ARTICLE

<https://doi.org/10.1057/s41599-021-00926-3>

OPEN

The pregnancy drop: How teaching evaluations penalize pregnant faculty

Ronke M. Olabisi ¹✉

The “leaky pipeline” and the “maternal wall” have for decades described the loss of women in STEM and the barriers faced by working mothers. Of the studies examining the impact of motherhood or pregnancy on faculty in higher education, most focus on colleagues’ attitudes towards mothers; few studies explore pregnancy specifically, only a handful examine student evaluations in particular, and none include female faculty in engineering. This study is the first to compare student evaluations across fields from female faculty when they were pregnant against when they were not. Two scenarios were considered: (1) the lived experiences of faculty who taught classes while pregnant and while not pregnant and (2) an experiment in which students submitted teaching evaluations for an actress whom half the students believed was pregnant while the other half did not. Among faculty respondents, women of colour received lower scores while pregnant and these scores lowered further when women were in engineering and/or had severe symptoms. Depending on their demographics, students who participated in the experiment were awarded teaching evaluation scores that differed when they believed the instructor was pregnant. Findings suggest that in fields with fewer women, the maternal wall is amplified and there is a unique intersectional experience of it during pregnancy. These findings may be useful for Tenure and Promotion committees to understand and therefore account for pregnancy bias in teaching evaluations.

¹University of California, Irvine, CA, USA. ✉email: ronke.olabisi@uci.edu

Introduction

A number of studies have documented the negative impact of motherhood on the careers of women in science, technology, engineering, and mathematics (STEM). Williams et al. described the impact of motherhood as a no-win proposition on the careers of women in STEM (Williams et al., 2014, p. 5). Studies examining bias against mothers have found that when identical resumes were randomly presented as women with or without children, those with children were 79% less likely to be hired, offered an average of \$11,000 less in salary, half as likely to be promoted, and held to higher standards when it came to punctuality and performance (Correll et al., 2007; Williams et al., 2014). Studies have also confirmed that mothers walk a “tightrope” (Williams et al., 2014, p. 3), where they are assumed by their colleagues to be less competent and committed (Benard et al., 2007), and further, those who are irrefutably competent and committed are seen as bad mothers, and hence bad people (Benard and Correll, 2010; Williams et al., 2014, p. 28). Men with young children and women without children are, respectively, 35% and 33% more likely to secure tenure track positions than women with young children (Waxman and Ispa-Landa, 2016). In fact, having children was found to be *beneficial* to the careers of men while *detrimental* to the careers of women (Ginns et al., 2007; Marsh, 2007; Onwuegbuzie et al., 2007). This “baby penalty” or a “baby tax,” can have a more profound impact on pregnant women.

Pregnancy discrimination can appear in many forms. Pregnant job applicants experience greater interpersonal hostility, are less likely to be hired or promoted, and receive lower starting salary recommendations than nonpregnant women (Bragger et al., 2002; Hebl et al., 2007; Heilman and Okimoto, 2008; Morgan et al., 2013). Approximately 250,000 pregnant workers are denied requests for temporary accommodations each year, such as lifting lighter loads or not working with toxic chemicals, while others are outright fired for becoming pregnant (Ellmann and Frye, 2018; The Childbirth Connection, 2014). There is a clear pattern of the disproportionate impact of pregnancy discrimination on women of colour: although accounting for only 14.3% of the workforce, African American women filed 28.6% of the pregnancy discrimination charges with the U.S. Equal Employment Opportunity Commission (EEOC) (Ellmann and Frye, 2018; The Childbirth Connection, 2014). In addition to hostile discrimination, pregnant women are also often targets of benign discrimination. Without their consultation or consent, pregnant women have been demoted from prestigious roles within their organizations to roles with “fewer responsibilities and less prestige” in efforts “to be nice and to give [her] a desk job” (“Holland v. Gee,” 2012, pp. 13–14). The assumption is a diminished capacity to perform work due to physical inability and/or emotional instability due to pregnancy hormones, which differs from the bias against non-pregnant mothers. Although all women are subject to the stereotype that women are irrational and emotional, pregnancy intensifies the stereotypes (Ollilainen, 2019, p. 964). This underlying assumption of diminished capacity ignores the broad range of pregnant women’s experiences and capabilities. The “benign” actions to modify duties without any discussion infantilize pregnant women in efforts to care for them. These forms of discrimination often result in pressure to take reduced or diminished maternity leave (Kavya and Kramer, 2020).

The concept of intersectionality describes a framework for understanding how an individual’s racial, social, political—or any characteristic that places them in a minoritized group—intersect and combine to create experiences of discrimination that differ from what might be assumed by an overlap (Collins, 2000; Crenshaw, 1990; Davis, 2014). For instance, Black women’s experiences are not merely the sum of those of white women and

Black men. Thus, the experiences of working pregnant women might be described as intersectional in that some experiences are shared with those of non-pregnant mothers while other experiences are wholly unique to being pregnant. The term “the maternal wall” has been used to describe discrimination against mothers since 2003 (Williams and Segal, 2003), yet despite its profound effect on women at a pivotal point in their careers, only recently has the differential impact of motherhood vs. pregnancy on women in academia been explored (Ollilainen, 2019). Of the studies that explored pregnancy specifically, they did not include women in engineering nor the earth or physical sciences, fields with fewer proportions of women.

Personal reflections from women in STEM describe the negative impact of pregnancy on how women are perceived by colleagues and students, particularly on students’ teaching evaluations. Student evaluations of teaching are used at the majority of U.S. higher education institutions, have become the primary source of information in the evaluation of faculty teaching, and are given more weight than classroom visits or exam scores (Miller and Seldin, 2014; Seldin, 1998; Stroebe, 2016). Teaching evaluations have a substantial influence on hiring, firing, reappointment, tenure, promotion, post-tenure review, and salary raise decisions (Schimanski and Alperin, 2018; Uttl and Smibert, 2017). Many institutions follow the same general process for review, promotion, and tenure procedures. Months in advance, faculty candidates submit information, which includes publications, grants, and teaching evaluations into their dossier. These materials are then evaluated by multiple faculty external to the candidate’s institution. These evaluations are added to the candidate’s dossier and in many institutions candidates never see those evaluations, nor know who wrote them nor what they contain (Strunk, 2020). After a faculty vote by the candidate’s department, the faculty write a collective letter and the department chair writes an individual one. The candidate then writes a response to this letter and the dean adds another letter, all of which enter into the candidate’s dossier. A university committee receives this dossier, votes on whether to promote or award tenure and the candidate is eventually notified of the outcome. At each step along the way, bias in teaching evaluations could colour the perceptions of evaluators and lead to cumulative disadvantage.

Proponents of using teaching evaluations as a metric acknowledge their imperfections while arguing their usefulness as a record of instructor progress or lack thereof (Marsh, 2007; Wang and Gonzalez, 2020). Opponents of teaching evaluations argue that they measure student bias rather than teaching effectiveness or student learning (Bavishi et al., 2010; Boring, 2017; Boring et al., 2016; Hornstein, 2017). This bias differentially impacts STEM faculty—professors teaching quantitative classes (e.g., math vs. English) were significantly more likely to receive lower teaching evaluation ratings and were far more likely not to receive tenure, promotion, and/or merit pay when their performance was evaluated against common standards than professors teaching qualitative classes (Uttl and Smibert, 2017). Bias can impact a range of critical factors that can determine the success of female faculty in STEM, many of which have nothing to do with teaching. For instance, to be successful, faculty in STEM must obtain funding to support their research and publish the results of that research; however, gender bias has been demonstrated in both granting agencies and the publication process (Chawla, 2018; Hengel, 2017; Witteman et al., 2019). Faculty with labs must attract undergraduates, graduate students, and postdoctoral trainees, who perform the bulk of the experiments needed to produce research results. Teaching evaluations do not capture invisible forces such as bias in the peer-review process when applying for grants or submitting manuscripts, nor whether graduate students are choosing to join or forego a lab because the

principal investigator is pregnant. For instance, qualitative comments from the survey revealed some engineering faculty received statements from students indicating that their pregnancy influenced their decisions to work with them. One respondent described an encounter in which a male colleague expressed surprise at her second pregnancy because he thought she was serious about her work. Such sentiments may reflect which colleagues are less likely to recommend students join the labs of pregnant women. As one surveyed engineering junior faculty member noted, in making the decision to stay or leave her lab, a student stated that her impending maternity leave was the deciding factor, then proceeded to join the lab of a male colleague who was taking sabbatical over the same time period as her maternity leave.

Despite not capturing such issues, a benefit of using teaching evaluation data in this study is that they reveal student perceptions of faculty who are (impending) mothers. Research describing the maternal wall largely focuses on the perceptions of colleagues (Williams et al., 2014; Williams and Segal, 2003). Part of the reason maternal bias among students is not as fully described may arise from the fact that colleagues are more likely than students to know which women faculty are mothers. Although maternal status may or may not be shared with students depending on the preference of the instructor, pregnancy status may reveal itself. Despite evidence that teaching evaluations are biased against all women, and women of colour, in particular, they continue to be used in many colleges and universities for tenure and promotion across all faculty positions. Hence, it is important to ascertain whether pregnant women face additional hurdles in tenure track positions to counteract the impact of such biases.

The consistency of student evaluation data enables comparison of evaluations across humanities and STEM fields to dissect whether there is bias against pregnant faculty in general or just those in certain fields, to uncover whether this bias can be predicted by student attributes, and to establish whether such bias mirrors that observed in pregnancy discrimination data, where women of colour are differentially impacted. Although researchers have used student evaluations of teaching to demonstrate bias against people with accents (Rubin and Smith, 1990; Subtirelu, 2015), faculty in quantitative fields (Uttl and Smibert, 2017), and gender, racial, and ethnic minorities (Boring, 2017; Gutiérrez y Muhs et al., 2012; Lazos, 2012; Wang and Gonzalez, 2020), most were between groups comparisons, e.g., math vs. English, Black women vs. white men. One experiment that compared instructors against themselves showed that male instructors received lower evaluation scores when students thought they were female, while female instructors received higher scores if students believed them to be male (Boring et al., 2016). Excepting interventional studies, few reports have compared in-person teaching evaluations of instructors against themselves. Using teaching evaluations as a measure herein is innovative because it allows the same woman to be compared against herself without any intervention. This exploratory study examines teaching evaluation scores for identical women when pregnant or not in the form of (1) self-reported evaluation scores from women when pregnant and when not pregnant and (2) evaluations by students participating in an experiment during which students watched an instructional video where half the students believed an instructor to be pregnant while the other half did not.

Methods

Study design. All surveys were conducted after obtaining informed consent and in accordance with protocols approved by the Institutional Review Board (IRB) of Rutgers University,¹ which provides ethical review for all human subjects research.

There were two different situations considered: (1) the lived experiences of faculty who taught classes while pregnant and while not pregnant and (2) a simulation in which students submitted teaching evaluations for an actress whom half the students believed was pregnant while the other half did not. Data from these two separate situations were then considered together to capture whether the intersecting identities of the faculty or student respondents led to harsher evaluation scores. This study employed a quantitative design that included a quantitative survey (see Supplementary information) in the form of questionnaires (for faculty) and abbreviated student evaluation forms (for students). Faculty online matrix questionnaires were employed with 32 questions ranging from 1 to 5 as strongly disagree (1) or strongly agree (5). Faculty respondents were asked about their perceptions of student treatment, student characteristics, student evaluation scores, as well as their own characteristics and pregnancy symptoms. Student subjects were shown a video recording of a short lesson on a topic for which they had limited experience and then were asked to complete an abbreviated 7-question student evaluation form to rate instructor effectiveness with or without information indicating that the instructor was pregnant.

Recruitment. Convenience sampling was used to recruit faculty participants through emails and social media groups targeted to women in academia. Survey respondents were included in the study if they taught a university-level class while pregnant and while not pregnant and if they received student evaluations during both of these times; graduate students were not excluded. For the experimental portion of the study, biomedical engineering students were recruited by the announcement of an experiment entitled: “Five dollars for five minutes.” In the last 10 min of a core junior biomedical engineering class, a faculty instructor offered students the opportunity to participate in a study to receive \$5 for 5 min of their time, after which the faculty instructor distributed consent forms and evaluations, discussed the consent form, made it clear participation was completely voluntary and anonymous, and that students could leave the room at any time, and then the faculty instructor left the room to avoid any undue pressure and the experiment was proctored by a graduate student unrelated to the class. The graduate student played the video, collected consent forms and evaluations, then distributed the money.

Data collection. One hundred and three faculty responded to the online anonymous survey. Respondents were excluded if they did not have student evaluations for both pregnancy and non-pregnancy. Respondents were also excluded if surveys were incomplete. After exclusion, 50 surveys were included for analysis and 53 survey respondents were dropped.

After consenting, to receive compensation, students were instructed to remain silent during the study, and to rewrite the following statements to ensure they had read them: *this study is to determine whether students rate (pregnant) faculty the same, more harshly, or more leniently when teaching in a video compared to when teaching in person.*² The two surveys were distributed: one with the word *pregnant* and one with the word omitted. Surveys were alternated between male and female students to ensure equal numbers of male and female students received each version. Surveys were not alternated based on other demographics such as race or ethnicity and were randomly distributed along these demographics. To avoid any prior student encounter with existing faculty, an actress was chosen to teach a unit on a topic for which students were novices. To prevent artefact due to potential actor mistakes during presentation, the unit was pre-

recorded and evaluated by instructors familiar with the subject matter. The video was edited to 5 min; the actress happened to be African American. After copying the statements, students were shown the video. Afterwards, they were instructed to fill out a short survey (supplementary information) modelled after the teaching evaluations they are regularly distributed. Surveys were anonymous and were collected in an envelope by the graduate student. There were 83 student surveys collected, and all were complete and included in the analysis.

Statistics. Statistical data analysis was performed using the Statistical Package for Social Sciences (SPSS) statistical software (version: 28.0.0.0 (190)). Pairwise comparisons were conducted with one-tailed Student's *t*-tests (paired samples for faculty, independent samples for students). The independent correlation of various risk stratifiers to lowered evaluations in faculty data was determined by means of logistic regression analysis with

change in course quality and teaching evaluation scores as the dependent variables. A binary generalized linear logistic model for main effects and interaction by means of a stepwise analysis was used. Faculty were analysed as a combined group, then stratified for race and symptoms. Significance was reported for $p < 0.05$.

Results

Of the 83 participating students 31% were white, 46% Asian, 12% underrepresented minority, 8% Middle Eastern, and 3% other, and 52% were female while 48% were male (Fig. 1). Faculty respondents were from humanities (6%), medicine (12%), engineering (32%) and sciences (50%), with the sciences further broken down into the life, physical, and earth sciences (22%) and the social sciences (22%) not including education (6%), which was separated into a distinct category. Of the 50 faculty in the study, women reported that when pregnant they

Fig. 2 The mean rating scores as reported by faculty of course quality and instructor effectiveness, separated by general field, student knowledge of pregnancy, student knowledge of the illness, the severity of symptoms, weight gain, and instructor race. Error bars show standard error and asterisks indicate statistical significance ($p < 0.05$).

Fig. 1 Breakdown of student participants in the video simulation and faculty respondents to the survey of lived experiences included in the study. From left to right, student demographics ($n = 83$), student gender, faculty field ($n = 6$), faculty race/ethnicity ($n = 50$).

Table 1 Stratified binary logistic regression results for women in Engineering or Education.

Group	B	SE B	Wald χ^2	p	OR	95% CI OR
All Women in Engineering or Education	1.815	0.7257	6.254	0.012	6.140	[1.481, 25.465]
White Women in Engineering	1.966	0.7365	7.123	0.008	7.139	[1.805, 34.026]
Women of Colour in Education or Engineering	1.144	0.9019	1.610	0.204	3.141	[0.554, 20.757]
White Women in Engineering + weight gain or severe symptoms	3.280	1.3761	5.681	0.017	26.577	[2.600, 863.999]
White Women in Engineering + weight gain or severe symptoms	1.130	0.8678	1.695	0.193	3.095	[0.585, 19.087]
Women of Colour in Engineering or Education + weight gain or severe symptoms	2.853	1.2885	4.901	0.027	17.333	[1.802, 417.535]

All groups are compared against white women in non-engineering or education fields. For All Women in Engineering or Education, the first row denotes statistics for the teaching effectiveness scores and the second row denotes statistics for the course quality scores. For all other groups, these statistics were identical.
SE B standard error of B, OR odds ratio, CI confidence interval.

were postdocs (6%), graduate students (10%), Non-tenure track faculty (16%), Assistant Professors (54%), or Associate Professors (14%).

Survey of lived experiences. Across all fields the instructors' teaching effectiveness ratings went down (from a nonpregnant score of 4.29 ± 0.06 to 4.14 ± 0.12 when pregnant; paired $t(47) = 1.7$, $p = 0.047$) for all women when they were pregnant (Fig. 2). Instructor effectiveness averages dropped further if women had severe symptoms (from nonpregnant 4.50 ± 0.09 to 3.9 ± 0.42 when pregnant; paired $t(9) = 1.8$, $p = 0.051$) or gained more than 15 lbs (from nonpregnant 4.38 ± 0.26 to 4.11 ± 0.78 when pregnant; paired $t(39) = 2.2$, $p = 0.018$). When disaggregating the data by broad field, teaching effectiveness ratings dropped for women in STEM fields (from nonpregnant 4.34 ± 0.25 to 4.05 ± 0.72 when pregnant; paired $t(39) = 2.5$, $p = 0.009$), but not for women in humanities (from nonpregnant 4.75 ± 0.25 to 4.75 ± 0.25 when pregnant; standard error of the difference was 0) and not significantly for women in medicine (from nonpregnant 4.63 ± 0.13 to 4.58 ± 0.15 when pregnant; paired $t(5) = -0.4$, $p = 0.34$). To disaggregate pregnant women's teaching effectiveness ratings further by STEM discipline revealed a substantial drop for pregnant women in engineering (from nonpregnant 4.22 ± 0.29 to 3.84 ± 0.52 when pregnant; paired $t(15) = 1.7$, $p = 0.055$) and education (from nonpregnant 4.25 ± 0.13 to 2.50 ± 1.50 when pregnant; paired $t(2) = 1.4$, $p = 0.19$), but not for pregnant women in the life, physical, or earth sciences (from nonpregnant 4.45 ± 0.12 to 4.45 ± 0.12 when pregnant; paired $t(10) = 0$, $p = 0.5$). Comparing faculty demographics revealed that across all fields instructor effectiveness ratings dropped for women of colour (from nonpregnant 4.33 ± 0.35 to 3.78 ± 1.27 when pregnant; paired $t(17) = 2.4$, $p = 0.01$), but not for white women (from nonpregnant 4.41 ± 0.16 to 4.37 ± 0.19 when pregnant; paired $t(28) = 0.5$, $p = 0.31$).

Therefore, the engineering data was stratified by race to determine whether women of colour were driving the drop in teaching evaluation scores in engineering. To stratify the data by race, the engineering and education data were combined to increase statistical power. The stratified analysis by race revealed that when separated into women of colour (from nonpregnant 4.167 ± 0.46 to 2.83 ± 1.36 when pregnant; paired $t(5) = 2.3$, $p = 0.03$) and white women (from nonpregnant 4.25 ± 0.20 to 4.13 ± 0.18 when pregnant; paired $t(11) = 0.9$, $p = 0.19$), the drop in engineering teaching evaluation scores was statistically significant only when the instructor was a woman of colour. It is important to note that descriptive analysis of the data revealed that all respondents in the education field were women of colour. This field was the only social science field that demonstrated a

Fig. 3 The mean rating scores as reported by faculty of course quality and instructor effectiveness during pregnancy for white women and women of colour for all fields, for women in engineering, for women in engineering with severe symptoms, and for women in engineering with a high (over 30 pounds) weight change. Being in engineering causes women's scores to drop. Compounding being in engineering with severe symptoms or a high weight change causes a greater drop in evaluations for all women. This drop is larger for women of colour. Error bars show standard error and asterisks indicate statistical significance ($p < 0.05$).

drop in teaching effectiveness and course quality ratings due to pregnancy, as reported in Fig. 2.

Given the significant stratified results by race between pregnant women of colour and pregnant white women in teaching effectiveness scores, where pregnant women of colour had a significantly greater decrease in teaching effectiveness scores, a binary logistic regression (generalized linear model, GLM) was used to determine the specific factors that influenced teaching effectiveness scores for women of colour. GLM Logistic regression analysis revealed that for pregnant women of colour, there was an interaction between the severity of symptoms, the level of weight gain, and being in engineering or education (Table 1), which resulted in lower ratings in both course quality and teaching effectiveness (Fig. 3). When all races were considered, women in engineering or education had 6.14 greater odds of receiving lower teaching evaluation scores when pregnant than women in other fields; for course quality, their odds ratio was 7.139. When pregnant women in engineering were stratified by race, the analysis revealed that the increased odds of having lower evaluation scores were driven by women of colour. Pregnant white women in engineering or education did not have statistically greater odds of receiving lower teaching evaluation or course quality scores than pregnant white women in other fields.

Fig. 4 Comments left by students for pregnant faculty in their evaluations. Comments from engineering evaluations were more negative than other fields.

Conversely, pregnant women of colour in engineering or education had 26.577 greater odds of receiving lower teaching evaluation or course quality scores than pregnant white women in other fields. If a pregnant woman of colour in engineering also had severe symptoms or weight gain over 15 pounds, her odds ratio of receiving lower teaching evaluation or course quality scores was 17.333 times that of pregnant white women in other fields.

In short, across all fields being a woman of colour caused teaching evaluation scores to drop while pregnant. This pregnancy drop was worse if the woman of colour was in engineering or education, from nonpregnant 4.17 ± 0.47 to 2.83 ± 1.37 when pregnant ($p = 0.017$, Table 1). Finally, having severe symptoms (requiring missing work) and having a weight change of more than 15 pounds further lowered women's teaching evaluations to 2.33 ± 0.88 and 2.00 ± 0.58 ($p = 0.027$, Table 1), respectively. There was no effect due to other factors such as instructor age, instructor age when pregnant, how long ago the pregnancy happened, experience teaching (in years), faculty level, class size, whether instructors taught the same class before or after their pregnancy, semester or quarter taught (e.g., fall vs spring), or institution type (e.g., R1, R2, teaching, or medical).

Shifting from analysing instructor characteristics to analysing student characteristics reveals that certain student attributes affected whether pregnant instructors would receive drops in their scores. For instance, although teaching effectiveness scores dropped regardless of whether students knew their instructors were pregnant, the difference was only significant when students did know (from nonpregnant 4.38 ± 0.20 to 4.20 ± 0.59 when pregnant; paired $t(41) = 2.1$, $p = 0.02$). Though not significant, the drop in teaching effectiveness scores was larger when instructors believed students did not know (from nonpregnant 4.33 ± 0.46 to 3.75 ± 1.07 when pregnant; paired $t(5) = 1$, $p = 0.18$). Both male and female students awarded roughly 0.25 points less to women when they were pregnant, but the mean of the lowest scores awarded by female students (from nonpregnant 4.42 ± 0.23 to 4.16 ± 1.00 when pregnant; paired $t(18) = 1.6$, $p = 0.06$) were higher than the mean of the highest scores awarded by male students (from nonpregnant 4.14 ± 0.40 to 3.90 ± 0.59 when pregnant; paired $t(10) = 0.8$, $p = 0.23$), though these differences were not significant. In essence, although both male and female students penalized faculty similarly for pregnancy, the nonpregnant scores awarded by male students were lower than those awarded by female students, thus the penalty is more noticeable in classes with a greater male-to-female student ratio. In classes where there was an equal ratio of male and female students, the pregnancy drop in teaching evaluation scores was lowest (from nonpregnant 4.43 ± 0.10 to 4.30 ± 0.28 when pregnant; paired $t(14) = 1.2$, $p = 0.13$). First-year students gave lower scores and a greater pregnancy penalty

Fig. 5 The mean rating scores as reported by students of video quality and instructor effectiveness awarded to the actress instructor by student race, gender, and level of interest in the class. Error bars show standard error and asterisks indicate statistical significance ($p < 0.05$).

(from nonpregnant 4.32 ± 0.22 to 4.12 ± 0.39 when pregnant; paired $t(16) = 1.9$, $p = 0.03$) than graduate students (from nonpregnant 4.42 ± 0.27 to 4.29 ± 0.61 when pregnant; paired $t(11) = 0.8$, $p = 0.19$), but the largest drop in evaluation scores was in classes with a wide range in student ages. When classes had mixed levels of students (from sophomores to graduate students), the non-pregnant scores were among the highest and the pregnancy drop was largest (from nonpregnant 4.40 ± 0.23 to 4.10 ± 0.94 when pregnant; paired $t(19) = 1.4$, $p = 0.09$), though it was not significant.

Course quality ratings mirrored that of instructor effectiveness ratings, with the addition of student ignorance that the instructor was feeling unwell. Although the drops in course quality scores were similar regardless of whether the students knew the instructor was feeling unwell, the variation was lower when students did not know and thus the difference was significant when students were unaware (from nonpregnant 4.41 ± 0.22 to 4.17 ± 0.51 when pregnant; paired $t(38) = 2.6$, $p = 0.007$) the instructor was feeling unwell than when students were aware (from nonpregnant 4.39 ± 0.04 to 4.17 ± 0.75 when pregnant; paired $t(8) = 0.9$, $p = 0.18$).

When asked if anything in their evaluations surprised them, women reported comments left for them by students (Fig. 4). In non-engineering and non-education fields, these comments were overwhelmingly positive. In engineering and education, the comments were overwhelmingly negative, and several women noted that students reported them to Deans for perceived rude or disrespectful behaviour.

Student evaluations of teaching from video simulations. Scores awarded by students watching the actress instructor also varied by student characteristics, specifically, the race and gender of the student (Fig. 5). When students thought the instructor was pregnant, every group awarded *higher* instructor effectiveness scores, with the exception of students who had a low prior interest in the class (from nonpregnant 3.62 ± 0.59 to 3.16 ± 2.56 when pregnant; independent $t(6) = -0.65$, $p = 0.27$) and white male (from nonpregnant 3.40 ± 0.49 to 2.80 ± 1.20 when pregnant; independent $t(6) = -1.1$, $p = 0.31$) and Middle Eastern students (from nonpregnant 4.25 ± 0.25 to 2.67 ± 0.33 when pregnant; independent $t(4) = -3.8$, $p = 0.009$), who awarded lower scores when they believed the instructor was pregnant. The

higher scores were significant only for Asian (from nonpregnant 3.25 ± 0.62 to 3.78 ± 0.73 when pregnant; independent $t(36) = 2.0$, $p = 0.02$) and underrepresented minority students (from nonpregnant 3.33 ± 0.27 to 4.5 ± 0.33 when pregnant; independent $t(6) = 3.3$, $p = 0.008$), while the lower scores were only significant for Middle Eastern students.

For video quality scores, the gap narrowed for white male (from nonpregnant 3.4 ± 0.71 to 3.2 ± 1.70 when pregnant; independent $t(6) = -0.3$, $p = 0.38$), Asian (from nonpregnant 3.22 ± 0.65 to 3.55 ± 0.68 when pregnant; independent $t(36) = 1.2$, $p = 0.11$), and underrepresented minority students (from nonpregnant 3.67 ± 0.67 to 3.75 ± 0.92 when pregnant; independent $t(6) = 0.1$, $p = 0.44$), resulting in a lower penalty from white male students and a lower bonus from Asian and underrepresented minority students. Conversely, the gap increased for Middle Eastern students (from nonpregnant 4.00 ± 0.67 to 2.33 ± 0.33 when pregnant; independent $t(5) = -3.2$, $p = 0.01$), resulting in a greater penalty. For those with low interest in the class the scores mirrored their teaching effectiveness ratings and also went down (from nonpregnant 3.50 ± 0.72 to 3.2 ± 0.94 when pregnant; independent $t(16) = -0.6$, $p = 0.26$).

Discussion

These findings demonstrate that when women faculty teach while pregnant, their teaching evaluation scores drop, particularly for women of colour. This “pregnancy drop” grows with the level the pregnancy is affecting them, whether through greater weight gain or more severe symptoms. Further, the drop is influenced by the gender and ethnicity of the students rating them. Numerous studies have demonstrated that women, in general, receive lower teaching evaluation scores than their male counterparts (Boring, 2017; Marsh, 2007; Mengel et al., 2018), and these scores are further lowered for women of colour. In a quasi-experimental dataset comprising 19,952 teaching evaluations for which students had been randomly assigned male or female instructors, women, particularly junior female instructors, received systematically lower teaching evaluation scores than their male counterparts, and these lower evaluations were driven by male students (Mengel et al., 2018). In the present study, both effects were observed—lowered scores for women of colour compared to white women and lowered scores driven by male students. The lowered scores driven by male students were observed in actual evaluations reported by women faculty (Fig. 1), though in the simulation, they were only observed in evaluations awarded by a white male and Middle Eastern students—male students from other ethnicities awarded greater scores when they believed the instructor to be pregnant (Fig. 5). It is possible this pregnancy drop or bonus reflects student values. For instance, many white males and Middle Eastern students value stay at home mothers while many Latin and African American students value working mothers (Halpin and Teixeira, 2010; LeMaster et al., 2004; Moghadam, 2004; Orbuch and Custer, 1995; Pepin and Cotter, 2018).

A major cause for the lowered scores reported by faculty women of colour is more difficult to pinpoint. The theory of intersectionality helps to problematize these results. Women of colour already are the targets of greater bias when teaching and greater discrimination when pregnant, thus their intersecting identities of being people of colour, being women, and being pregnant compound in an intersectional dimension that amplifies the bias. For instance, pregnant women of colour from certain racial, ethnic, or economic backgrounds face biases stemming from stereotypes concerning the number of children they have (or should or should not have) (Ellmann and Frye, 2018). Pregnant Black women are more likely to be perceived as single mothers in need of public assistance than

pregnant white women (Rosenthal and Lobel, 2016). A Black respondent described an encounter in which a colleague expressed surprise, not at her pregnancy but that she had a husband. Regardless of the cause of bias, review, promotion, and tenure committees should have an awareness of the level of pregnancy bias in student teaching evaluations when assessing faculty dossiers. As each level of review is further removed from faculty candidates, heavily biased student evaluations may in turn bias these committees, resulting in a cumulative negative impact on the victims of bias. The data in this study show substantial and significant bias towards pregnant women of colour.

The data further show that the discipline of the pregnant faculty plays a role. The findings demonstrate that women in engineering and education receive lower teaching evaluation scores than women in other fields, regardless of pregnancy status. Several women in engineering and education described student comments on instructor attitude or behaviour, such as perceived rudeness or disinterest. It is possible that these perceptions may be due to shifting instructor attention and/or student immaturity. It has been documented that women faculty are expected to exhibit more nurturing behaviour than their male counterparts, in the form of counselling, mentoring, and favour requests (El-Alayli et al., 2018). Such time and energy-consuming requests have been termed emotional labour, and it is possible that such expectations are not met at the same level during a woman's pregnancy. Besides the impact of pregnancy on a woman's body, there are at times anxiety surrounding a pregnancy. For instance, though there was no specific question about it, 8% of study participants reported their miscarriages. This suggests a lasting impact of the loss, which many women experience as profound grief, which in turn may further make women less emotionally available for students. This may be perceived as rudeness by students.

Beyond possible differences in women's behaviour, while pregnant, it may simply be errors in student perception. For instance, the neural networks that process emotion in faces continue to develop structurally and functionally throughout adolescence (McClure, 2000; Monk et al., 2003; Vetter et al., 2018), which is defined as early (11–13 years), middle (14–17 years), and late (18–22 years) (Steinberg, 2002). Late adolescence encompasses the college years, and immaturity in processing facial expressions may in part play a role in perceived slights by students. For example, one faculty respondent described students interpreting her decreased mobility as disinterest and her holding her head during frequent headaches as boredom. In addition, female children, adolescents, and adults outperform their male counterparts in facial expression processing, which may in part explain why lower evaluation scores are driven by male students. It also may explain in part the effect observed in this study, that graduate students award higher evaluation scores than first-year students and gave smaller pregnancy penalties. Graduate students are typically out of the late adolescent phase and are better able to process emotions than undergraduate students who are still late adolescents.

Nevertheless, while adolescent misattribution of facial expression may in part explain the drop in scores during a woman's pregnancy, it does not explain the consistent and systematically lower scores awarded to female instructors against their comparable male counterparts. It has been demonstrated that the driving factor for these lower scores is best explained by bias (Bavishi et al., 2010; Boring, 2017; Marsh, 2007; Mengel et al., 2018; Storage et al., 2016; Subtirelu, 2015; Uttl and Smibert, 2017). Even several of the more positive non-engineering comments reflect bias, with students commenting on the instructor's future mothering capabilities rather than her teaching abilities or expressing qualified praise (e.g., “even though she was pregnant”), as though pregnancy precludes competence (Fig. 4).

Finally, the results showed that women in humanities, medicine, and the life, physical, and earth sciences had little to no drop in their

evaluation scores. It is possible that lower teaching evaluations may also be correlated to illness affecting performance; however, these women did not have fewer symptoms, rather they were not penalized for the symptoms they did have. Furthermore, women who have undergone chemotherapy while teaching has reported excellent teaching evaluations (Netz-Fulkerson, 2016; Wisenberg, 2009). Several studies have demonstrated that pregnancy is viewed as a lifestyle choice (Sieverding et al., 2018), while illness is beyond one's control. Hence, in certain fields, women who "choose" to get pregnant receive no sympathy for their symptoms while those who become ill do. This bias is cultural, as a different interpretation of chemotherapy-induced illness vs pregnancy-induced illness is that a cancer patient often must choose to take chemotherapy if she wishes to live longer, just as a woman must choose to get pregnant if she wishes to have biological children. Although having children and the pursuit of life are both considered basic human rights, symptomatic pregnant women—particularly women of colour in engineering—may be considered "deserving" of poorer teaching evaluations while those who are ill are not.

Given that of all the STEM fields, engineering has the fewest female tenure track faculty (16%), compared to the sciences (34%) and mathematics (25%) (National Science Foundation National Center for Science and Engineering Statistics, 2019), these results may help explain why. Becoming pregnant is perceived as detrimental to one's career and as it stands many women delay—or receive advice to delay—having children until they receive tenure (Kavya and Kramer, 2020). This study demonstrates that this perception may be accurate: pregnancy bias impacts teaching evaluations, which in turn impacts tenure and promotion committees. Engineering's low numbers of women may both reflect and contribute to a culture that is less accepting of pregnancy. The implications of these findings may be that the various institutional programmes designed to help women persist in STEM may not be as critical as dismantling the systemic biases that drive them out. The loss of girls and women from STEM fields begins early, as young as age six, and progresses at each stage of a woman's career. In an attempt to describe this female attrition, observers of this phenomenon have likened it to a "leaky pipeline" caused by a "chilly climate." This perspective is often from the outside, looking in. More recently, women within these fields have described this experience as a "gauntlet," with challenges and obstacles along the way that often feel designed to drive them out (Rodrigues and Clancy, 2020; Urry, 2015). While not all bias is deliberate, choosing to use a demonstrably biased evaluation for review, promotion, and tenure is. Since the junior faculty stage often coincides with prime childbearing age, delaying motherhood may not be feasible for all women and may result in profound regret should the delay result in infertility. Several survey respondents indicated multiple pregnancies for which they received poor evaluations, which a review, promotion, and tenure committee might view as inconsistency in teaching rather than a reflection of student bias. Beyond the potential pitfalls of student bias, student evaluations of teaching effectiveness have been shown to be unrelated to student learning, and that a better measure would be direct observation by a non-student evaluator. The results from this study further indicate that when review, promotion, and tenure committees assess female instructors' teaching evaluations, care must be taken to avoid validating student bias against pregnant faculty and to ensure that the "pregnancy drop" in teaching evaluations does not remain yet another obstacle for women in the gauntlet.

Recommendations. The abrupt shift from in-person to online teaching during the initial 2020 shutdown in response to the COVID-19 pandemic left many instructors scrambling to create

online content overnight. In response, many universities offered faculty the option of excluding teaching evaluations from this period. Understanding that there may be profound bias during pregnancy, institutions could also offer this option to women who teach during their pregnancies. It is a simple fix, has been demonstrated to be possible on a large scale, and places the decision with the pregnant women, giving them the control to exclude pregnancy bias.

Study limitations. It is possible that the higher pregnancy scores awarded during the experimental portion of the study compared to the lower pregnancy scores collected from actual female faculty reflects social desirability effects, where the study participants give politically correct answers because they suspect they are being tested for bias. Additionally, the diminished gaps observed when students were evaluating "video quality" rather than "instructor effectiveness," may reflect that students were rating the technical aspects of the video rather than interpreting the rating as a corollary for "course quality." Finally, the stronger bias against women who gained the most weight might indicate bias against women who are more "obviously" pregnant; however, it could also result from a bias against women who are deemed overweight.

Data availability

The datasets generated and/or analysed during the current study are not publicly available because even when deidentified, the low numbers of certain demographics and their responses could be used to identify participants. However, where possible, data are available from the corresponding author upon reasonable request.

Received: 31 October 2020; Accepted: 7 October 2021;

Published online: 29 October 2021

Notes

- 1 This study was performed at Rutgers University prior to the author's move to UC Irvine.
- 2 The study was completed before 2020 when online instruction became more familiar.

References

- Bavishi A, Madera JM, Hebl MR (2010) The effect of professor ethnicity and gender on student evaluations: judged before met. *J Divers High Educ* 3(4):245–256
- Benard S, Correll SJ (2010) Normative discrimination and the motherhood penalty. *Gender Soc* 24(5):616–646
- Benard S, Paik I, Correll SJ (2007) Cognitive bias and the motherhood penalty. *Hastings LJ* 59:1359–1387
- Boring A (2017) Gender biases in student evaluations of teaching. *J Public Econ* 145:27–41
- Boring A, Ottoboni K, Stark P (2016) Student evaluations of teaching (mostly) do not measure teaching effectiveness. *ScienceOpen Res* 0(0):1–11
- Bragger JD, Kutcher E, Morgan J, Firth P (2002) The effects of the structured interview on reducing biases against pregnant job applicants. *Sex Roles* 46(7–8):215–226
- Chawla D (2018) Peer review fails equity test. *Nature* 561(7723):295–296
- Collins PH (2000) *Intersecting oppressions*. Sage Publishing. http://www.uk.sagepub.com/upm-data/13299_Chapter_16_Web_Byte_Patricia_Hill_Collins.pdf. Accessed 29 Aug 2021
- Correll SJ, Benard S, Paik I (2007) Getting a job: is there a motherhood penalty? *Am J Sociol* 112(5):1297–1338
- Crenshaw K (1990) Mapping the margins: intersectionality, identity politics, and violence against women of color. *Stan L Rev* 43:1241–1299
- Davis K (2014) Intersectionality as critical methodology. In: Lykke N (ed) *Writing academic texts differently: intersectional feminist methodologies and the playful art of writing*. 1st ed. Routledge, New York, pp. 17–29
- El-Alayli A, Hansen-Brown AA, Ceynar M (2018) Dancing backwards in high heels: female professors experience more work demands and special favor requests, particularly from academically entitled students. *Sex Roles* 79(3–4):136–150
- Ellmann N, Frye J (2018) Efforts to combat pregnancy discrimination: confronting racial, ethnic and economic bias. *The Center for American Progress*. <https://>

- www.americanprogress.org/issues/women/news/2018/11/02/460353/efforts-combat-pregnancy-discrimination/. Accessed 28 Aug 2021
- Giuns P, Prosser M, Barrie S (2007) Students' perceptions of teaching quality in higher education: the perspective of currently enrolled students. *Stud High Educ* 32(5):603–615
- Gutiérrez y Muhs, G, Niemann, YF, González, CG, & Harris, AP (Eds.) (2012) *Presumed incompetent: The intersections of race and class for women in academia*. University Press of Colorado
- Halpin J, Teixeira R (2010) Latino attitudes about women and society. Racial equity and justice. The Center for American Progress. CDN. https://cdn.americanprogress.org/wp-content/uploads/issues/2010/07/pdf/latino_attitudes.pdf. Accessed 28 Aug 2021
- Hebl MR, King EB, Glick P, Singletary SL et al. (2007) Hostile and benevolent reactions toward pregnant women: complementary interpersonal punishments and rewards that maintain traditional roles. *J Appl Psychol* 92(6):1499
- Heilman ME, Okimoto TG (2008) Motherhood: a potential source of bias in employment decisions. *J Appl Psychol* 93(1):189
- Hengel E (2017) Publishing while female. Are women held to higher standards? Evidence from peer review. *Cambridge Working Papers in Economics* 1753. Faculty of Economics, University of Cambridge
- Holland v. Gee, No. 11-11659 (11th Cir 2012)
- Hornstein HA (2017) Student evaluations of teaching are an inadequate assessment tool for evaluating faculty performance. *Cogent Educ* 4(1):1304016
- Kavya P, Kramer MW (2020) The impact of maternity leave advice within the academy on work-life balance of women faculty and administrators. In: Cabbage J (ed) *Developing women leaders in the academy through enhanced communication strategies*, 1st edn. Lexington Books, Maryland, pp. 75–101
- Lazos SR (2012) Are student teaching evaluations holding back women and minorities?: the perils of “doing” gender and race in the classroom. In Gutiérrez y. In: Muhs G, Niemann YF, González CG, Harris AP (eds) *Presumed incompetent: The intersections of race and class for women in Academia*, 1st edn. Utah State University Press, Boulder, pp. 164–185
- LeMaster J, Marcus-Newhall A, Casad BJ, Silverman N (2004) Life experiences of working and stay-at-home mothers. In: Chin JL (ed.) *The psychology of prejudice and discrimination: gender and sexual orientation*, 1st edn, vol 3. Praeger, New York, pp. 61–91
- Marsh HW (2007) Students' evaluations of university teaching: dimensionality, reliability, validity, potential biases and usefulness. In: Perry R, Smart J (eds) *The scholarship of teaching and learning in higher education: an evidence-based perspective*. Springer, Dordrecht, pp. 319–383
- McClure EB (2000) A meta-analytic review of sex differences in facial expression processing and their development in infants, children, and adolescents. *Psychol Bull* 126(3):424
- Mengel F, Sauermann J, Zölitz U (2018) Gender bias in teaching evaluations. *J Eur Econ Assoc* 17(2):535–566
- Miller JE, Seldin P (2014) Changing practices in faculty evaluation. *Academe* 100(3):35–38
- Moghadam VM (2004) Patriarchy in transition: women and the changing family in the Middle East. *J Comp Fam Stud* 35(2):137–162
- Monk CS, McClure EB, Nelson EE, Zarahn E et al. (2003) Adolescent immaturity in attention-related brain engagement to emotional facial expressions. *Neuroimage* 20(1):420–428
- Morgan WB, Walker SS, Hebl MMR, King EB (2013) A field experiment: reducing interpersonal discrimination toward pregnant job applicants. *J Appl Psychol* 98(5):799
- National Science Foundation National Center for Science and Engineering Statistics (2019) Women, minorities, and persons with disabilities in science and engineering: special report NSF 19-304. <https://ncses.nsf.gov/pubs/nsf19304/>. Accessed 28 Aug 2021
- Netz-Fulkerson JA (2016) *Investigating residual impacts of teachers with cancer*. Dissertation, University of Denver
- Ollilainen M (2019) Ideal bodies at work: faculty mothers and pregnancy in academia. *Gender Educ* 32(7):1–16
- Onwuegbuzie AJ, Witcher AE, Collins KM, Filer JD et al. (2007) Students' perceptions of characteristics of effective college teachers: a validity study of a teaching evaluation form using a mixed-methods analysis. *Am Educ Res J* 44(1):113–160
- Orbuch TL, Custer L (1995) The social context of married women's work and its impact on Black husbands and White husbands. *J Marriage Fam* 57(2):333–345
- Pepin JR, Cotter DA (2018) Separating spheres? Diverging trends in youth's gender attitudes about work and family. *J Marriage Fam* 80(1):7–24
- Rodriguez MA, Clancy KB (2020) A comparative examination of research on why women are underrepresented in some STEM disciplines compared to others, with a particular focus on computer science, engineering, physics, mathematics, medicine, chemistry, and biology. NASEM Commissioned Report. https://www.nap.edu/resource/25585/Commissioned_Paper_Rodriguez.pdf. Accessed Jun 16 2020
- Rosenthal L, Lobel M (2016) Stereotypes of Black American women related to sexuality and motherhood. *Psychol Women Q* 40(3):414–427
- Rubin DL, Smith KA (1990) Effects of accent, ethnicity, and lecture topic on undergraduates' perceptions of nonnative English-speaking teaching assistants. *Int J Intercult Rel* 14(3):337–353
- Schimanski LA, Alperin JP (2018) The Query ID=“Q12” Text=“Please provide the page range or article number for reference Schimanski and Alperin (2018).”-->evaluation of scholarship in academic promotion and tenure processes: past, present, and future. *F1000Res* 7:1605
- Seldin P (1998) How colleges evaluate teaching: 1988 vs. 1998: Practices and trends in the evaluation of faculty performance. *AAHE Bull* 50:3–7
- Sieverding M, Eib C, Neubauer AB, Stahl T (2018) Can lifestyle preferences help explain the persistent gender gap in academia? The “mothers work less” hypothesis supported for German but not for US early career researchers. *PLoS ONE* 13(8):e0202728
- Steinberg LD (2002) *Adolescence*. McGraw-Hill, New York
- Storage D, Horne Z, Cimpian A, Leslie S-J (2016) The frequency of “brilliant” and “genius” in teaching evaluations predicts the representation of women and African Americans across fields. *PLoS ONE* 11(3):e0150194
- Stroebe W (2016) Why good teaching evaluations may reward bad teaching: on grade inflation and other unintended consequences of student evaluations. *Perspect Psychol Sci* 11(6):800–816
- Strunk KK (2020) Demystifying and democratizing tenure and promotion. Inside Higher Ed. <https://www.insidehighered.com/advice/2020/03/13/tenure-and-promotion-process-must-be-revised-especially-historically-marginalized>. Accessed March 13 2021
- Subtirelu NC (2015) “She does have an accent but...”: race and language ideology in students' evaluations of mathematics instructors on RateMyProfessors.com. *Lang Soc* 44(1):35–62
- The Childbirth Connection (2014) Listening to mothers: the experiences of expecting and new mothers in the workplace. A program of the National Partnership for Women and Families. <https://www.nationalpartnership.org/our-work/resources/economic-justice/pregnancy-discrimination/listening-to-mothers-experiences-of-expecting-and-new-mothers.pdf>. Accessed 29 Aug 2021
- Urry M (2015) Science and gender: scientists must work harder on equality. *Nat News* 528(7583):471–473
- Uttl B, Smibert D (2017) Student evaluations of teaching: teaching quantitative courses can be hazardous to one's career. *PeerJ* 5:e3299
- Vetter NC, Drauschke M, Thieme J, Altgassen M (2018) Adolescent basic facial emotion recognition is not influenced by puberty or own-age bias. *Front Psychol* 9:956–968
- Wang L, Gonzalez JA (2020) Racial/ethnic and national origin bias in SET. *Int J Organ Anal* 28(4):843–855
- Waxman S, Ispa-Landa S (2016) Academia's ‘Baby Penalty’. *US News and World Report*
- Williams JC, Phillips KW, Hall EV (2014) Double jeopardy: gender bias against women in science. Tools for change: boosting the retention of women in the STEM pipeline. <http://worklifelaw.org/publication/double-jeopardy-gender-bias-against-women-of-color-in-science/>. Accessed 28 Aug 2021
- Williams JC, Segal N (2003) Beyond the maternal wall: relief for family caregivers who are discriminated against on the job. *Harv Women's Law J* 26:77–162
- Wisenberg SL (2009) *The adventures of cancer bitch*. University of Iowa Press, Iowa City
- Witteman HO, Hendricks M, Straus S, Tannenbaum C (2019) Are gender gaps due to evaluations of the applicant or the science? A natural experiment at a national funding agency. *Lancet* 393(10171):531–540

Acknowledgements

Extreme thanks and gratitude to Shauna Elbers Carlisle, PhD, at the University of Washington Bothell, for her critical review of the manuscript and assistance with SPSS.

Competing interests

The author declares no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1057/s41599-021-00926-3>.

Correspondence and requests for materials should be addressed to Ronke M. Olabisi.

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2021