical knowledge	0.31	(−0.06, 0.68)	.10	.14
History taking	0.37	(0.03, 0.71)	.04	.05
Physical exam	0.36	(−0.01, 0.73)	.06	.14
Procedures	0.60	(0.12, 1.08)	.02	.01
Relationships	0.49	(0.14, 0.84)	.01	.06
Medical care	0.38	(0.01, 0.75)	.04	.21
Attitude	0.24	(−0.08, 0.55)	.14	.72
Overall	0.43	(0.10, 0.76)	.01	.03

*p *Value from Student's* t *test.*
†p *Value from mixed linear model.*

one-half point, for example, carries substantially more relevance in a 4-point scale than in a 9-point grading scale. These gender differences are in the range of small-to-medium-sized effects, as characterized by Cohen, in which small and medium effect sizes (mean score divided by the standard deviation) are defined as 0.2 and 0.5, respectively.[6] For example, among male residents, for whom the differences are most pronounced, the effect size was 0.43 for the overall dimension; as noted, the effect size among female residents was considerably smaller. Although all of the differences noted were less than 1 point, the presence of significant gender-influenced evaluation indicates potential problems with the grading process.

None of the subgroup analyses revealed important interactions, specifically with type of attending physician (generalist vs subspecialist) or faculty rank. Previous analyses identified three general domains evaluated on the 9-item ABIM form (biomedical knowledge, interpersonal qualities, and technical abilities),[3,4] but the magnitude of the gender differences in the evaluation process did not seem to vary across these domains.

There have been few prior studies on gender-related issues in the medical training and evaluation process. Day et al. investigated how program directors (gender unspecified) evaluated medical residents using the ABIM form, and found that men were rated higher than women on procedures and medical knowledge, and women were rated higher than men on humanistic qualities.[7] A study by Smith et al. showed that physicians consistently rated women residency applicants more favorably than male applicants.[8] More recently, Colliver et al. found no interaction between examinee gender and patient gender on standardized patients' ratings of the examinees' interpersonal and communication skills.[9] Solomon et al. examined the interaction between the gender of preceptors and the gender of third-year medical students and found no significant interaction of gender in the evaluation process.[10] Hayward et al. also found no evidence of gender bias in evaluations of surgical residents, but this analysis was done between gender groups and did not account for within-subject differences.[11]

Our data indicate that complex gender-related effects may exist in the medical training environment. Such effects may have serious negative consequences for students and housestaff seeking to learn in a stressful environment and may have repercussions for residents in their future training and employment opportunities. Training programs have an obligation to provide an unbiased learning and evaluation experience and to explore means for ensuring that all individuals are treated fairly and with respect.[1] The current president of the Association of American Medical Colleges noted that "Gender and diversity issues are becoming more rather than less troublesome to address."[12] According to Myers, some faculty feel justified in perpetuating a standard of behavior to which they became inured as students, failing to recognize the power imbalance in the attending–trainee relationship.[12]

Although many academic training programs have begun to initiate programs to promote the equitable status and treatment of all students and faculty, the focus has been mainly on issues surrounding sexual harassment. For example, our university's policy states explicitly that "all persons who participate in university programs and activities should be able to work in an atmosphere free from all forms of harassment." Organizational commitment to such policies may help institutionalize unbiased treatment of all medical students and housestaff and assist in creating a learning environment in which gender issues do not adversely affect program trainees. However, gender bias is more subtle than sexual harassment, and, while there are programs to deal with the former, there are no specific programs that address this issue. Providing more guidance to attending physicians on rating criteria and educating faculty on the need for objective evaluation may help promote an unbiased grading process. Investigating alternative grading methods, such as having attending physicians meet as a group to discuss housestaff performance or increasing resident involvement in the evaluation process, may also help diminish problems. Programs may also want to monitor their own circumstances using methods such as those employed in these analyses, and evaluate the effectiveness of interventions designed to enhance gender-neutral evaluation.

Several limitations of this study should be noted. First, in general, residents were not evaluated by the same attending physicians. Therefore, we cannot state conclusively that our results were due to gender differences alone. However, the consistency of the findings for both male and female residents on nearly all evaluation dimensions suggests that the observed differences were unlikely to be chance effects. Second, some bias may have occurred in the process of pairing residents and evaluators, who may choose with whom they work. However, if residents are choosing attending physicians in part because they perceive a problem with obtaining objective evaluations or to obtain an unfair advantage, this would further indicate a

problem in the grading process. Third, only 2 years of evaluations were examined for each resident, and most evaluations are now several years old. It is possible that the effects observed may have begun to dissipate in response to an increased emphasis on promoting diversity in medical education. Finally, this study was carried out within a single department at one academic institution. These results may not generalize to all such institutions or other departments.

In sum, we observed small but statistically significant effects of resident–attending physician gender pairing on the evaluation of medical trainees. Our study implies that gender influences what is intended to be an objective performance evaluation. As teaching institutions strive to incorporate greater sensitivity to diversity in their programs, subtle gender-related issues that may inappropriately influence the evaluation process may require greater attention in the future.

REFERENCES

1. Gordon GH, Labby D, Levinson W. Sex and the teacher-learner relationship in medicine. J Gen Intern Med. 1992;7:443–8.
2. Komaromy M, Bindman AB, Haber RJ, Sande MA. Sexual harassment in medical training. N Engl J Med. 1993;328:322–6.
3. 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; 5:214–7.
4. Haber RJ, Avins AL. Do ratings on the American Board of Internal Medicine Resident Evaluation Form detect differences in clinical competence? J Gen Intern Med. 1994;9:140–5.
5. Rothman KJ. No adjustments are needed for multiple comparisons. Epidemiology. 1990;1:43–6.
6. Cohen J. Statistical Power Analysis for the Behavioral Sciences. Revised ed. New York, NY: Academic Press; 1977.
7. Day SC, Norcini JJ, Shea JA, Benson JA Jr. Gender differences in the clinical competence of residents in internal medicine. J Gen Intern Med. 1989;4:309–12.
8. Smith CJ, Rodenhauser P, Markert RJ. Gender bias of Ohio physicians in the evaluation of the personal statements of residency applicants. Acad Med. 1991;66:479–81.
9. Colliver JA, Vu NV, Marcy ML, Travis TA, Robbs RS. Effects of examinee gender, standardized-patient gender, and their interaction on standardized patients' ratings of examinees' interpersonal and communication skills. Acad Med. 1993;68:153–7.
10. Solomon DJ, Speer AJ, Ainsworth MA, DiPette DJ. Investigating gender bias in preceptors' ratings of medical students. Acad Med. 1993;68:703.
11. Hayward CZ, Sachdeva A, Clarke JR. Is there gender bias in the evaluation of surgical residents? Surgery. 1987;102:297–9.
12. Association of American Medical Colleges. Enhancing the Environment for Women in Academic Medicine: Resources and Pathways. Washington, DC: Association of American Medical Colleges; 1996.