American Sociological Association

American Sociological Review
1–29


DOI:10.1177/00031224231184264
journals.sagepub.com/home/asr
Sage

Taking the Time: The Implications of Workplace Assessment for Organizational Gender Inequality

Laura K. Nelson,[a] Alexandra Brewer,[b] Anna S. Mueller,[c] Daniel M. O'Connor,[d] Arjun Dayal,[e] and Vineet M. Arora[f]

Abstract

Gendered differences in workload distribution, in particular who spends time on low-promotability workplace tasks—tasks that are essential for organizations yet do not typically lead to promotions—contribute to persistent gender inequalities in workplaces. We examined how gender is implicated in the content, quality, and consequences of one low-promotability workplace task: assessment. By analyzing real-world behavioral data that include 33,456 in-the-moment numerical and textual evaluations of 359 resident physicians (subordinates) by 285 attending physicians (superordinates) in eight U.S. hospitals, and by combining qualitative methods and machine learning, we found that, compared to men, women attendings wrote more words in their comments to residents, used more job-related terms, and were more likely to provide helpful feedback, particularly when residents were struggling. Additionally, we found women residents were less likely to receive substantive evaluations, regardless of attending gender. Our findings suggest that workplace assessment is gendered in three ways: women (superordinates) spend more time on this low-promotability task, they are more cognitively engaged with assessment, and women (subordinates) are less likely to fully benefit from quality assessment. We conclude that workplaces would benefit from addressing pervasive inequalities hidden within workplace assessment, equalizing not only who provides this assessment work, but who does it well and equitably.

Keywords

gender, workplace inequality, assessment and feedback, work and occupations, medical education, mixed methods, qualitative methods, supervised machine learning

Despite decades of reforms to remove gendered policies and practices from the workplace, and despite the fact that women now outperform men in higher education, women remain underrepresented in leadership positions in both industry and academia. Substantial research has revealed how gender inequality is built into many organizations' promotion pipelines (e.g., Acker 1990; Correll et al. 2020), and these patterns can be found in even the most egalitarian organizations (Ridgeway 2011). The persistence of the gendered promotion gap has led

[a]University of British Columbia
[b]University of Southern California
[c]Indiana University Bloomington
[d]Mass General Brigham Wentworth-Douglass Hospital
[e]Rush Copley Medical Group
[f]University of Chicago

Corresponding Author:
Laura K. Nelson, Department of Sociology, University of British Columbia, AnSo 2111, 6303 NW Marine Drive, Vancouver, BC V6T 1Z1, Canada
Email: laura.k.nelson@ubc.ca

researchers to examine whether the "leaky pipeline" is in part generated by small behavioral differences between men and women that, as they accumulate, can produce gendered organizational outcomes.

Workplace task allocation is one of these processes: small differences in how men and women spend their time, or differences in how they are *asked* to allocate their time, can serve as a barrier to women's productivity and thus their advancement in the workplace. Workplace tasks fall on a spectrum of promotability, from high-promotability tasks that are framed as most important to an organization's bottom line and can directly lead to promotions, such as research, product development, or sales; to low-promotability tasks that are necessary for an organization to function but do not weigh as heavily in an individual's promotion, such as serving on committees and advising and mentoring; to non-promotable tasks that are incidental to an organization's core functioning yet still occupy employee time, such as organizing a special event for an office or helping coworkers or students with personal problems. Compared to men, women spend, on average, more time on lower-promotability tasks at the expense of time spent on higher-promotability tasks, contributing to gender inequality in promotions (Babcock et al. 2017; El-Alayli, Hansen-Brown, and Ceynar 2018; Guarino and Borden 2017; Miller and Roksa 2020; Mitchell and Hesli 2013; O'Meara et al. 2017; Winslow 2010) and implicating task allocation in the persistence of workplace gender inequality.

The causes of these behavioral differences are complex, ranging from gendered socialization to different implicit incentive structures for men and women to gendered interactions at work, but one frequently proposed solution to this form of inequality is to encourage women to change their behavior: women should simply spend less time on low-promotability tasks and more time on high-promotability tasks (e.g., Babcock et al. 2022). Yet we also know that many of the low-promotability tasks to which women are increasingly encouraged to "say no," such as mentoring and feedback, employee training, curriculum development, and committee work, are themselves implicated in workplace inequality. Tenure and promotion committees, for example, can be a barrier to women's advancement through university hierarchies, as committees often apply tenure criteria unequally to men and women (Weisshaar 2017). If there are differences in how men and women carry out their work on these committees, having women opt out of this crucial (and time-consuming) form of service may not improve gender equality in an organization, although it may help the woman opting out pursue her own next promotion. Indeed, research shows that women report more personal, emotional, and cognitive engagement with low-promotability tasks. This may compound the burden of this work on women superordinates, but it also suggests women may be approaching some tasks differently, and perhaps more thoughtfully, than men (Bellas 1999; Eagan and Garvey 2015). Having women opt-out, then, may deny subordinates access to thoughtful, emotionally engaged supports, contributing to the persistence of gender inequality.

Workplace task allocation thus presents a puzzle to those interested in understanding and addressing workplace inequality. On the one hand, that women spend more time on these tasks compared to men is one reason for the underrepresentation of women in leadership positions, suggesting women should minimize time spent on these tasks. On the other hand, ensuring these tasks are not only completed, but completed well, without bias, and with attention to underlying dynamics of bias and discrimination, is also crucial to addressing workplace inequality. While this puzzle is broadly acknowledged as important, it has been difficult to find data that allow researchers to unpack these complexities and link them to outcomes indicative of gender inequality. To be sure, qualitative research has suggested meaningful differences in how men and women approach low-promotability work and the self-reported consequences it

has for their stress (Bellas 1999; Eagan and Garvey 2015). But this research is heavily reliant on self-reports, which can carry perception biases, and generally struggles to connect these qualitative differences in work to consequences they may have for gender inequality in an organization as a whole. In short, we do not yet fully understand the scope of these potential differences or their role in gender inequality, and therefore, prior work has struggled to come up with appropriate and effective solutions.

We contribute to addressing these gaps in the literature by using the case of academic emergency medicine and novel longitudinal digital trace data that capture real-time engagement with one ubiquitous low-promotability task: workplace assessment of subordinates by superordinates. Our data directly capture in-the-moment evaluations of performance milestones for resident physicians (physicians in training) by emergency medicine (EM) attending physicians (fully-licensed physicians in charge of patient care and resident education). Our data came from eight accredited teaching hospitals in the United States over a two-year period (2013 to 2015). These data offer several unique strengths. First, the data include both textual commentary and numerical ratings of residents on nationally standardized and purportedly objective performance milestones set by the Accreditation Council for Graduate Medical Education (ACGME), the body responsible for accrediting all graduate medical training programs for physicians in the United States. Second, prior research using these data has established that over time, women residents are rated lower than their men counterparts, on average, suggesting evidence of gender inequality for residents produced by attendings in these hospitals (Dayal et al. 2017). Third, although attendings were directly asked to fill out these evaluations as part of their day-to-day tasks, the number of evaluations they submitted and the length or content of their comments were unlikely to directly affect their likelihood of promotion. Fourth, these evaluations matter to residents' future career trajectories: positive evaluations can contribute to being named "chief" resident during residency (a prestigious position) or being encouraged to go into academic medicine after residency concludes (a prestigious next step in their careers), while multiple negative evaluations could lead to remediation. Fifth, and perhaps most important, because we have data on the gender of both the attending and the resident physicians and longitudinal evaluations, we can examine gender differences in how attendings engage in this low-promotability task, and whether these differences shift based on the resident's gender. These data thus offer insights into the consequences this low-promotability task has for gender inequality—for both attendings and residents.

To do this, we combined qualitative and quantitative methods with machine learning to analyze the amount, content, and context of feedback that 285 attending physicians chose to provide 359 residents across 33,456 evaluations. First, we examined whether the extra engagement reported by women in qualitative interviews is reflected in empirically measurable differences in the way women approached this assessment work compared to men. Second, we investigated whether there were gender differences in the quality of that assessment work. Third, we reflected on the potential implications of these empirical differences for the supervisors spending time on these tasks, as well as the students and employees most in need of quality feedback and other organizational services.

We found a gendered pattern in who takes the time to provide high-quality and helpful assessments. Specifically, we found that women were more likely to be motivating in their assessments: not only did they provide more feedback, but their comments were more often helpful and offered task-specific content and reassurance (e.g., when mistakes were made) compared to men attendings. Men's assessments were more minimalist: they were more likely to provide a numerical evaluation with no written feedback, or to provide short feedback that was not particularly helpful to either the resident or the

training program. We also found evidence that these assessment practices likely mattered to gender inequality in residents' experiences of evaluation. While women attending physicians were more likely to provide helpful feedback to struggling residents, both women and men attendings demonstrated a positive bias toward men residents in their feedback: men residents were more likely than women residents to receive helpful feedback or reassuring comments, from both men and women attendings. Thus, the empirical evidence captured by our digital trace data suggests a gendered double jeopardy (Owens 2022): for many low-promotability tasks, including standardized performance evaluations captured in our data, more women superordinates take the time and dedicate the cognitive energy to carry out the task well, yet women subordinates do not always receive the full benefit of their supervisor's work.

GENDER INEQUALITY AND LOW-PROMOTABILITY WORKPLACE TASKS

Across both industry and academia there is a gendered promotion gap. Women hold almost 52 percent of all management and professional-level positions and are 52 percent of the college-educated workforce, yet they remain underrepresented in leadership positions (Warner, Ellmann, and Boesch 2018). In the legal profession, for example, women are 45 percent of associates but only 23 percent of partners and 19 percent equity partners; in finance, they constitute 53 percent of financial managers and 37 percent of financial analysts, but only 13 percent of chief financial officers in Fortune 500 companies (Warner et al. 2018). Women are also underrepresented in the upper and more prestigious levels in academia. In 2018, women were awarded 58 percent of undergraduate degrees, 58 percent of master's degrees, and 53 percent of doctoral degrees. Women held nearly half (50 percent) of all tenure-track positions in 2018, yet they held just 39 percent of tenured positions, and only a third (34 percent) of full professors were women (National Center for Education Statistics 2018).

In academic medicine this inequality is even more stark: even though women now enroll in medical school at a higher rate than men, they hold just 40 percent of faculty positions and represent only 25 percent of full professors (Association of American Medical Colleges 2019). Women face even greater barriers to career advancement in historically male-dominated fields within medicine. In emergency medicine, a mere 28 percent of all faculty are women (Bennett et al. 2019), despite studies showing women physicians have equal or better quality-of-care outcomes (Meier et al. 2019; Tsugawa et al. 2017). Women started earning at least half of all master's degrees in 1980 and at least half of all doctorate degrees in 2010. The underrepresentation of women in leadership positions is thus not an educational pipeline issue, but results in part from gendered promotion gaps as women advance in their careers (see, e.g., Marcotte, Arora, and Ganguli 2021).

Scholars have studied the multiple and complex causes of the gender promotion gap, including work-life balance (e.g., Ecklund and Lincoln 2016), tenure (Weisshaar 2017) and teaching (MacNell, Driscoll, and Hunt 2015) evaluation processes, the general conditions under which women work (Acker and Armenti 2004), and promotion criteria itself (Marcotte et al. 2021). Differences in workload distributions is one of the more well-documented gendered dimensions of workplace life and is likely an important cause of persistent gender inequality in leadership positions. Women spend significantly more time on lower-promotability tasks compared to men (Babcock et al. 2017; Mitchell and Hesli 2013; Winslow 2010), including committee work (Guarino and Borden 2017), advising and mentoring (O'Meara et al. 2017), support work for labs (Miller and Roksa 2020) and hospitals (Gupta et al. 2019), and more mundane citizenship tasks such as posing for promotional photos (Armijo et al. 2021), all at the expense of time spent on tasks more likely

to lead to promotions and raises. In addition to time spent on low-promotability tasks, personal accounts from men who are racial or ethnic minorities and women regardless of race/ethnicity detail the personal, emotional, and cognitive burden of supporting both their organizations and their subordinates (Bellas 1999; El-Alayli et al. 2018; Misra et al. 2021; Shayne 2017), which can lead to lower productivity and higher rates of stress and burnout (Eagan and Garvey 2015; Hart and Cress 2008).

Workplace assessment tasks are one ubiquitous type of low-promotability work. The process of assessing workers' performance for the purposes of hiring, firing, promotions, and pay, as well as developing in-house talent and improving workplace processes, is often traced back to the U.S. military's "merit rating" system during World War I. By the end of World War II, 60 percent of U.S. companies were using some form of merit ratings to make decisions about their employees; by the 1960s it was close to 90 percent (Cappelli and Tavis 2016). The process of assessing workplace performance has changed over the years, from simply assessing employees to improving employee talent and workplace processes and culture, but virtually every company now uses multiple forms of assessment, expending millions of hours on assessment every year (Buckingham and Goodall 2015).

Assessment is widely used in academia as well as industry. One survey documenting time faculty spend on low-promotability tasks (Ziker et al. 2013) found that faculty spend around four hours per week on advising and mentoring, five hours per week on service (e.g., committee work), and close to seven hours per week on administrative work (e.g., reporting, filling out forms). The survey did not have a separate category specifically for workplace assessment, but we can assume at least some of these 16 hours per week spent completing these lower-promotability tasks were devoted to workplace assessment, as academics assess virtually everyone in their day-to-day activities. Faculty and instructors, for example, assess students (e.g., grading, commenting on papers), MA and PhD advisees (e.g., commenting on work, in-person advising, writing letters of recommendation), colleagues (e.g., tenure and promotion and merit committees, teaching peer reviews), and their academic programs (e.g., graduate and undergraduate program assessments, accreditation reviews). These day-to-day assessment tasks rarely, if ever, directly contribute to promotions or tenure.

Regardless of the form it takes, assessment is often essential to organizational operations. Some form of evaluation and review is used by organizations across industry and academia at almost every decision-making point, including crucial decisions affecting career paths. Employers use information from assessments when determining salaries, pay raises, and promotions; universities rely on assessments for tenure and promotion decisions and when selecting candidates for admission to undergraduate and graduate programs.

Despite its ubiquity and importance, standardized assessment, particularly rote performance evaluations, are nearly universally disliked by both superordinates and subordinates (Castilla 2008). One management study found that 72 percent of organizations believed their performance evaluation process was not effective, and 58 percent of companies claimed that performance management (via evaluations) is not an effective use of time (Brandon Hall Group 2016). Providing assessment labor often comes at a professional cost, as it diverts employee hours away from core workplace tasks. In particular, the day-to-day work of providing workplace assessment is typically done by employees whose primary tasks are not assessment: faculty at research universities who primarily do research but who also assess their peers and their students; lawyers whose primary task is working on billable hours but who also assess paralegals and interns; doctors whose primary tasks are patient care and medical research but who also assess and train resident physicians. Even for individuals in management positions who assess their employees as a regular part of their job, they are themselves assessed not on

the number or content of the evaluations they complete, but on their ability to undertake challenging (rather than rote) tasks (Babcock et al. 2017:715; King et al. 2012). Workplace assessment is thus a ubiquitous and important low-promotability task that may be contributing to the overall unequal distribution of workplace tasks, yet research on who carries out assessment work—and, importantly, how and with what consequences to subordinates—is surprisingly sparse.

Challenges to Studying Workplace Assessment

The majority of research on the role of workplace assessment in reproducing inequality has focused on how performance feedback affects those being evaluated. Research has shown, for example, that quality assessments, in the forms of mentoring and feedback, are important for the success of women and racial/ethnic minorities in particular (Correll and Simard 2016). Yet research also shows stark inequality in who receives valuable assessment resources. For example, women are more likely than men to receive vague feedback that hurts their careers (Chopra, Arora, and Saint 2018; Correll and Simard 2016; Sambunjak, Straus, and Marusić 2006), and managers often evoke gendered frames as they assess workplace performance, leading to men and women workers being evaluated and valued using different criteria, particularly when evaluation criteria are not clear (Correll et al. 2020). What is less understood about workplace assessment, yet is potentially equally consequential given the large number of employee hours spent per year on this task, is how subtle differences in who provides these evaluations and how they carry out this task may also contribute to social inequalities. Who provides assessment labor, and, importantly, who does it well and equitably?

Analyzing the provision of workplace assessment has been stymied by data limitations. Quantitative data collected via self-reports has found that women spend more time on low-promotability tasks more generally, although none of these studies have focused on workplace assessment in particular. Moreover, self-reports, such as interviews and time-use surveys, are unreliable accounts of actual behavior (Jerolmack and Khan 2014). And quantitative differences in time spent on tasks do not capture engagement with that work and potential differences in how assessment work is carried out. Qualitative interviews suggest there is discretion in how workers approach these tasks, with differing cognitive effects of time spent on low-promotability tasks (Bellas 1999; Eagan and Garvey 2015). These qualitative interviews have been crucial in elaborating the nature of these differences, but they do not always allow us to understand the broad extent of these potential differences or their consequences for inequality in an organization.

Despite these challenges, theory suggests how gender may shape assessment work. Women, for example, report higher levels of engagement with low-promotability tasks, particularly tasks focused on serving subordinates, which may lead to more detailed and helpful feedback from women superordinates compared to men (Armijo et al. 2021). Additionally, in academia, students tend to perceive and expect women professors to be more nurturing than men professors and as such, will make additional demands of their women professors compared to their men professors (El-Alayli et al. 2018). Women often respond to these perceptions and additional demands by spending more time managing the feelings of their students, which may lead to women providing more reassurance in their feedback to students. There is competing evidence on how subordinates' gender may interact with the feedback provided. One study found that women faculty perceive women PhD students as less serious (Ellemers et al. 2004), suggesting women may be just as biased against women subordinates as their men colleagues. On the other hand, when women are in management positions, gendered pay gaps are reduced, suggesting women may be more egalitarian in supervisory roles (Shin 2012).

In short, research suggests that gender likely influences not only the amount of time superordinates will spend on assessment, but that women will likely be more engaged, helpful, and reassuring in their feedback compared to men. Due to challenges in collecting behavioral data, however, we have not been able to test whether and how gender affects the way workplace assessment is carried out. Our research addresses this need by leveraging the case of academic medicine, a novel behavioral dataset of real-time, real-world workplace assessment data, and by combining computational and qualitative methods.

THE CASE OF ACADEMIC MEDICINE

To advance our understanding of differences in the amount and quality of workplace assessment work done by women and men and its consequences, we used the case of on-the-job evaluations of resident physicians done by supervising attending physicians in U.S. hospitals. Residency is the stage of medical training that comes immediately after medical school, during which aspiring physicians apprentice in a medical specialty, such as emergency medicine. Residents possess medical degrees and work as doctors but are not fully licensed practitioners: they practice medicine under the supervision of an attending physician for several years before either entering independent practice or further specializing via a fellowship. Attending physicians are responsible for on-the-job mentoring and training of residents, in addition to other duties such as patient care, which contributes to the core mission of the hospital and brings in revenue, and, in academic hospitals, research.

Academic medicine is thus like many workplaces where senior colleagues are responsible for providing feedback to, mentoring, and supporting junior ones, but these tasks are relatively unrewarded. Even though resident evaluation, feedback, and training is centrally important to the medical profession—it not only helps residents advance in their careers, but medical schools and residency programs are evaluated by accreditors on the quality of feedback received—it is compensated less than other kinds of medical work and contributes less to promotions (Beasley, Simon, and Wright 2006; Mayer et al. 2014). Focusing on the case of academic emergency medicine (EM), we used a real-time capture of one type of workplace assessment task to analyze the way women and men attendings approach this low-promotability task, and the potential effects of these differences on their residents.

DATA

Study data came from EM residency training programs at eight hospitals across the United States between July 2013 and July 2015. We analyzed a unique dataset of 33,456 in-the-moment evaluations of residents by attendings, including numerical ratings on specific procedures and optional text providing feedback about resident performance to both residents and the hospital (a subset of 13,567 evaluations included textual comments). The evaluations were collected from 285 attending physicians (194 men and 91 women) based on their assessment of 359 EM residents (237 men and 122 women) through a smartphone application called InstantEval (version 2.0, Monte Carlo Software LLC, Annandale, VA, designed by two of our co-authors).

In line with the recent agile approach to assessment adopted by many companies, this app facilitates real-time, ongoing direct-observation evaluations of EM residents by allowing attending physicians to assess resident performance via a numerical score (on a scale of 1 to 5 that allowed for half points) and textual commentary. Each evaluation encouraged the assessment of 1 of 23 nationally standardized EM subcompetencies, or milestones, as set by the ACGME (ACGME 2015). Subcompetencies include skills that medical residents develop during their residency, including emergency stabilization, physical examination, and diagnosis. Attendings could choose whom to evaluate

and when (although most programs encouraged one to three evaluations per shift) and whether and how much written feedback to provide (with a limit of 1,000 characters). In addition to contributing to (re)accreditation, this assessment task helped the training program evaluate its own effectiveness and served as a form of feedback and mentoring to residents. Physicians were not professionally rewarded for providing more detailed evaluations, making this evaluation task similar to the types of assessment work faculty and managers are asked to do more generally.

The data are digital trace data, a form of found—as opposed to research ready—data that is becoming increasingly important in the social sciences (Salganik 2019). Like all data, digital trace data have benefits and limitations. Because our data capture real-time behavior, not after-the-fact accounts of behavior, it is not subject to the biases and subjectivity inherent in self-reported data based on perceptions of behavior. Because it is digital trace data, however, it was created for practical, not research, purposes. It is thus missing many details, including demographic details, that would be desirable for quantitative research. No information, for example, about nationality, race, ethnicity, seniority, or other demographics were collected by the application; this is a limitation of this study, as prior research indicates these factors intersect with gender in ways that may shape evaluations (Miller and Roksa 2020; Tiako, South, and Ray 2021). Attendings may also have provided in-person feedback directly to residents, which would not be captured as a digital trace. Digital trace data, similar to time-use surveys, also do not capture intentionality—the data capture what someone does, but not why they did it. We view our data and analyses as an important complement to, and extension of, existing studies based on research-ready data, while recognizing their limitations.

Mirroring other quantitative data on low-promotability tasks (e.g., Guarino and Borden 2017), there was a wide variance in the amount of time spent on this task and in the specific ways attendings carried it out. The number of textual comments submitted by each attending over the two-year period ranged from 0 to a maximum of 736 comments. The skew was also large. Fifteen percent (42) of attendings submitted numerical evaluations without submitting any textual comments, and an additional 9 percent (25) submitted only one textual comment; 15 percent (42) of the attendings provided over 100 comments each, comprising 70 percent of the total submitted comments. By analyzing the textual comments submitted via this app, we were able to examine in detail what, if any, of this variance was captured by gendered differences in the amount of evaluations completed and the content of those evaluations.

The textual corpus used in our primary analyses includes all the evaluations reported across all eight sites and over the two-year data collection period and the accompanied metadata, including attending gender, resident gender, resident year (they could be in year 1, 2, or 3 of a three-year program), hospital generic ID, type of hospital (whether it was academic or community), and numerical evaluation. The textual comments were directed at residents, to evaluate them and help them improve their practice, and to training program administrators seeking to evaluate the progress of specific residents and the training program as a whole.

We complemented a quantitative analysis of key features of these data with a more in-depth, qualitative analysis of one hospital (that we call "University Hospital") and all the evaluations with text from this hospital (n = 2,765) directed at first- and third-year residents. We chose University Hospital because it is the largest academic Emergency Department (ED) with two full years of data in our sample and had the largest number of textual comments. However, we also did extensive single-coder qualitative coding of three other EDs (two academic and one community ED) to ensure there was nothing idiosyncratic about University Hospital; we found the same patterns in these other hospitals. Using these 2,765 comments as training data, we used

supervised machine learning to quantitatively assess the quality of feedback across all comments, a process we describe below.

All names used in this text are pseudonyms to protect confidentiality. Additionally, we corrected typos and replaced medical abbreviations in all quotes presented in the results to improve readability for a general audience. We indicate this by placing brackets around the language we edited. This study was approved as exempt research by the University of Chicago Institutional Review Board.

METHODS AND MEASUREMENT

We used a combination of computational text analysis techniques and qualitative analysis to examine differences in how much feedback women provided compared to men and the type of feedback they provided. To measure the amount of feedback, we simply counted the number of evaluations and the number of words in each written evaluation provided. We considered this a proxy for the time spent on providing feedback. Second, we know from prior research that specific feedback is more valuable and helpful compared to abstract feedback for individuals trying to advance in their careers (Correll and Simard 2016). As this feedback was meant to evaluate residents on their achievement of milestones that largely related to medical competencies, we used a list of medical terms from the website MedicineNet[1] to count the number of medical terms in each written evaluation as a partial proxy for whether the feedback was specific to the residents' clinical performance.[2]

Words on their own, of course, do not convey the full content of the text, and simply writing more medical terms does not necessarily mean the comment provided high-quality feedback. To assess gender differences in the quality of feedback, we relied on our qualitatively coded data from University Hospital to guide supervised machine learning to classify all the evaluations. Our analytic procedure for qualitative data coding involved a multi-stage and multi-analyst process. Importantly, all qualitative coding was done by a mixed team of physicians and sociologists. We conducted iterative rounds of open and axial coding (Saldaña 2016). In this process, we identified a subset of 734 comments about resident errors from the first and third postgraduate year (PGY1 and PGY3) at University Hospital. At every stage of data analysis, more than one analyst coded all comments. When the two coders did not agree with a classification, it was discussed collectively, and consensus was reached in all cases.

This process resulted in several themes relevant to evaluating the quality of feedback. First, we identified instances of feedback in response to errors residents had made. These typically referred to mistakes in medical knowledge, judgment, patient care, or operations (e.g., efficiency, use of the electronic medical record, or note-taking). Errors are a normal part of residency, and attending feedback can be helpful in contextualizing what can be learned from errors and whether errors indicate a lack of skill that will be detrimental to a resident's medical career (Bosk 1979 [2003]). As a second step, we coded these comments as either helpful or unhelpful, depending on whether the attending included any suggestion for how the resident might improve their performance. A total of 367 of the 734 comments were coded as helpful. We also coded responses to errors for whether they provided any reassurance to the resident: language that let residents know that even though they had made a mistake they were still doing a good job. A total of 488 of the 734 comments were coded as containing reassurance. The physician co-authors confirmed that all comments identified as mentioning medical errors did in fact mention a medical error, and comments identified as containing helpful feedback in fact contained helpful information for residents.

With this subset of qualitatively coded comments, we used supervised machine learning to classify the remaining un-coded comments from all sites in two steps. We first transformed the hand-coded comments

from University Hospital into a comment/term matrix using TFIDF (term frequency/inverse-document frequency) as our feature space. We used TFIDF rather than word counts to account for common words, such as *patient*. We allocated 70 percent (1,935) of the 2,765 rows to the training set and the rest to the test set. Randomly assigning the training and test set each time, we used three supervised machine learning algorithms—Naive Bayes, Linear Support Vector Machines, and Radial Basis Function Kernel Support Vector Machines—to classify comments in three iterations: whether the comment was in response to an error, whether the comment was helpful, and whether it was reassuring. All algorithms achieved a strict exact match accuracy rate between .78 and .88 for the error category, .87 and .92 for the helpful category, and .87 and .89 for the reassuring category, all generally accepted accuracy rates on complex text coding tasks (Caruana and Niculescu-Mizil 2006). Linear Support Vector Machines produced the most accurate classification after 100 cross-fold validations; we used this trained model to classify all comments, including the hand-coded comments.[3]

We calculated the precision, recall, and F1 score—the harmonic mean between precision and recall—for all hand-coded comments. The F1 score was .73 for the error category, .70 for the helpful category, and .69 for the reassuring category (see Table 1). For each category, recall was much higher than precision. The trained model was thus more likely to generate false positives compared to false negatives, but this was the case for comments from both men and women. Our model thus likely over-counted the number of comments positively coded in each category, but this over-counting did not affect our calculations of the relative differences between men and women.[4] To ensure our accuracy was not affected by the choice of which comments were hand-coded, we did a final accuracy test by randomly selecting 50 comments the algorithm coded as helpful and 50 coded as reassuring from the seven sites that were not hand-coded and hand-checked the precision. The precision was .94 for helpful comments and .58 for reassuring comments.

Table 1. Precision, Recall, and F1 Score by Comment Code Category for the Final Supervised Machine Learning Model

	Precision	Recall	F1
Error	.61	.91	.73
Helpful	.58	.89	.70
Reassuring	.55	.94	.69

Source: Comments collected using InstantEval (V2.0 Monte Carlo Software LLC, Annandale, VA), across eight U.S. hospitals, 2013 to 2015.
Note: Three comment code categories: whether the comment was in response to an error (*error*), whether the comment was likely helpful (*helpful*), and whether the comment provided reassurance (*reassuring*). F1 is the harmonic mean of precision and recall.

A number of features of our data make standard, even non-parametric, statistical tests inappropriate for evaluating quantitative differences. Each attending contributed multiple comments and each resident was evaluated by multiple attendings. Furthermore, the data were nested in eight sites that were correlated with gender differences. It was difficult, if not impossible, to structure the data in a way that ensured the independence assumption was met. To account for the dependencies as much as possible, in our primary analysis we treated the individual attending as the unit of analysis, aggregating all comments by attending. To examine how gendered differences in evaluation may affect students, in a secondary analysis we treated each individual resident as the unit of analysis, aggregating all comments received by each resident. We constructed confidence intervals using the jackknife resampling method, aggregating the parameter estimates from each subsample of size $n - 2$, removing one woman and one man for each subsample. The jackknife is similar to and predates the more popular bootstrapping method. Bootstrapping, however, adds extra variance and is sensitive to outliers, particularly for smaller datasets. Jackknifing is

thus more appropriate for small, constructed datasets such as ours, and it is accurate for estimating confidence intervals for continuous measures such as means (Efron 1982).

To examine the effect of the context in which a comment was written, we repeated all our calculations separately for comments associated with low scores and comments associated with high scores, again using the jackknife method to construct confidence intervals. We designated a comment as low-scoring if it was associated with a score in the lowest quantile for the year of the resident receiving the comment, and high-scoring if it was associated with a score in the highest quantile for the year of the resident receiving the comment.

The nested nature of our data, combined with the fact that we did not have key demographic details about the attendings and residents in our data, means we could not use traditional statistical tests to evaluate quantitative differences that we found. Instead, we view our study as a rich, descriptive analysis of directly-captured behavioral data, finding, as we will report, meaningful and compelling empirical associations between gender and key outcome measures.

RESULTS

Overall, the mean (median) woman attending submitted 105 (44) evaluations compared to 130 (34) for the mean (median) man (see Table 2). As these descriptive statistics suggest, the skew was quite large, with 90 percent of attendings submitting 345 or fewer evaluations each. Of these 90 percent, the mean (median) woman submitted 72 (43) evaluations compared to 62 (29) for the mean (median) man. Of the 285 attending physicians in our data, women attendings were much less likely to submit a numerical evaluation without leaving a comment: 36 of the 194 men attendings (18 percent) submitted only numerical evaluations without leaving any comment, compared to only 6 of the 91 women attendings (7 percent). This suggests that on the most simple, objective measures, women on the whole spent more time and cognitive energy on this evaluation task. The textual comments provide more details about the way men and women approached this evaluation task. The data used in our analysis include the 13,567 evaluations that contain a comment contributed by 243 attendings who left at least one comment over the two years the data were collected; 65 percent were men ($n = 158$) and 35 percent women ($n = 85$).

Table 2 and Figures 1 and 2 summarize our main findings. Table 2 shows the summary statistics for all our measures, aggregated by attending and resident. Figure 1 shows the kernel density estimation plots for each of our measures by attending gender, truncated at the data limits for clarity of presentation, and Figure 2 shows our measures by comment, conditional on the score submitted with each comment. Together, these aggregate findings suggest that, across our five measures (i.e., total number of words, number of medical terms, comments in response to an error, helpful comments, and reassuring comments), although a few women and men performed the bulk of the assessment labor, including the labor of providing helpful and reassuring feedback, more men did the bare minimum on this evaluation task and more women did more (and more detailed) labor, and the labor was comparatively more distributed among the women. Figure 2 demonstrates that the differences in textual quantity and quality we found were not due to women and men scoring residents differently: for virtually every score attached to a comment, women provided more (and more helpful) feedback. Additionally, these differences were more pronounced for comments associated with low scores—times when a resident may have been struggling to complete a task. An in-depth quantitative and qualitative analysis of our five measures supports these findings.

Word Counts

Figure 3 shows the differences in means by the gender of the attending across all our measures, for all comments, comments

Table 2. Descriptive Statistics Aggregated by Attending and Resident Gender

	Mean	Median	Min.	Max.	Std.Dev.	Count	Mean	Median	Min.	Max.	Std.Dev.	Count
By Attending			*Women Attendings*						*Men Attendings*			
Word Count	1,393	425	4	23,305	3044	85	1,169	212	2	14,489	2160	158
Medical Terms Count	101	28	0	1,902	235	85	86	17	0	2,331	215	158
Error Count	17	5	0	382	47	85	13	3	0	181	24	158
Helpful Count	11	2	0	275	35	70	8	2	0	179	19	122
Reassuring Count	8	2	0	240	29	70	5	1	0	155	16	122
Comment Count	53	19	1	504	87	85	57	11	1	736	103	158
By Resident			*Women Residents*						*Men Residents*			
Word Count	462	259	4	3,252	512	222	466	307	1	2,452	452	430
Medical Terms Count	34	17	0	229	42	222	34	20	0	171	36	430
Error Count	5.7	3.0	0	55	6.5	222	5.3	3.0	0	43	5.8	430
Helpful Count	3.1	2.0	0	15	3.0	181	3.2	2.0	0	19	3.1	356
Reassuring Count	2.0	1.0	0	14	2.2	181	2.4	2.0	0	16	2.6	356
Comment Count	20	11	1	109	22	222	21	13	1	116	22	430

Source: Comments collected using InstantEval (V2.0 Monte Carlo Software LLC, Annandale, VA), across eight U.S. hospitals, 2013 to 2015.
Note: Measures are aggregated by attending or resident, by gender, over the two-year data collection period.

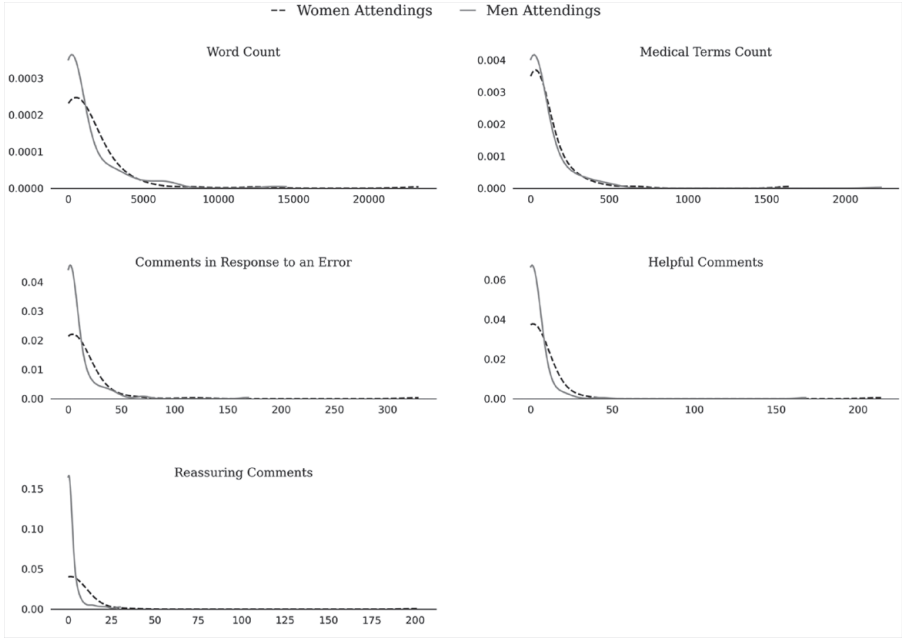

Figure 1. Truncated Kernel Density Estimation Plots across Five Measures of Feedback Quantity and Quality by Attending Gender
Source: Comments collected using InstantEval (V2.0 Monte Carlo Software LLC, Annandale, VA), across eight U.S. hospitals, 2013 to 2015.
Note: Curves truncated at the data limits.

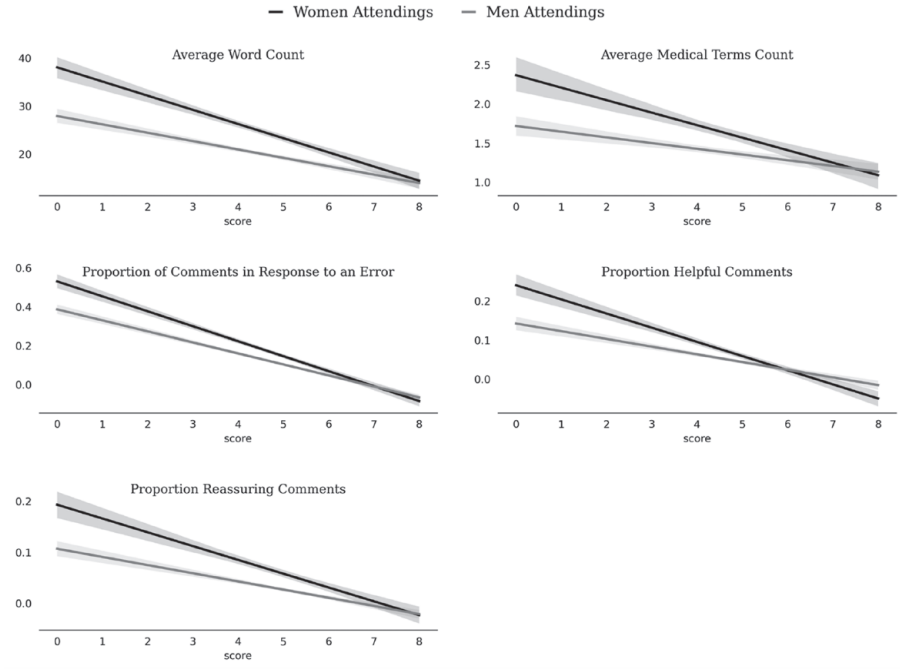

Figure 2. Regression Plot of Five Measures of Feedback Quantity and Quality by Comment, Conditional on Score and the Gender of the Attending
Source: Comments collected using InstantEval (V2.0 Monte Carlo Software LLC, Annandale, VA), across eight U.S. hospitals, 2013 to 2015.
Note: Shaded bands indicate the 95 percent confidence interval.

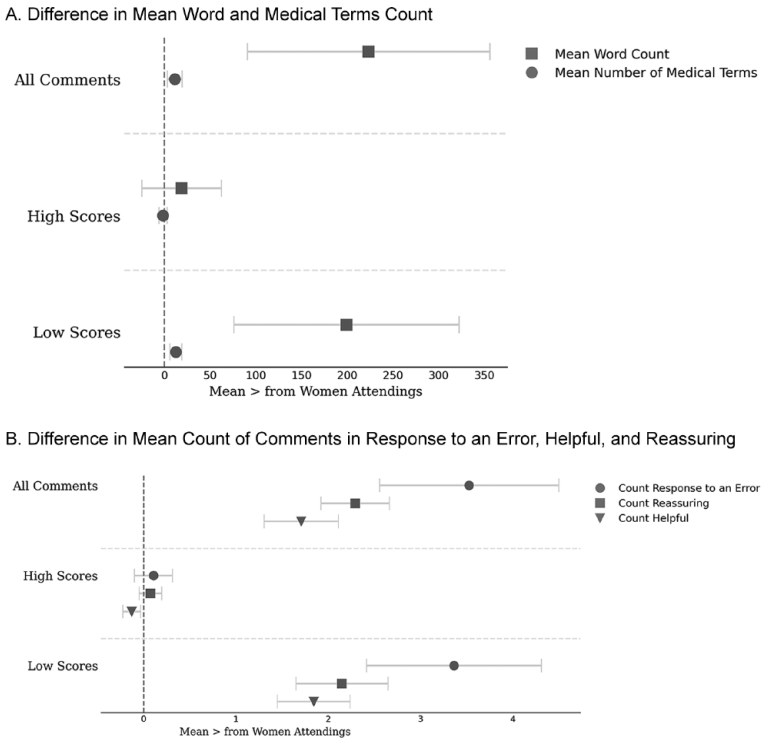

Figure 3. Mean for Women Attendings Minus the Mean for Men Attendings across Five Measures of Feedback Quantity and Quality

Source: Comments collected using InstantEval (V2.0 Monte Carlo Software LLC, Annandale, VA), across eight U.S. hospitals, 2013 to 2105.

Note: This figure shows the mean for men attendings subtracted from the mean for women attendings across five measures, for all comments, comments associated with high scores, and comments associated with low scores. Error bars represent the 95 percent confidence interval produced via the jackknife method. A positive difference indicates the mean for women is larger; a negative difference indicates the mean for men is larger.

associated with high scores, and comments associated with low scores, with the jackknife confidence interval. Women attendings tended to write longer comments compared to men attendings: the mean (and median) woman attending wrote 1,393 (425) words across the two-year data collection period, compared to 1,169 (212) words for the mean (and median) man (see Figure 3A). In our qualitative analyses, it was clear that longer comments were almost always more informative for residents than shorter comments. As a typical example, compare the following two comments directed at Gavin, a first-year (PGY-1) resident, by two different attendings (Harrison and Sofia). Both comments concerned Gavin's ability to assess patients and make plans for their care, but Sofia's longer comment contained much more information about what happened than did the shorter comment by Harrison:

> *Harrison for Gavin:* Gavin seems to be progressing well, but there are some concerns about focused assessments, sorting out key issues, and organized plans that seem to be lagging behind peers.
>
> *Sofia for Gavin:* Missed the need to do a full workup on a patient on chronic steroids with [congenital adrenal hyperplasia] who is immunosuppressed when she presented with fever. Although well appearing, she needed stress steroids and a [work-up]. Also missed the need to do a [urinalysis]/

Table 3. Sample Comments with Word Count

Comment ID	Attending	Resident	Word Count[a]	Comment
1	Mary	Faith	5	Great job at the intubation.
2	Brian	Faith	5	Great job with intubation today.
3	Greg	Faith	18	did a good intubation followed procedure well, asked appropriate questions, worked thru BURP [backwards upwards rightwards pressure] method to get better visualization
4	Sabrina	Faith	66	Quick to see patients. Asks appropriate questions in order to further learning (e.g., [X-Ray] vs MRI for evaluation of patient with back pain). Remember that asking patient about prior physician visits for same complaint as well as prior [workup] & results can be helpful when deciding what you want to do during this [Emergency Department] visit. Stayed late & attempted to finish/wrap care for her patients.
5	Richard	Faith	21	did a nice job with a no-medication intubation during compressions in a code; made a strong contribution in a busy shift
6	Eric	Faith	6	Good job on intubation during code.
7	Kristin	Faith	88	Busy University Hospital night shift: Faith worked hard to carry many patients (off service intern was tied up and saw two patients to her 8–10 in the last 4 hours of the shift) and did so well. She followed up on her patients, talked through systems-based practice issues that were new to her, was attentive in her re-evaluations and follow up on patients, accepted feedback readily, and was accurate in her assessments and plans. Very good job for being new to University Hospital on a pretty chaotic night.

Source: Comments collected using InstantEval (V2.0 Monte Carlo Software LLC, Annandale, VA), across eight U.S. hospitals, 2013 to 2015.
[a]Original, not revised, text.

> [urine culture] on another patient who presented with fever (with kidney abnormality) although also looked well. Wanted to send both patients home without evaluation. Didn't seem familiar with the notion of occult bacteremia or occult [urinary tract infection].

The longer comment provides more opportunity for learning and improvement, for both the resident receiving the feedback and the hospital seeking to improve its training program.

Even when comments contained mostly praise, longer comments still tended to be more useful for residents' learning and for the hospitals' insight into resident performance and the program's effectiveness. Compare, as a typical example, the comments directed at Faith, a resident, by attending physicians about her ability to intubate patients over the course of her first postgraduate year (PGY1) presented in Table 3. The shortest comments (Comments 1 and 2, Table 3) convey clear

Table 4. Most Frequently Used Medical Terms in Attending Comments to Residents

Word	Count
patient	3,786
trauma	534
resident	524
pain	449
intubation	392
medical	377
clinical	371
airway	338
intern	331
diagnosis	310
family	221
ultrasound	177
pediatric	170
feedback	165
EM[a]	161
acute	156
sense	153
chest	151
central	142
physician	140

Source: Comments collected using InstantEval (V2.0 Monte Carlo Software LLC, Annandale, VA), across eight U.S. hospitals, 2013 to 2015. Medical terms come from two lists maintained on MedicineNet (http://www.medicinenet.com): the Medical Dictionary List and the List of Procedures and Tests.
[a]EM is the common abbreviation for Emergency Medicine.

positive feedback: "Great job at the intubation." But they provide little context surrounding the intubation. Comment 3 is longer (18 words), and while similarly clear and positive, it lets Faith know she not only did the procedure well, but the way she asked questions of the attending was "appropriate" (asking appropriate questions is an important part of resident performance [Brewer et al. 2020; Mueller et al. 2017]). For the other attendings in the hospital, the comment lets them know that during this intubation, Faith tried the BURP (backwards upwards rightwards pressure) method; this is useful as it lets program leadership know the variety of skills the resident displayed while earning this particular evaluation. Longer comments (e.g., Comments 3, 4, 5, and 7) almost always provided more information to the residents (so they can maximize their learning) and the training program (so they can have an accurate sense of their residents' progress). Thus, they arguably represent higher-quality evaluations.

These comments also reveal another important difference in utility between feedback. Some comments (e.g., "Great job at the intubation") used minimal medical terminology (one word: intubation), whereas others, like Comment 4 in Table 3, used quite a few more (e.g., X-Ray, MRI, back pain, workup). Indeed, using data from all sites, our quantitative analysis revealed that women attendings wrote more medical terms compared to men attendings: the mean (median) woman wrote 92 (26) medical terms across the two-year period, compared to 80 (17) for the mean (median) man (see Figure 3A and Table 2). In general, our qualitative analysis shows that comments with more medical terminology provided more information for residents' ability to learn and the hospital's ability to evaluate. Table 4 lists the most frequently used medical terms. These words, such as *patient, trauma, resident, pain*, and *intubation*, suggest comments with more medical terms were focused on the way residents cared for patients and specific medical procedures. The higher word count combined with more medical terms per comments suggest women attendings not only provided a greater quantity of feedback, but they provided more medically specific, and thus potentially more informative and higher-quality, feedback.

Feedback Quality

To delve further into the quality of feedback rather than just the quantity, we examined how men and women attendings responded to errors made by residents. Errors are a normal and important part of residency (Bosk [1979] 2003); residents are often doing things for the first time and mistakes are inherent in learning complicated skills. They also represent important learning opportunities (Bosk [1979]

2003). Quantitatively, we found that women wrote slightly more comments in response to resident errors compared to men. The mean (and median) woman attending wrote 17 (5) comments in response to an error across the two-year period, compared to 13 (3) for the mean (median) man attending (see Figure 3B).

Noting an error in and of itself can lead to useful feedback, but we further examined whether the attending provided helpful, actionable information in response to the error, or whether they just noted the error without also providing instructions for improvement. For example, in both the following comments to Joe and Alec (first-year residents), Peter (the attending) notes they did not remember to follow up on orders. The comment from Peter to Joe notes this error, but without providing a strategy to do better in the future: "Occasionally forgets orders. Overall did well." We coded this comment as "unhelpful" because Peter describes potentially serious problems Joe is having in the ED, but does not provide further information about how he might improve his clinical skills. Compare this comment to the longer and more helpful one Peter wrote for Alec:

> Great job challenging yourself on a shift with sick patients. Remember to follow up on the labs that you order so that you can assess your management (frequent lactates in the gastrointestinal bleeding patient). You should be asking yourself, "is my resuscitation working?" or in other words, is the lactate correcting? As you continue to progress through intern year, try to plan ahead, so that if the lactate doesn't correct, you should already be planning your next step. In general, it is a great idea whenever you order a test to decide before it results what you will do if it is normal, and what you will do if it is abnormal. Read up on your patients every night, it is the best way to solidify the experience you gain during your shifts.

In this comment, Peter discusses the importance of following up on orders, provides details about the kind of thought processes the ideal emergency medicine physician has going on during the care of a patient experiencing a significant threat to their life (in ED jargon, a "sick" patient), and the role of recurring lab tests in that process. Table 5 provides additional examples of helpful versus unhelpful comments. Our quantitative analysis suggests women were more likely to write helpful comments compared to men: the mean (median) woman wrote 5 (1) helpful comments compared to 3 (0) written by the mean (median) man (see Figure 3B).

Reassurance is another important dimension of feedback quality. In our dataset, reassurance occurs when attendings provide positive commentary about residents' abilities alongside critical feedback; see comments 3 ("Good intubation") and 7 ("Good to work with her, competent and trustworthy") in Table 5. For residents attempting to learn and move on from mistakes, such comments might encourage them and reinforce the aspects of their skillset where they are performing well. Reassurance may also help communicate to residents that a certain number of mistakes are forgivable and normal, helping establish a pedagogical environment in which attendings can use errors as teaching moments (Bosk [1979] 2003). In our data, women were more likely to provide reassurance in their comments. The mean (and median) woman wrote 4 (0) comments that offered reassurance, compared to 2 (0) for the mean (median) man (see Figure 3B).

Feedback Context

As suggested in Figure 2, we found that the differences identified above were much more likely in comments associated with low performance scores (see Figure 3). For comments associated with low scores, the mean (median) woman wrote 924 (262) words and 66 (19) medical terms over the two-year period, compared to 725 (134) words and 60 (17) medical terms over the two-year period (see Figure 3A row 3). Ten (2) of the comments associated with low scores were written in response to an error for the mean

Table 5. Sample Comments Coded as Helpful and Not Helpful

Comment #	Attending	Resident	Word Count[a]	Comment	ML Prediction (error/helpful/reassuring)
				Helpful	
1	Steven	Landon	78	Could comment here – I think it's fair to limit procedure attempts at 3 so long as there is a backup option or an alternative approach. i.e. [lumbar puncture] or vascular access or some other procedure. If you get to 3 good attempts, adjusting your approach each time and it's still not happening, move on to another provider or another option. This has benefit of both limiting the amount of time you sink into something and preventing iatrogenic complications.	yes/yes/no
2	Lisa	Bennett	22	Remember to consider indications, contraindications, adverse effects when deciding on best agent for procedural sedation as 1 drug will not fit all.	yes/yes/no
3	Lisa	Ethan	53	Made good use of [endotracheal tube]. Pushed for results. Picked up patients (on with intern). Good intubation on moribund patients – it's okay to come out & bag if you have trouble passing [endotracheal tube] (while prepping a smaller [endotracheal tube]). Also remember that you should probably document events if greater than 1 pass to intubate.	yes/yes/yes
4	Lisa	Claire	36	Busy trauma shift – understandably got behind on charting but was able to complete in timely manner overall. Quickly performed [ultrasound] on patient with [gun shot wound] to flank (but don't forget need for [prim]ary survey as well).	yes/no/yes
				Not Helpful	
5	Victoria	Patricia	31	Didn't know the dose of succinylcholine or rocuronium for intubation. Lost focus during medical resuscutation. Wasn't truthful about physical examination component. Do not see the progress that she is reportedly making.	no/no/no
6	Charlotte	Patricia	16	Missed pul[monary] edema and 4/6 holosystolic murmur on patient who told her he had a murmur.	yes/no/no
7	Scott	Sonja	33	We were a little slow in instituting [early goal-directed therapy] in a patient with severe sepsis. Otherwise, she managed patients appropriately during a four-hour . . . shift. Good to work with her – competent and trustworthy.	yes/no/yes
8	Eric	Patricia	24	Mostly appropriate though I fear she tends to over order to compensate for lack of knowledge and understanding (example: fussiness in infant, "febrile" infant).	yes/no/no

Source: Comments collected using InstantEval (V2.0 Monte Carlo Software LLC, Annandale, VA), across eight U.S. hospitals, 2013 to 2015.
Note: ML prediction is the predicted code for each code category from the trained Linear Support Vector Machine classifier. The primary helpful/not helpful categories are from the qualitative coding.
[a]Original, not revised, text.

(and median) woman, compared to 7 (2) for the mean (median) man. When associated with low scores, the mean (median) woman wrote 5 (1) helpful comments and 4 (0) reassuring comments, compared to 3 (0) helpful comments and 2 (0) reassuring comments for the mean (median) man. For each of these measures, there was little or no difference between men and women attendings for comments associated with high scores. Together, these findings suggest women attendings, in contrast to men attendings, were taking the time and thought to focus on and help struggling students in particular.

Substantive Significance for Attendings

The difference in the amount, specificity, helpfulness, and supportiveness of feedback provided by women and men attendings is substantively important. Across the two years, men were more than twice as likely to provide numerical evaluations without any written feedback (18 compared to 7 percent). Of those who provided written feedback, the average woman wrote 17 percent more words (1,393 compared to 1,169), 14 percent more medical terms (92 compared to 80), 40 percent more comments in response to an error (12 compared to 8), 32 percent more helpful comments (12 compared to 8), and 50 percent more comments that offered reassurance (5 compared to 3). In other words, on this one narrow workplace task, the median woman was taking on substantially more of the feedback work (17 percent more in simple word counts), and between 14 and 50 percent more of the effort to provide substantive, helpful, and reassuring feedback.

Substantive Significance for Residents

What do these differences mean for residents? As a final step, we examined feedback amount and quality from the resident point of view. Research suggests quality feedback is more important for women compared to men (Correll and Simard 2016), so we analyzed differences in what type of feedback residents received by both resident and attending gender.

When residents received comments from women attendings, the mean resident received more words per comment compared to men (26 from women compared to men's 20) and more medical terms per comment (1.7 from women compared to men's 1.4). Residents who received comments from women were also more likely to receive comments in response to an error (22 compared to 14 percent). Of these comments, residents who received comments from women were more likely to receive helpful and reassuring comments (9 percent from women compared to 6 percent from men were helpful, and 8 percent compared to 4 percent were reassuring).

These differences are likely substantively important. Each resident received an average of 21 evaluations with text from faculty attendings over the two years the data were collected. If a hypothetical resident received all their evaluations from women attendings (an unlikely scenario but helpful for comparison), they would have received 26 percent more words compared to if they had received all their feedback from men (546 words versus 420 words[5]); 22 percent more medical terms (36 versus 29); 50 percent more comments mentioning a specific error they made (5 versus 3); 66 percent more helpful comments (2 versus 1), and 66 percent more reassuring comments (2 versus 1). In sum, residents who received comments from women received between 26 and 66 percent more chances at specific, helpful, reassuring, and performance-related feedback compared to residents who received feedback from men.

Additionally, we found differences in the type of evaluations received conditional on resident gender. Figure 4 shows differences in means when grouped by resident gender rather than attending. Women residents were more likely to receive comments in response to an error, but men residents were more likely to receive helpful comments from other men and reassuring comments from both men and women (see Figure 4B). This preference for men residents by both women and men

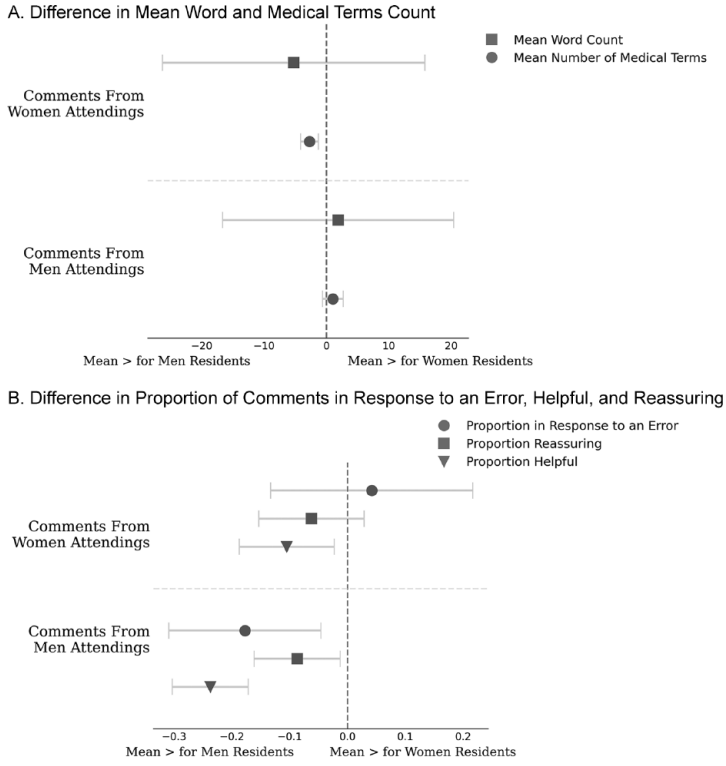

Figure 4. Mean for Women Residents Minus the Mean for Men Residents across Five Measures of Feedback Quantity and Quality by Attending Gender

Source: Comments collected using InstantEval (V2.0 Monte Carlo Software LLC, Annandale, VA), across eight U.S. hospitals, 2013 to 2015.

Note: This figure shows the mean for men residents subtracted from the mean for women residents across five measures, by gender of the attending. Error bars represent the 95 percent confidence interval produced via the jackknife method. A positive difference indicates the mean for women residents is larger; a negative difference indicates the mean for men residents is larger.

attendings is not surprising, as past research shows a statistically significant and substantial gender gap in resident evaluations using this same dataset (Dayal et al. 2017). If feedback is important for residents, they would have been helped more, all else equal, by receiving this feedback from women compared to men, but particularly if they themselves were men.

DISCUSSION AND CONCLUSIONS

Despite making many gains over the past several decades, women are still underrepresented in leadership positions across academia and industry, particularly in male-typed fields. The unequal gender distribution of low-promotability and high-promotability workplace tasks is one partial explanation for this feature of the stalled gender revolution (England 2010). The impact of workplace task allocation on organizational gender inequality is more complex than simply hours spent on different tasks, however. Qualitative research suggests women engage with these tasks differently than men in ways that may affect both superordinates and subordinates, but it has been difficult to operationalize and directly measure the extent and potential consequences of differences in the way the

same task may be carried out. In this article, we used observational data from one industry, academic emergency medicine, and one low-promotability task, workplace assessment, to better understand differences in how men and women approach low-promotability work tasks and the consequences it has for gender inequality in organizations. In so doing, we contribute to the understanding of persistent workplace inequality.

We analyzed behavioral data collected in real time via an app used by attendings in eight U.S. hospitals. This passively collected digital trace data allowed us to measure, without researcher interference, the actual way this workplace assessment task was completed, rather than subjective (and often faulty) accounts of time spent on workplace tasks. Our data included numerical evaluations of residents and, if a textual comment was submitted, the content of each comment. We then used a combination of qualitative coding and supervised machine learning to analyze these rich, numerical and textual data, which allowed us to scale our in-depth qualitative analyses of one hospital to the entire sample of eight hospitals in a way not feasible through qualitative methods alone.

Our analysis generated four key empirical findings relevant to workplace inequality. First, we found differences in how much men and women attendings wrote in their resident assessments as well as the content of their feedback. Women provided more feedback, more detailed feedback on relevant skills, and more helpful and reassuring feedback compared to men. Men's assessments were more minimalist in comparison: men more often provided scores without comments, or sparse comments less likely to be helpful. Women were thus not only taking the time to write more words, but their words were more helpful and reassuring, suggesting additional thought was put into writing those words. Second, we found that this gender difference in who took the time was the largest for comments associated with low scores—tasks where the learner struggled and where the hospital could potentially make the most improvements to their program. Third, we found that while a few attendings provided the majority of written feedback, the burden of providing this feedback was more evenly distributed among the women compared to the men. Fourth, we found that both men and women attendings were biased toward men residents, with men providing more helpful feedback to other men, and both men and women providing more reassuring feedback to men.

Implications for Research on Gender in the Workplace

These findings make several contributions to scholarship on gender inequality in the workplace. First and foremost, we advance this body of research by suggesting that simply measuring time spent on different workplace tasks does not fully capture the impact of workplace task allocation on workplace inequality. In addition to time spent on these tasks, the type of engagement required to complete these tasks can matter for workplace inequality in (at least) two important ways. First, women may approach carrying out these tasks differently than men, with potential implications for the cognitive burden of these tasks and thus their potential effect on employee burnout. Second, many low-promotability tasks contribute to structures of inequality in the workplace. If some of these low-promotability tasks are done with care, they can promote equality; if they are done carelessly, or with bias, they instead contribute to persistent inequality in the workplace. Understanding the implications of workplace task allocation on workplace inequality, then, requires not just understanding time spent on tasks, but how these tasks are carried out. We analyzed one ubiquitous low-promotability task, workplace assessment, to show how the *how* might matter.

By analyzing the full text of workplace assessment comments alongside accompanying quantitative scores, we found that women attendings in our data were more effective at supplying residents with feedback, in terms of

the quantity and length of evaluations as well as the content of their written assessments. This was particularly true for feedback for residents who they perceived as struggling. The extra effort expanded by women, either because of gendered socialization (Bellas 1999; Bowles and Babcock 2013; Shayne 2017) or gendered demands (El-Alayli et al. 2018), very likely leads to increased cognitive burden on the individuals providing these services—a cognitive burden that is difficult to directly measure yet has been shown to lead to more stress and burnout in white women and racial minorities (Eagan and Garvey 2015; Hart and Cress 2008). At the same time, we found the comments provided by women were likely more helpful to the residents receiving the evaluation. Comments written by women more often provided targeted and specific feedback, and women more often provided guidance and reassurance to residents when they made mistakes, contextualizing errors and providing residents the information they needed to correct those errors in the future. Specific and supportive feedback is particularly crucial for women looking to advance in their careers (Correll and Simard 2016).

That said, both men and women attendings demonstrated some bias in favor of men residents. Although not the primary focus of our analysis, we found that attendings (both men and women) were more likely to comment on errors if the resident was a woman, and both men and women attendings were more likely to provide reassuring feedback, and men to provide helpful feedback, if the resident was a man. This may help explain why other studies have found a significant and substantial gender gap in the ways subordinates are evaluated, in resident evaluations using this same data (Brewer et al. forthcoming; Dayal et al. 2017), and in academia and organizations more broadly (e.g., Correll et al. 2020; Weisshaar 2017). We found the effects of workplace assessment on gender inequality come from both sides: in our data, women attendings were more likely to take the time to carry out this workplace labor well, and women residents were less likely to fully benefit from motivating feedback.

Together, our empirical findings help clarify the mechanism through which differences in engagement with seemingly gender-neutral workplace processes, such as small, unobtrusive assessment tasks, may translate into macro patterns of workplace inequality. Specifically, this assessment task is an example of just one of the many low-promotability workplace tasks employees and supervisors are asked to do above and beyond their core workplace duties. As a rough estimate, if workers do up to 78 of these low-promotability tasks per year, as suggested by one quantitative study (Guarino and Borden 2017), small differences in time spent on, and approach taken with, each task, such as the differences we found here, may give the illusion of near equality, while masking pernicious differences that compound to impose significant extra burdens on women. Workplace assessment is just one of the many low-promotability tasks employees engage in, but it is similar to many other tasks in the imposition on supervisors' time and mental capacity, and the effect on other people in workplaces. Promotion (and tenure) committees, developing training programs, diversity and inclusion committees, and faculty governance all have similar dynamics as workplace assessment: they are considered low-promotability tasks, yet doing these tasks well has implications for workplace inequality more broadly.

Finally, our approach demonstrates the benefits of using passively-collected data produced through real, applied tasks carried out in the workplace, and the affordances of combining rigorous qualitative, computational, and quantitative techniques. The large amount of data produced and recorded in the process of workers carrying out their everyday tasks can provide valuable insights into behaviors that may contribute to workplace inequality (see, e.g., Correll et al. 2020). Unlike self-reports and interview data, these data directly capture behaviors, providing a different (and arguably more accurate) measure of workplace practices. Additionally,

these data tend to be unstructured and complex but also rich, necessitating a strategic bricolage of both qualitative and computer-assisted methods to fully analyze (Nelson 2021). In this article, we modeled one way to combine multiple techniques to analyze passively-collected complex behavioral data to extract insights into an important everyday workplace task.

Practical Implications

Based on our findings, we suggest three practical implications for how we might achieve more equity in work. First, solutions to the workplace assessment gap ought to address not just the amount of time spent on assessment, but the way this work is carried out. Simply equalizing the number of hours spent on each task ignores the extra emotional and cognitive burden of helping and reassuring struggling employees, subordinates, and students, and providing suggestions for how to improve programs to meet the needs of these employees (which in the case of medicine, also likely improves patients' experiences).

Second, when proposing solutions to the workplace assessment gap, we do not want to unwittingly shift the burden of gender inequality onto a different population. Non-promotable tasks may not be necessary for a workplace to function. Low-promotability tasks, such as workplace assessment, on the other hand, do often directly benefit individual careers and organizational cultures and equity. If women are indeed providing more effective feedback as our evidence suggests, simply coaching women to do less of this work—the popular "just say no" solution to gaps in low-promotability work more generally (e.g., Babcock et al. 2022; Bernstein 2017; Kara 2016; Bray, McLaren, and Ocampo 2020)—may disproportionately affect vulnerable employees and those who are struggling, potentially exacerbating inequalities at the training level. Rather than the "just say no" solution, or solutions aimed solely at equalizing the number of hours spent on low-promotability tasks, our findings reaffirm the argument that everyone needs training on how to do workplace assessment well, and on how to prevent implicit bias from slipping into evaluations and feedback.

Third, one of the most striking inequalities we found in our data was its enormous skew, with a few attendings shouldering the vast majority of feedback labor; many women did not provide great feedback, and a few men and women provided an enormous amount of consistently high-quality feedback. Solutions and training should be geared toward distributing the amount of (effective) assessment labor more evenly throughout the population, without leaving crucial support gaps for early-career employees and students. This could be achieved, for example, by more clearly specifying what is expected for supervisors and managers providing feedback to others and to institutions. Providing simple guidelines or information related to specific workplace assessment tasks, such as "most people spend this amount of time on this assessment task," or "attendings tend to write an average of three sentences and reference two concrete medical tasks in each comment," may be more effective at addressing who provides feedback and how than simply focusing on the amount of time spent or number of tasks completed. Solutions aimed at the incentive structure—rewarding or requiring workplace assessment and other types of service work for promotions—should reward not just the amount of work done, but the quality of that work. Future research could analyze the effectiveness of these different types of interventions in producing more equitable workloads overall, in both quantity and quality.

Limitations and Future Directions

Of course, our study is not without limitations. Our research focused on one occupational setting and one workplace task. Further research could compare other settings and additional low-promotability work tasks. Moreover, while our data provided a unique window into in-the-moment written evaluations in real workplaces, they may not have

captured the full extent of assessment and feedback taking place. In most educational settings, attendings typically offer residents both written and verbal, face-to-face feedback on their performances. Our data did not capture gender inequalities in verbal assessments. Future research is needed to examine whether the quantity and quality of feedback that attendings offer residents in-person varies by attending and resident gender. Digital trace data, like time-use surveys, also do not capture the "why" behind behavior. We do not know, for example, if women wrote more words because they cared more about the resident, because they did not want to disappoint their own supervisors, or because they were potentially using these evaluations to try to demonstrate their own managerial skills. Existing research has examined some reasons why women spend more time on low-promotability tasks; additional interview-based research could better elucidate reasons why employees choose to spend time on different workplace tasks.

Our dataset had many strengths, but we were not able to verify whether the negative comments and lower scores attendings gave residents may have reflected poorer performances in the ED or attendings' biased perceptions of their clinical abilities. Future research might be able to examine gender disparities in in-person assessment via ethnographic methods. In addition, while we knew the gender of the attendings and residents in our data, we did not have access to other demographic data, including race, age, national origin, or citizenship status. Further research is needed to better understand how other demographic markers affect the distribution of workplace tasks, particularly the effect of intersecting identities (Tiako et al. 2021; cf. Miller and Roksa 2020). Finally, the task we examined is just one of the many assessment tasks physicians are asked to do every day. Further research is needed to understand the full amount of time spent on workplace assessment, in the case of academic medicine specifically but also across industry and academia more broadly.

Limitations aside, our research brought rich yet broad behavioral data to bear on understanding the quantity and quality of workplace assessment, addressing the question of who shoulders the burden of effective workplace evaluation. In doing so, we empirically demonstrated how micro-level engagement with unobtrusive and seemingly gender-neutral workplace tasks can still create (largely invisible) gendered outcomes and contribute to workplace inequality. Solutions to gendered differences in those providing assessment, we conclude, should address not only who performs this work but who performs it well, and should work toward distributing the amount and quality of assessment work more evenly throughout the population in a way that does not leave crucial support gaps for learners and early-career employees.

APPENDIX: DETAILED SUPERVISED MACHINE LEARNING ACCURACY MEASURES

Table A1. Precision, Recall, and F1 Score by Comment Code Category for the Final Supervised Machine Learning Model, by Gender of the Attending and for All

	All			Men Attendings			Women Attendings		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
Error	.61	.91	.73	.61	.89	.72	.61	.95	.75
Helpful	.58	.89	.70	.54	.86	.67	.61	.92	.74
Reassuring	.55	.94	.69	.54	.94	.69	.56	.94	.70

Table A2. Confusion Matrix for the Final Supervised Machine Learning Model, by Gender of the Attending and for All

All

Error			Helpful			Reassuring		
	no	yes		no	yes		no	yes
no	3348 (74%)	425 (9%)	no	523 (48%)	224 (20%)	no	442 (40%)	284 (26%)
yes	63 (1%)	671 (15%)	yes	38 (3%)	311 (28%)	yes	22 (2%)	348 (32%)

Women Attendings

Error			Helpful			Reassuring		
	no	yes		no	yes		no	yes
no	1047 (67%)	196 (12%)	no	201 (39%)	113 (22%)	no	185 (36%)	138 (27%)
yes	18 (1%)	313 (20%)	yes	16 (3%)	179 (35%)	yes	11 (2%)	175 (34%)

Men Attendings

Error			Helpful			Reassuring		
	no	yes		no	yes		no	yes
no	2274 (78%)	229 (8%)	no	322 (55%)	111 (19%)	no	257 (44%)	146 (25%)
yes	45 (2%)	358 (12%)	yes	22 (4%)	132 (22%)	yes	11 (2%)	173 (29%)

Source: Comments collected using InstantEval (V2.0 Monte Carlo Software LLC, Annandale, VA), across eight U.S. hospitals, 2013 to 2015.
Note: Three comment code categories: whether the comment was in response to an error (*error*), whether the comment was likely helpful (*helpful*), and whether the comment provided reassurance (*reassuring).* F1 is the harmonic mean of precision and recall.

Editors' Note

To avoid any possible conflict of interest, the *ASR* Editors were not involved in the evaluation of this paper. The entire review process was handled by a Deputy Editor who is not affiliated with Indiana University.

Acknowledgments

The authors extend deep gratitude to Chandra Muller and Jen Schradie for their helpful comments on drafts of this manuscript, and Melissa Osborne, Rebecca Ewert, Tania Jenkins, Miriam Midoun, and Emily Tcheng for their helpful assistance with early stages of data coding.

Funding

This project was supported by funding from the University of Chicago Diversity Small Grant (awarded to Vineet M. Arora) and a University of Chicago Gianinno Faculty Research Award (awarded to Anna S. Mueller).

ORCID iDs

Laura K. Nelson https://orcid.org/0000-0001-8948-300X
Alexandra Brewer https://orcid.org/0000-0003-1910-2739
Anna S. Mueller https://orcid.org/0000-0002-3220-8944
Daniel M. O'Connor https://orcid.org/0000-0001-5464-2031

Notes

1. MedicineNet (http://www.medicinenet.com [accessed November 6, 2018]) is owned by the WebMD Consumer Network.
2. To create the dictionary we combined two lists: the Medical Dictionary List (https://www.medicinenet.com/script/main/alphaidx.asp?p=a_dict [accessed September 26, 2018]) and the List of Procedures and Tests (https://www.medicinenet.com/procedures_and_tests/alpha_a.htm [accessed September 7, 2018]). We removed duplicate entries, separated acronyms from the full phrase, and re-ordered comma separated entries.
3. To keep the hand-coding process manageable, only comments written in response to an error were hand-coded with the helpful and reassuring codes. We then used the trained machine learning algorithm to predict whether all comments were helpful and/or reassuring, whether or not they were in response to an error.
4. We provide the precision and recall measures by gender of attending and the full confusion matrices in the Appendix.
5. Calculated by taking the mean word count per comment and multiplying by 21.

References

Accreditation Council for Graduate Medical Education (ACGM) and American Board of Emergency Medicine. 2015. "The Emergency Medicine Milestone Project" (https://www.acgme.org/Portals/0/PDFs/Milestones/EmergencyMedicineMilestones.pdf).

Acker, Joan. 1990. "Hierarchies, Jobs, Bodies: A Theory of Gendered Organizations." *Gender & Society* 4(2):139–58.

Acker, Sandra, and Carmen Armenti. 2004. "Sleepless in Academia." *Gender and Education* 16(1):3–24.

Armijo, Priscila Rodrigues, Julie K. Silver, Allison R. Larson, Philomena Asante, and Sash Shillcutt. 2021. "Citizenship Tasks and Women Physicians: Additional Woman Tax in Academic Medicine?" *Journal of Women's Health* 30(7):935–43 (http://doi.org/10.1089/jwh.2020.8482).

Association of American Medical Colleges. 2019. "Diversity in Medicine: Facts and Figures 2019" (https://www.aamc.org/data-reports/workforce/report/diversity-medicine-facts-and-figures-2019).

Babcock, Linda, Brenda Peyser, Lise Vesterlund, and Laurie R. Weingart. 2022. *The No Club: Putting a Stop to Women's Dead-End Work*. New York: Simon & Schuster.

Babcock, Linda, Maria P. Recalde, Lise Vesterlund, and Laurie Weingart. 2017. "Gender Differences in Accepting and Receiving Requests for Tasks with Low Promotability." *American Economic Review* 107(3):714–47.

Beasley, Brent W., Stephen D. Simon, and Scott M. Wright. 2006. "A Time to Be Promoted: The Prospective Study of Promotion in Academia." *Journal of General Internal Medicine* 21(2):123–29.

Bellas, Marcia L. 1999. "Emotional Labor in Academia: The Case of Professors." *Annals of the American Academy of Political and Social Science* 561:96–110.

Bennett, Christopher L., Ali S. Raja, Neena Kapoor, Dara Kass, Daniel M. Blumenthal, Nate Gross, and Angela M. Mills. 2019. "Gender Differences in Faculty Rank among Academic Emergency Physicians in the United States." *Academic Emergency Medicine* 26(3):281–5.

Bernstein, Robin. 2017. "The Art of 'No.'" *The Chronicle of Higher Education*, March 19 (https://www.chronicle.com/article/The-Art-of-No-/239508).

Bosk, Charles L. [1979] 2003. *Forgive and Remember: Managing Medical Failure*, 2nd ed. Chicago: University of Chicago Press.

Bowles, Hannah Riley, and Linda Babcock. 2013. "How Can Women Escape the Compensation Negotiation Dilemma? Relational Accounts Are One Answer." *Psychology of Women Quarterly* 37(1):80–96.

Brandon Hall Group. 2016. "Performance Management 2016: People Over Process." (http://go.brandonhall.com/l/8262/2016-07-27/5qkggw).

Bray, Sarah, Rachel McLaren, and Anthony Ocampo. 2020. "How to Stop Yourself From Being Overworked and Bitter." *Inside Higher Ed* (https://www

.insidehighered.com/advice/2020/02/05/avoiding-disgruntlement-and-burnout-too-much-service-work-opinion).

Brewer, Alexandra, Laura K. Nelson, Rebecca Ewert, Anna S. Mueller, Arjun Dayal, Daniel M. O'Connor, and Vineet M. Arora. Forthcoming. "Gender and Inconsistent Evaluations: A Mixed Methods Analysis of Feedback for Emergency Medicine Residents" *The Western Journal of Emergency Medicine*.

Brewer, Alexandra, Melissa Osborne, Anna S. Mueller, Daniel M. O'Connor, Arjun Dayal, and Vineet M. Arora. 2020. "Who Gets the Benefit of the Doubt? Performance Evaluations, Medical Errors, and the Production of Gender Inequality in Emergency Medical Education." *American Sociological Review* 85(2):247–70.

Buckingham, Marcus, and Ashley Goodall. 2015. "Reinventing Performance Management." *Harvard Business Review* (https://hbr.org/2015/04/reinventing-performance-management).

Cappelli, Peter, and Anna Tavis. 2016. "The Performance Management Revolution." *Harvard Business Review* (https://hbr.org/2016/10/the-performance-management-revolution).

Caruana, Rich, and Alexandru Niculescu-Mizil. 2006. "An Empirical Comparison of Supervised Learning Algorithms." Pp. 161–68 in *Proceedings of the 23rd International Conference on Machine Learning*. ICML '06. New York: ACM.

Castilla, Emilio J. 2008. "Gender, Race, and Meritocracy in Organizational Careers." *American Journal of Sociology* 113(6):1479–526.

Chopra, Vineet, Vineet M. Arora, and Sanjay Saint. 2018. "Will You Be My Mentor? Four Archetypes to Help Mentees Succeed in Academic Medicine." *JAMA Internal Medicine* 178(2):175–76.

Correll, Shelley J., and Caroline Simard. 2016. "Vague Feedback Is Holding Women Back." *Harvard Business Review*, April 29.

Correll, Shelley J., Katherine R. Weisshaar, Alison T. Wynn, and JoAnne Delfino Wehner. 2020. "Inside the Black Box of Organizational Life: The Gendered Language of Performance Assessment." *American Sociological Review* 85(6):1022–50.

Dayal, Arjun, Daniel M. O'Connor, Usama Qadri, and Vineet M. Arora. 2017. "Comparison of Male vs. Female Resident Milestone Evaluations by Faculty during Emergency Medicine Residency Training." *JAMA Internal Medicine* 177(5):651–57.

Eagan, M. Kevin Jr., and Jason C. Garvey. 2015. "Stressing Out: Connecting Race, Gender, and Stress with Faculty Productivity." *The Journal of Higher Education* 86(6):923–54.

Ecklund, Elaine, and Anne E. Lincoln. 2016. *Failing Families, Failing Science: Work-Family Conflict in Academic Science*. New York: New York University Press.

Efron, Bradley. 1982. "The Jackknife, the Bootstrap, and Other Resampling Plans." Philadelphia, PA: Society for Industrial and Applied Mathematics.

El-Alayli, Amani, Ashley A. Hansen-Brown, and Michelle Ceynar. 2018. "Dancing Backwards in High Heels: Female Professors Experience More Work Demands and Special Favor Requests, Particularly from Academically Entitled Students." *Sex Roles* 79(3):136–50.

Ellemers, Naomi, Henriette Van den Heuvel, Dick de Gilder, Anne Maass, and Alessandra Bonvini. 2004. "The Underrepresentation of Women in Science: Differential Commitment or the Queen Bee Syndrome?" *British Journal of Social Psychology* 43(3):315–38.

England, Paula. 2010. "The Gender Revolution: Uneven and Stalled." *Gender & Society* 24(2):149–66.

Guarino, Cassandra M., and Victor M. H. Borden. 2017. "Faculty Service Loads and Gender: Are Women Taking Care of the Academic Family?" *Research in Higher Education* 58(6):672–94.

Gupta, Kiran, Sara G. Murray, Urmimala Sarkar, Michelle Mourad, and Julia Adler-Milstein. 2019. "Differences in Ambulatory EHR Use Patterns for Male vs. Female Physicians." *NEJM Catalyst* 5(6).

Hart, Jennifer L., and Christine M. Cress. 2008. "Are Women Faculty Just 'Worrywarts'? Accounting for Gender Differences in Self-Reported Stress." *Journal of Human Behavior in the Social Environment* 17(1–2):175–93.

Jerolmack, Colin, and Shamus Khan. 2014. "Talk Is Cheap: Ethnography and the Attitudinal Fallacy." *Sociological Methods & Research* 43(2):178–209.

Kara, Helen. 2016. "Do Yourself a Favour – Learn to Say 'No.'" *Times Higher Education*, June 3 (https://www.timeshighereducation.com/blog/do-yourself-favour-learn-say-no).

King, Eden B., Whitney Botsford, Michelle R. Hebi, Stephanie Kazama, Jeremy F. Dawson, and Andrew Perkins. 2012. "Benevolent Sexism at Work: Gender Differences in the Distribution of Challenging Developmental Experiences." *Journal of Management* 38(6):1835–66.

MacNell, Lillian, Adam Driscoll, and Andrea N. Hunt. 2015. "What's in a Name: Exposing Gender Bias in Student Ratings of Teaching." *Innovative Higher Education* 40(4):291–303.

Marcotte, Leah M., Vineet M. Arora, and Ishani Ganguli. 2021. "Toward Gender Equity in Academic Promotions." *JAMA Internal Medicine* 181(9):1155–56.

Mayer, Anita P., Janis E. Blair, Marcia G. Ko, Sharonne N. Hayes, Yu-Hui H. Chang, Suzanne L. Caubet, and Julia A. Files. 2014. "Gender Distribution of U.S. Medical School Faculty by Academic Track Type." *Academic Medicine: Journal of the Association of American Medical Colleges* 89(2):312–17.

Meier, Angela, Jenny Yang, Jinyuan Liu, Jeremy R. Beitler, Xin M. Tu, Robert L. Owens, Radhika L. Sundararajan, Atul Malhotra, and Rebecca E. Sell. 2019. "Female Physician Leadership during Cardiopulmonary Resuscitation Is Associated with Improved Patient Outcomes." *Critical Care Medicine* 47(1):8–13.

Miller, Candace, and Josipa Roksa. 2020. "Balancing Research and Service in Academia: Gender, Race, and Laboratory Tasks." *Gender & Society* 34(1):131–52.

Misra, Joya, Alexandra Kuvaeva, KerryAnn O'Meara, Dawn Kiyoe Culpepper, and Audrey Jaeger. 2021. "Gendered and Racialized Perceptions of Faculty Workloads." *Gender & Society* 35(3):358–94.

Mitchell, Sara McLaughlin, and Vicki L. Hesli. 2013. "Women Don't Ask? Women Don't Say No? Bargaining and Service in the Political Science Profession." *PS: Political Science & Politics* 46(2):355–69.

Mueller, Anna S., Tania M. Jenkins, Melissa Osborne, Arjun Dayal, Daniel M. O'Connor, and Vineet M. Arora. 2017. "Gender Differences in Attending Physicians' Feedback to Residents: A Qualitative Analysis." *Journal of Graduate Medical Education* 9(5):577–85.

National Center for Education Statistics, IPEDS Data Center. 2018. "Full-Time Instructional Staff, by Faculty and Tenure Status, Academic Rank, Race/Ethnicity, and Gender (Degree-Granting Institutions): Fall 2018." Fall Staff 2018 Survey.

Nelson, Laura K. 2021. "Cycles of Conflict, a Century of Continuity: The Impact of Persistent Place-Based Political Logics on Social Movement Strategy." *American Journal of Sociology* 127(1):1–59.

O'Meara, KerryAnn, Alexandra Kuvaeva, Gudrun Nyunt, Chelsea Waugaman, and Rose Jackson. 2017. "Asked More Often: Gender Differences in Faculty Workload in Research Universities and the Work Interactions That Shape Them." *American Educational Research Journal* 54(6):1154–86.

Owens, Jayanti. 2022. "Double Jeopardy: Teacher Biases, Racialized Organizations, and the Production of Racial/Ethnic Disparities in School Discipline." *American Sociological Review* 87(6):1007–48.

Ridgeway, Cecilia L. 2011. *Framed by Gender: How Gender Inequality Persists in the Modern World*. New York: Oxford University Press.

Saldaña, Johnny. 2016. *The Coding Manual for Qualitative Researchers*, 3rd ed. Los Angeles: Sage Publications.

Salganik, Matthew J. 2019. *Bit by Bit: Social Research in the Digital Age*. Princeton, NJ: Princeton University Press.

Sambunjak, Dario, Sharon E. Straus, and Ana Marusić. 2006. "Mentoring in Academic Medicine: A Systematic Review." *JAMA* 296(9):1103–15.

Shayne, Julia. 2017. "Recognizing Emotional Labor in Academe." *Inside Higher Ed* (https://www.insidehighered.com/advice/2017/09/15/importance-recognizing-faculty-their-emotional-support-students-essay).

Shin, Taekjin. 2012. "The Gender Gap in Executive Compensation: The Role of Female Directors and Chief Executive Officers." *ANNALS of the American Academy of Political and Social Science* 639:258–78.

Tiako, Max Jordan Nguemeni, Eugenia C. South, and Victor Ray. 2021. "Medical Schools as Racialized Organizations: A Primer." *Annals of Internal Medicine* 174(8):1143–44.

Tsugawa, Yusuke, Anupam B. Jena, Jose F. Figueroa, E. John Orav, Daniel M. Blumenthal, and Ashish K. Jha. 2017. "Comparison of Hospital Mortality and Readmission Rates for Medicare Patients Treated by Male vs Female Physicians." *JAMA Internal Medicine* 177(2):206–13.

Warner, Judith, Nora Ellmann, and Diana Boesch. 2018. "The Women's Leadership Gap: Women's Leadership by the Numbers." Center for American Progress (https://cdn.americanprogress.org/content/uploads/2018/11/19121654/WomensLeadershipFactSheet.pdf).

Weisshaar, Katherine. 2017. "Publish *and* Perish? An Assessment of Gender Gaps in Promotion to Tenure in Academia." *Social Forces* 96(2):529–60.

Winslow, Sarah. 2010. "Gender Inequality and Time Allocations among Academic Faculty." *Gender and Society* 24(6):769–93.

Ziker, John, Kathryn Demps, David Nolin, and Matt Genuchi. 2013. "Time Allocation Workload Knowledge Study, Phase 1 Report" (https://www.researchgate.net/publication/308761975_Time_Allocation_Workload_Knowledge_Study_Phase_1_Report).

Laura K. Nelson is an Assistant Professor of sociology at the University of British Columbia. She uses computational methods to study social movements, culture, gender, and institutions, and she develops transparent and reproducible text analysis frameworks for combining computational and qualitative methods using open-source tools. She has given talks, lectures, and workshops on computational social science research and methods across the United States, Canada, and Europe. Her research has appeared in journals such as the *American Journal of Sociology, Sociological Methods & Research, Gender & Society*, and *Poetics*, among other outlets.

Alexandra Brewer is an Assistant Professor of sociology at the University of Southern California. Her research examines the reproduction of social inequalities in the U.S. healthcare system, focusing primarily on the everyday practices of healthcare workers. Her work has won several awards and has been published in the *American Sociological Review, Journal of Health and Social Behavior*, and *Social Science & Medicine*. She is currently writing a book about how organizational constraints in hospitals shape physicians' decisions about pain management.

Anna S. Mueller is the Luther Dana Waterman Associate Professor of Sociology at Indiana University Bloomington. Mueller's research agenda investigates (1) the production of social inequality in medicine, with a focus

on emergency and pediatric medicine; and (2) how social forces shape youth vulnerability to suicide. Her award-winning research has appeared in journals such as the *American Sociological Review, American Journal of Sociology*, and *Sociological Theory*. She is currently the PI of an NIH-funded project to improve suicide prevention in schools and the author of a forthcoming book titled *Life under Pressure: The Social Roots of Adolescent Suicide* (Oxford 2024).

Daniel M. O'Connor is a micrographic surgeon at Dermatology and Skin Health and dermatologist at the Mass General Brigham Wentworth-Douglass Hospital in Dover, New Hampshire. He completed residency and fellowship training at Harvard Medical School and medical school at the University of Pennsylvania. His research focuses on cutaneous oncology and how gender influences the evaluation of medical trainees.

Arjun Dayal is a board certified dermatologist at the Rush Copley Medical Group in Aurora, Illinois. He completed residency at the University of Chicago Medicine Dermatology Residency Training Program and medical school at the University of Chicago Pritzker School of Medicine, where he was awarded the Dean's Scholarship for Outstanding Promise in Medicine. His current research focuses on novel applications of technology in dermatology, and how gender influences the evaluation of medical trainees. His recent work has been published in *JAMA Internal Medicine, Journal of Graduate Medical Education*, and *The Journal of Neuroscience*.

Vineet M. Arora is the Herbert T. Abelson Professor of Medicine and Dean of Medical Education at the University of Chicago. Her scholarship on improving the learning environment and care delivered to patients in teaching hospitals has been cited over 10,000 times. She is a vocal advocate for advancing gender equity in healthcare, and is the PI of an NIH and foundation grants to improve equity and opportunity for female and minority aspiring physician scientists. She is an elected member of the National Academy of Medicine and published in journals such as *JAMA, Annals of Internal Medicine*, and *Academic Medicine.*