Student evaluations of physics teachers: On the stability and persistence of gender bias

Geoff Potvin[1,2] and Zahra Hazari[1,2,3]
[1]STEM Transformation Institute, Florida International University, 11200 SW 8th Street, Miami, Florida 33199, USA
[2]Department of Physics, Florida International University, 11200 SW 8th Street, Miami, Florida 33199, USA
[3]Department of Teaching & Learning, Florida International University, 11200 SW 8th Street, Miami, Florida 33199, USA
(Received 7 February 2015; published 1 August 2016)

[This paper is part of the Focused Collection on Gender in Physics.] There is a long history of research which confounds the simple interpretation that evaluations in an educational context are purely measures of competency. One such issue is that of gender bias in student evaluations of their teachers. In our prior work, we found that male students underrated female high school teachers in biology and chemistry while all students underrated female high school teachers in physics. In the current work, we independently checked and extended this earlier work to examine the effect of physics identity on student evaluations and gender bias. Employing multiple regression on survey data from a representative sample of 6772 college students across the U.S., attending both 2-year and 4-year post-secondary institutions (including STEM and non-STEM majors), we find the core physics effect is unchanged despite a gap between studies of nearly 10 years. Namely, both male and female students underrate their female high school physics teachers, even after controlling for physics grades and classroom experiences. Our new focus on physics identity reveals that students with a strong physics identity show a larger gender bias in favor of male teachers than those with less of a physics identity. These results may help to explain how structures that privilege certain groups and marginalize others are prevalent amongst the youngest members of a defined physics community and may serve to uphold the status quo as these young members traverse to higher levels of physics community membership. Furthermore, biased evaluative feedback structures may be one of the propagators of women's lower competency beliefs in physics, a result that has been found by many prior studies.

DOI: 10.1103/PhysRevPhysEducRes.12.020107

I. INTRODUCTION

The way in which individuals see themselves is dependent, at least in part, on the way they are seen by others [1–4]. This is important because how individuals see themselves with respect to certain science fields has implications for their participation and persistence [2,3]. In education, there are very few formal structures to receive feedback about others' perceptions on a wide set of domains such as approachability, communication skills, clarity, knowledge, etc. However, one such feedback structure is the evaluations provided by students, which are often used to assess a teacher's effectiveness. As such, they have the power to not only color individuals' intrinsically held self-perceptions (e.g., students' self-beliefs), but also the extrinsic advancement of others (e.g., teachers). Given the fact that women in physics have consistently been found to have depressed self-perceptions and have less of a history of advancement [5–8], this work considers the issue of how gender affects student evaluations in physics.

Student evaluations also provide an opportunity to study underlying issues that may exist within a community and may be difficult to detect quantitatively since they may lie at an unconscious level. In considering a similar issue with regards to how gender and race affect teacher evaluations, Pitmann [9] proposes that "…our classrooms undoubtedly reflect the oppression of society" where oppression includes "the system of obstacles and the individual acts that maintain the privilege and authority of a dominant group." Thus, oppression may be acted out in the form of bias when students evaluate female physics teachers, even if this bias is not on a conscious level. This type of oppression, Pittmann would argue, serves as a mechanism for discouraging women, both teachers and students, from persisting or advancing, and contributes to maintaining women's marginalized status in physics. In the current study, we reconsider the question of gender bias in student evaluations in physics and add to the current body of work by also incorporating additional student characteristics (specifically, physics identity) to develop a more nuanced understanding of bias in evaluations.

II. PRIOR WORK ON GENDER BIAS IN STUDENT EVALUATIONS

While many people simply interpret student evaluations as a measure of teaching effectiveness, a long history of research assessing the validity and interpretation of their meaning has raised concerns which may confound such a straightforward interpretation. As mentioned, one such issue is that of gender—both of the student and of the teacher. Research on student evaluations has found many contradictory results with respect to possible gender bias [10]. While some studies have found same-sex biases, in which students rated teachers of the same sex as themselves more highly [11], others have found that both male and female students underrated female teachers on different evaluation measures including overall effectiveness and academic competence [12–14]. Additional summaries and meta-analyses of the literature on student evaluations have claimed little to no evidence of gender bias in any sense [15,16], small biases in favor of female teachers [17], biases when teachers do not fulfill their expected gender role [18], and a general ambiguity on findings related to gender bias in student evaluations [19].

Unfortunately, much of the work on gender bias in student evaluations has not accounted for the classroom or disciplinary context of the evaluation. For example, the specific discipline in which the evaluation was occurring is not usually factored in as a potential explanatory variable; rather, students' evaluations are treated as measures that mean the "same" thing in any classroom. As we have argued previously [10], attention to disciplinary context is important given that gender-role expectations may vary in different science disciplines and that such expectations have been found to be important for student evaluations [18]. For example, while Centra and Gaubatz [11] considered the natural sciences separately from other disciplines, they never distinguished between various natural sciences despite the fact that gender representation and societal stereotypes vary widely between them [20]. In our prior work on gender bias in student evaluations [10], we found distinct gender-based patterns across biology, chemistry, and physics: on average, male students underrated female high school teachers in all three sciences while female students only underrated female high school teachers in physics. Furthermore, these effects persisted even after controlling for the context and experiences within the classrooms, e.g., time spent lecturing, use of whole class discussions, real-world examples, etc., and despite the fact that the students of male and female teachers performed equally well (on average) in a subsequent introductory science course in college and were equally likely to persist in their college science career intentions. Thus, we showed that the female teachers who were underrated were equally successful in preparing students for their next physics course and at encouraging students of all genders to persist in science studies [10].

Beyond the disciplinary context, particular classroom experiences, and student educational outcomes, additional sources of variation that are often left unaccounted for when considering gender bias in evaluations are specific characteristics of the student evaluators. In fact, Linse [19] pointed out that many quantitative studies that compared student evaluations of male and female faculty did not account for student gender at all, simply assuming that there would be no differences between male and female students.

III. INTERPRETING STUDENT AFFINITY THROUGH A PHYSICS IDENTITY FRAMEWORK

Although our prior work did differentiate by the gender of the student, we did not consider other important student characteristics such as students' disciplinary identities. However, one might ask, could it be that students who identify more strongly with physics have a greater gender bias in their evaluations, thus serving to replicate the dominant culture and stereotypes if they become more influential members of the community? By framing this work around students' physics identity, we may better understand whether the structures of privilege are being upheld in subtle ways by the views of those who are at the cusp of entering into our community and not solely by those who have already risen to the top.

Identity is central to understanding membership within a community of practice, such as a physics community, and can serve to differentiate between core and peripheral members, as well as nonmembers [21]. Thus, using a physics identity approach allows us to differentiate between students who see themselves as core members of the group who are "physics people" (in the common phrasing of students); that is, it differentiates between those who see themselves as belonging to such a group and those who dissociate themselves from it. Other approaches may not necessarily allow us to see this variation in students' affiliation. For example, performance differences in physics would not necessarily reveal differences in affiliation since some high performing students, particularly female students, have been found to have negative associations with learning physics [22]. Self-efficacy frameworks suffer a similar problem since feeling confident in performing tasks in a subject (e.g., being able to solve problems or conduct experiments) does not always translate to feelings of membership in a community [4,23]. While self-efficacy may be a necessary precursor, it is not a sufficient condition for belongingness. Thus, a physics identity framework is appropriate given our need to understand how early community members might reinforce or challenge structures of privilege.

Another reason for using this framing is that students' physics identity not only indicates their current affiliations with physics but also represents an increased likelihood of their future membership within the community. Measures of students' physics identity at the introductory physics

level in high school and college have been found to predict students' physics-related career choices [3,24,25]. While not surprising, this does give credence to the idea that if the beliefs of potential community members at the early levels support structures of privilege (even unconsciously so), these members will likely reinforce these structures in the future if and when they become fully fledged members of the community.

IV. RESEARCH QUESTIONS

There are three main goals of this work: first, to confirm our previous findings for student evaluations in physics using a new nationally representative data set which was collected almost a decade after our initial work; second, to assess the added effect of students' physics identity on evaluations particularly with respect to gender differences; and, third, to examine whether male and female teachers are equally effective at preparing students. Thus, this paper addresses the following research questions:

- Is there a gender bias in students' evaluations of their high school physics teachers, accounting for both the gender of the student and that of the teacher?
- How do students' physics classroom experiences affect their evaluations and do these experiences account for any gender effects observed?
- How do students' physics identities affect student evaluations, particularly with respect to gender of the student and teacher?
- Are male and female physics teachers equally effective in engaging students so that they pursue physics-related careers in college? Are they equally effective in preparing students to solve physics problems, such as those appearing on AP Physics exams?

The third question is relevant to this discussion because it could indicate whether any gender differences observed in student evaluations may be revealing actual systematic differences in the effectiveness of their teachers; on the other hand, a finding that teachers are equally effective in helping students to pursue physics-related careers and perform in physics would provide evidence that any gender effects observed in evaluations of teachers are due to other effects, including internalized gender perceptions of students.

V. METHODS

In this work, we employ multiple regression methods on nationally representative data. Specifically, we are using data drawn from the Sustainability and Gender in Engineering (SaGE) survey, which surveyed a nationally representative sample of 6772 college students across the U.S. who were attending both 2-year and 4-year post-secondary institutions. This study focuses on 1943 students who took high school physics. The broad goals of the SaGE study were to probe the sustainability-related experiences of college-enrolled students while they were in high school science classes, with a particular interest in understanding how these experiences impacted women in their STEM-related (and, especially, their engineering-related) career interests. The scope of the survey included a range of questions on students' high school science experiences (including questions probing their teachers' practices and effectiveness, which we will discuss in detail below), students' own career interests and outcome expectations for their careers, and items probing students self-beliefs and their demographics. There were a total of 50 institutions that participated in this study which were selected as a result of a stratified random sampling of colleges and universities from a comprehensive database generated from the National Center for Educational Statistics. The sample was stratified by type of institution (2-year and 4-year) and by size (small, medium, and large with equal size bins by student enrollment). This stratification ensured that our sample would have a more representative population of students attending colleges and universities in the U.S. (Note that a pure random sample would result in a disproportionately high number of very small schools, e.g., those with enrollment of less than 2000, since there are many of these schools in the U.S. This would not be representative of the student population since the majority attend larger schools. Thus, stratification allows us to build a more representative sample at the student level.) In support of the representativeness of our sample, the American Institute of Physics (AIP) reported that 37% of high school physics teachers were female in 2013 [26], similar to the 36% reported in our sample in 2011. Figure 1 shows a map of respondents' reported home ZIP codes, which helps to illustrate the national representativeness of students' previous high school experiences.

Within these 50 institutions, students were surveyed in required introductory English courses in order to derive a student sample that includes both STEM and non-STEM majors alike. The instrument was developed and validated during 2010–2011 and the survey was conducted in the Fall semester of 2011. To develop hypotheses and content validity, a brief open-ended survey was administered online to members of the National Science Teacher Association

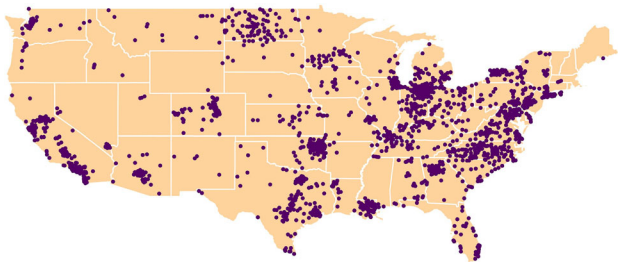

FIG. 1. Map of the home ZIP codes of student respondents. Note that each dot represents a ZIP code, not a student, so each dot may contain more than one student.

(NSTA), with 83 high school science teachers providing responses. The teachers' responses were included in the survey as items, as well as the responses to a similar instrument from 82 first-year engineering and 41 non-engineering majors. Once an initial draft of the instrument was written (also using prior instruments and prior literature), validation of the instrument was carried out through focus group interviews with 11 undergraduate STEM students to help establish face and content validity as well as basic interpretability amongst the study population. A test-retest study helped to establish the reliability of the survey items and included 62 undergraduate STEM and education majors enrolled at two different universities.

The primary outcome variable used in the current paper consists of students' responses to a set of seven, anchored, seven-point items that were written and tested to evaluate students' previous high school physics teachers. They were largely drawn from our previous study [10], but since extended to improve upon reliability and validity. The structure of the items was headed by "How would you rate your LAST high school PHYSICS teacher on the following characteristics?" and each item was anchored from "0—Low" to "6—High." The seven rating elements were "Enthusiasm for physics," "Treated all students with respect," "Explained ideas clearly," "Explained problems and answered questions in several different ways," "Was able to organize lessons and classroom activities," "Was able to handle discipline and manage the classroom," and "Was available to help students outside of class." Once the data were collected, these seven items were analyzed using exploratory factor analysis (EFA) to validate the underlying structure which was theorized to be an overall evaluation construct. EFA is a family of methodologies to explore the relationship between directly measured variables (e.g., the items on the survey) and underlying, indirectly assessed constructs (e.g., an overall evaluation score that predicts students' responses on several items). In fact, we found that a single factor (hereafter called a "teacher evaluation score") explained fully 75% of the total variance in all seven items. This single score, constructed out of the seven items, was used as the outcome in our subsequent regression analysis.

The primary predictors used in our multiple regression analysis included students' self-reported gender as well as the gender of their last high school physics teacher. Table I includes a cross tabulation of the student and teacher genders in the data. When testing for the relationship between evaluations, gender, and students' physics identity, we also used an anchored, five-point item "I see myself as a physics person" as a proxy for physics identity. This item has been found in several independent studies to be the single best proxy for physics identity [3,4]. In this case, the item predicts 23% of the total variance (adjusted $\mathbf{R}^2$) as the sole predictor in a regression onto physics career choice which was another item appearing on the SaGE survey.

TABLE I. Cross-table frequencies for gender of teacher and student.

	Physics teacher gender	
Student gender	Male	Female
Male	614	324
Female	639	366

This provides a measure of concurrent, criterion-related validity for the physics identity item. By comparison, students' grades in physics only predict 2% of the variance in physics career choice. Note that this single item is not intended to measure the *totality* of the nuance and meaning of students' physics identities; rather, in this case we have found, as previously, that this item acts as an excellent and simple stand-in for students' self-perceptions about physics.

Last, to assess the impacts of various teacher practices and classroom experiences, we used several other items from the SaGE survey such as the frequency of lecturing, whole class discussions, covering topics relevant to life, etc. The complete survey can be viewed at http://stem.fiu.edu/sage/.[1]

VI. RESULTS

To address our first research question, we regressed student and teacher genders onto the teacher evaluation score. See the "Model 1" column of Table II for a summary of the resulting linear regression model. We find the gender effect is the same as in our earlier work [10]. Specifically, the model shows that student gender is *not* a significant predictor of teacher evaluations, but teacher gender is, such that female teachers are rated on average 6.26% lower than male teachers ($p < 0.001$) by students of any gender. We also tested for a teacher-student gender interaction effect, which could indicate the presence of a same-gender or opposite-gender bias, but this is not a significant predictor. Despite an eight year gap between this work and our previous study with a completely independent sample (students enrolled in freshman physics courses during Fall 2003), the main effect of teacher gender is strikingly similar, even down to the regression coefficient. This is a strong corroboration of our earlier finding.

Second, to ascertain whether the gender bias effect is robust and not acting as proxy for some sort of systematic differences between the typical classroom choices of male and female teachers, we extended the first regression model to incorporate a number of teacher practices and classroom

[1]Note that all the analyses conducted in this paper were carried out in **R** [27], in particular, using the QuantPsyc [28], car [29], ggplot2 [30], maptools [31], nFactors [32], and plotrix [33] packages. Throughout these analyses, we have set alpha, the maximum acceptable chance of Type I error to be 5%.

TABLE II. Regression models incorporating student evaluations of physics teachers (100 point scale) with gender and classroom experiences. Note that ***: $p < 0.001$, **: $p < 0.01$, *: $p < 0.05$, ns: not significant.

	Model 1			Model 2		
Predictors	B	SE	sig	B	SE	sig
Intercept	74.51	1.05	***	19.69	3.16	***
Gender						
Student gender (SG) (0 = M, 1 = F)	0.37	1.46	ns	−0.74	1.08	ns
Teacher gender (TG) (0 = M, 1 = F)	−6.26	1.79	***	−4.88	1.11	***
SG * TG interaction	−0.11	2.46	ns			
Classroom experiences						
Physics grade (0,…,4.33)				3.65	0.63	***
Focus on conceptual understanding (0,…,4)				2.53	0.63	***
Concepts introduced before equations (0,…,20)				0.39	0.09	***
Demonstrations (0,…,20)				0.53	0.09	***
Topics relevant to life (0,…,20)				0.22	0.08	**
Students asked or answered questions, made comments (0,…,20)				0.45	0.08	***
Designed or built something (0,…,1)				4.85	1.22	**
Integrated ideas (info) from various sources (0,…,1)				3.21	1.14	**
Time studying (0, …, 60+)				1.37	0.37	***
Answered questions using data in tables (0,…,1)				5.13	1.25	***
Answered questions that required new insight and creativity (0,…,1)				3.65	1.16	**
Adjusted $\mathbf{R}^2$	0.01			0.30		

experiences that might be expected to impact teacher evaluations. The resulting model is summarized as "Model 2" in Table II. Most importantly, the gender effects of Model 1 do not change in substance: student gender remains non-significant while teacher gender is still significant at the $p < 0.001$ level, and the bias is in the same direction (with roughly the same regression coefficient). On top of this, a number of classroom experiences and teacher practices were found to be significant, such that the entire model explains 30% of the variance in teacher evaluation score, as measured by the adjusted coefficient of determination $\mathbf{R}^2$. Unsurprisingly, students' final grade in their high school physics class is a large and significant predictor of teacher evaluation: students who perform better also rate their teachers more highly ($3.65\% \pm 0.63\%$ per point of student GPA, $p < 0.001$). Since our results are correlational in nature, we cannot say whether students who have more highly rated teachers perform better or if students who perform better rate their teachers more highly. Nonetheless, this predictor is somewhat as expected, and is consistent with our previous study [10]. In terms of classroom focus, students who reported that their classes focused more heavily upon conceptual understanding ($2.53\% \pm 0.63\%$ per point out of 5) or introduced concepts before equations ($0.39\% \pm 0.09\%$ per point on a scale 0,…,20 measuring the number of class days per month this event happened) rated their teachers more highly (both $p < 0.001$). Teachers who more regularly conducted demonstrations were rated more highly ($0.53\% \pm 0.09\%$ per point on a 0,…,20 scale representing the number of days per month this event happened, $p < 0.001$), as were teachers who more frequently addressed topics relevant to students' lives (by $0.22\% \pm 0.08\%$ per point on the same scale, $p < 0.01$). As a measure of the level of student participation in the classroom, individuals who reported that students asked questions, answered questions, or made comments more frequently rated their teachers more highly ($0.45\% \pm 0.08\%$ on the same 0,…,20 scale, $p < 0.001$). Students who reported that they spent more time studying rated their teachers more positively ($1.37\% \pm 0.37\%$ per minute of studying per day, on average, $p < 0.001$). Last, the final three positive predictors that appear in this model are students reporting that they ever had the chance to design or build something ($4.85\% \pm 1.22\%$, $p < 0.01$), had to answer test questions that involved using data in tables ($5.13\% \pm 1.25\%$, $p < 0.001$), or had to answer test questions that required new insight or creativity ($3.65\% \pm 1.16\%$, $p < 0.01$).

Though the main purpose of the second, extended regression model was to test whether or not the gender bias effect held up, it is worthwhile to consider the importance of finding the classroom experience variables to be significant predictors of teacher evaluations. It is gratifying to find that several experiences that have been argued to be beneficial for students' learning are also associated with improved teacher evaluations. For example, making class relevant to students' lives, focusing on conceptual understanding rather than on equations or mathematics first, having students more actively engaged (commenting, asking, or answering questions), and having to integrate knowledge from various sources are all indicative of more active and/or reformed practices consistent within much physics education research.

Thus, these results collectively provide another piece of evidence in favor of these teaching practices in general—they are associated with improved teacher evaluations. This should be seen as an added incentive for individual educators to adopt them.

Third, and perhaps most interestingly, we were able to examine how students' own identification with physics relates to the gender bias effects (e.g., "Do students with significant physics interests or intentions show this bias to a greater or lesser degree than the general student population?"). This is something that we were unable to address in earlier work due to limitations of the prior data, and the state of identity research in physics at the time. Since then, we and others have shown that physics identity is a highly relevant construct for understanding students' physics-related career choices [4,5]. Hence, we were interested in the current work to address the second main research question. We built a third regression model which incorporated the gender items as well as the anchored, five-point item "I see myself as a physics person" described earlier as predictors. The results of this model appear in Table III. Note that due to the highly collinear nature of physics identity and classroom experiences (including student grades), it was not possible to also incorporate the classroom predictors from Model 2.

The results show new, compelling effects. First, in this model, student gender becomes significant at the $p < 0.01$ level such that female students rate their teachers higher by $3.56\% \pm 1.19\%$, on average. Second, teacher gender continues to be a significant predictor on its own, with similar bias as before ($-4.10\% \pm 1.84\%$, $p < 0.05$). Interestingly, students' physics identity is a positive, significant predictor of teacher evaluations ($5.71\% \pm 0.53\%$ per point on a 5-point anchored scale, $p < 0.001$) such that students who show a stronger self-identification in physics rate their teachers more highly, on average. In addition, there is an interaction effect between teacher gender and students' physics identity as follows: as students' physics identity proxy increases, their evaluation score for their teacher increases by a significantly *smaller* amount if their teacher is female. Separately (not shown in Table III), we tested the significance of the interaction with student gender and the physics identity proxy; this was found to be nonsignificant. The combined effects of this model are shown in Fig. 2. As can be seen, students at the high end of the identity proxy scale rate a male teacher more than 10% higher, on average, than a female teacher. The key observation is that while female students are slightly more generous in their average ratings (hence the two pairs of parallel lines), the pair associated to male teachers has a significantly larger slope than the pair associated to female teachers.

Finally, in order to help account for the possibility that male and female teachers are somehow systematically different in their effectiveness as teachers (in a way not captured by the classroom and pedagogical experience factors appearing in the second regression model), we examined whether or not students of a male teacher or female teacher were more likely to choose a physics career. We found no significant difference [$t(1443) = -0.74$, $p = 0.46$]. This was true when considering both male students [$t(658) = 0.31$, $p = 0.76$] and female students [$t(694) = -1.71$, $p = 0.09$] separately. This continues to

TABLE III. Regression model predicting student evaluations of physics teachers (100 point scale) incorporating gender and physics identity proxy. Note that ***: $p < 0.001$, **: $p < 0.01$, *: $p < 0.05$.

	Model		
Predictors	B	SE	sig
Intercept	64.02	1.34	***
Student gender (SG) (0 = M, 1 = F)	3.56	1.19	**
Teacher gender (TG) (0 = M, 1 = F)	−4.10	1.84	*
Physics identity (PI)	5.71	0.53	***
TG*PI interaction	−1.75	0.89	*
Adjusted $\mathbf{R}^2$	0.09		

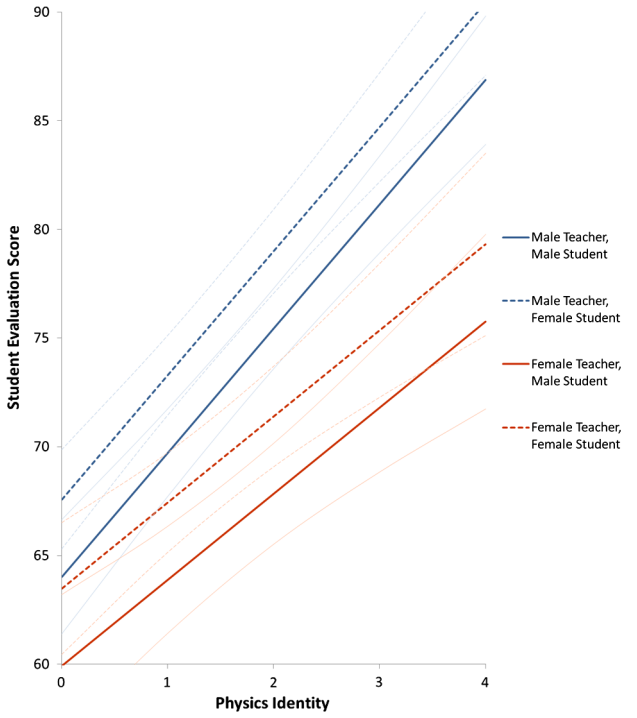

FIG. 2. Regression model prediction of student evaluations (100 point scale) versus physics identity proxy (0,…,4) disaggregated by the gender of physics teacher and student. (Faint lines represent 95% confidence intervals). Note that (a) female students rate teachers slightly higher than male students in this Model (the dashed lines are above the solid lines for each pair of lines, respectively), (b) students with a higher physics identity proxy rate their teachers more highly in general (all slopes are positive), and (c) that male teachers receive a larger increase in their evaluations from students with higher physics identities (the top pair of lines has a greater slope than the bottom pair).

be true even after controlling for students' prior interests in both physics careers and STEM careers at the beginning of high school. Although we did not have students' performance in their subsequent college physics course for this study, we did collect AP exam scores for the subset of respondents who took these exams. Since these scores were independent of the teachers' manipulation (they neither wrote nor graded them), we compared the scores of students who had male and female teachers. We found no significant differences either for AP Physics B [$t(129) = -0.39$, $p = 0.70$] or AP Physics C [$t(84) = -1.4$, $p = 0.17$]. Thus, the evidence indicates that male and female high school teachers are equally effective in encouraging students to choose or persist towards physics careers and in performing on AP exams, two other measures of teachers' effectiveness.

VII. DISCUSSION

Our findings indicate that gender bias continues to be a concern in student evaluations of physics teachers. There is some importance to our reconfirmation of the basic gender bias effect in teacher evaluations, which persists even when controlling for a number of classroom experiences and teacher practices (as indicated by the significant teacher gender effect in the first two regression models). This is a strong replication made with an independent measurement on an entirely new student sample, separated by eight years. The stability of this finding should worry educators concerned with improving the participation of women in physics in significant ways—recall that the bias exhibited here is true for both male and female students at the end of high school or beginning of college. Thus, we cannot assume that these biases are present only amongst a cadre of senior scientists who can act to marginalize women [34]; these attitudes that generally marginalize the competency of women in physics appear to be common amongst participants at the introductory or peripheral stages of physics participation.

In some ways, these findings counter other work that reports little to no gender bias in students' evaluations of teaching. However, as mentioned earlier, this prior work usually does not take into account disciplinary context and aggregates over disparate disciplines whereas our work looks at a specific discipline. We might ask, why should we assume that the same gender bias, or lack thereof, would exist in different fields that hold widely varying gender-related expectations and stereotypes? Gender research that examines beliefs across disciplines has found that stereotypes associated with fields can affect female representation [20], particularly when the stereotypes are also associated with gender identities (for example, the stereotype that physics requires greater visual-spatial abilities than other fields and males have greater visual-spatial abilities). In a recent Science article, the level to which the stereotype of "innate genius" was associated with 30 different fields predicted female representation in those fields—the more innate genius was associated to a field, the fewer females were found at doctoral levels in those fields [20]. This finding shows that popular beliefs about different fields have strong implications for participation in those fields and can separate the privileged from the marginalized. Our work shows that one way in which this privileging can manifest itself in physics is through formal feedback structures such as student evaluations, which could also bleed into peer-to-peer feedback and interactions.

One bright spot in this work is that we find further evidence in favor of more active or reformed classroom practices—in particular, factors that are associated with these practices are generally predictive of improved teacher evaluations. This should lend further credence to the desirability of adopting these teaching practices in physics classrooms, though there is extremely strong evidence in favor of active learning already [35].

The most compelling finding in this paper is the result showing that students who have the strongest affinity for physics (as indicated by the physics identity proxy) exhibit a larger bias against female teachers in their evaluations (as shown by the significant interaction between teacher gender and the physics identity proxy, see Fig. 2). This pattern is true for both male and female students (as indicated by the fact that the interaction between student gender and the physics identity proxy was not significant when tested). This further complicates the issue, because not only is the gender bias seen in the general student population, it is stronger amongst those who are more likely to become members of established physics communities in the future. Thus, this work likely indicates that structures that privilege one group over another can be replicated from both external driving forces (e.g., views of the general population) and internal ones (e.g., views of the disciplinary community members). How individuals are recognized and evaluated for their teaching is one such structure.

Other studies have found this type of gender bias in other internal structures, for example, gender differences in how potential science job candidates are evaluated by science faculty [34]. Qualitative research on the gendering of physics has also found that graduate students in physics reproduce gendered norms, such as not acting in stereotypically feminine ways, in order to maintain feelings of competence [36]. This is a more subtle manifestation of how one group is privileged over another since those who do not fit the gendered norms are either compelled to align with them or suffer from feelings that they are somehow less competent. Our research points to the early prevalence of a similar gendering (exhibited through a gender bias) amongst fledgling members of the physics community as they judge others (their physics teachers in this case). This is important since it is sometimes believed that these norms and accompanying views are primarily held by older members of science communities and, once these members are no longer central, the issues for women will dissipate. However, as

Urry [37] has insightfully pointed out, "We are almost all prejudiced in the sense that we have absorbed the gender and race stereotypes that prevail in our society," which includes our youngest members, both male and female, even if they are not conscious of it. This is reinforced by another aspect of our findings—gender bias in evaluations that favors male teachers is not uniquely held by male students. Both male and female students show the gender bias to the same degree, which facilitates the replication of structures of privilege by members of a community, even those who belong to the underprivileged group.

Finally, one may attempt to dismiss the importance of finding gender bias in teacher evaluations outside of the post-secondary environment. After all, it is rare that they be used in performance assessments of educators in the secondary school sphere. However, the importance of this work lies in its implications for our understanding of students' own gendered expectations of appropriate roles and related competencies for teachers as well as peers. In other words, the bias we observe towards the teacher is also likely to be elicited in subtle ways towards peers and others. Gendered views about physics amongst peers have previously been reported and were found to have strong implications for the science-related interests of students [38,39]. Furthermore, there is strong evidence that female students in introductory physics classes are more likely to draw on vicarious experiences (e.g., observing how others are treated) in the development of their beliefs about their own abilities [40].

The mechanism by which these gender biases translate into individuals' actions, however, continues to be unclear. For example, do women, after repeatedly experiencing these types of biases, choose to leave the field or relinquish advancement because they feel inadequate? Prior work does indicate that female teachers are more susceptible than their male counterparts to negative emotional responses ("anxious, disheartened, depressed, worried, frustrated, angry, irritated, and disgusted") after receiving student evaluations [41] and are more likely to internalize negative feedback as a reflection of their own abilities [42]. This serves as a doubly negative effect—not only is there an actual gender bias but those suffering from the bias may be more likely to blame themselves for negative feedback.

Some limitations of this work should be kept in mind. First, the correlational analyses presented are based on a retrospective, cohort design in data collection which cannot establish causality (though we could rule out some causal hypotheses when nonsignificant), nor can these analyses uncover the deep mechanisms connecting individuals' beliefs about physics competency and gender with their actions (evaluations) since these beliefs may largely be unconscious. Future qualitative research would be more appropriate for this purpose. Second, in keeping with U.S. census practice, students' gender responses (and that of their previous high school science teachers) was limited to dichotomous "male" or "female" options. This is a limited choice but it was made for interpretive considerations of quantitative data and for practical (e.g., data collection) purposes. Third, our results directly confirm the presence of a persistent bias against female physics high school teachers as assessed by their former students (after enrolling in college); while we believe that this bias is likely very pertinent in the interpretation of college course evaluations in general, future work will need to further investigate bias in formal course evaluations in college physics classes. Similarly, we have not measured gender bias in *peer assessment* (e.g., students evaluating one another), although this is one of the primary concerns that our research raises—unconscious gender bias may contribute to the depressed attitudes and feelings of competence of women in physics classes.

ACKNOWLEDGMENT

This work was supported by the National Science Foundation under Grant No. 1036617.

[1] J. P. Gee, Identity as an analytic lens for research in education, Rev. Res. Educ. **25**, 99 (2000).
[2] H. B. Carlone and A. Johnson, Understanding the science experiences of successful women of color: Science identity as an analytic lens, J. Res. Sci. Teach. **44**, 1187 (2007).
[3] Z. Hazari, G. Sonnert, P. M. Sadler, and M.-C. Shanahan, Connecting high school physics experiences, outcome expectations, physics identity, and physics career choice: A gender study, J. Res. Sci. Teach. **47**, 978 (2010).
[4] G. Potvin and Z. Hazari, The Development and Measurement of Identity across the Physical Sciences, *Proceedings of Physics Education Research Conference 2013* (American Association of Physics Teachers, College Park, MD, 2013).
[5] Z. Hazari, P. M. Sadler, and G. Sonnert, The science identity of college students: Exploring the intersection of gender, race, and ethnicity, J. Coll. Sci. Teach. **42**, 82 (2013).
[6] P. Murphy and E. Whitelegg, Girls and physics: Continuing barriers to 'belonging', Curriculum J. **17**, 281 (2006).
[7] W. Adams, K. Perkins, N. Podolefsky, M. Dubson, N. Finkelstein, and C. Wieman, New instrument for measuring student beliefs about physics and learning physics: The Colorado learning attitudes about science survey, Phys. Rev. ST Phys. Educ. Res. **2**, 010101 (2006).
[8] G. Potvin and R. H. Tai, Examining the relationships among doctoral completion time, gender, and future salary

prospects for physical scientists, J. Chem. Educ. **89**, 21 (2012).

[9] C. T. Pittman, Race and gender oppression in the classroom: The experiences of women faculty of color with white male students, Teaching Sociology **38**, 183 (2010).

[10] G. Potvin, Z. Hazari, R. H. Tai, and P. M. Sadler, Unraveling bias from student evaluations of their high school science teachers, Sci. Educ. **93**, 827 (2009).

[11] J. A. Centra and N. B. Gaubatz, Is there gender bias in student evaluations of teaching?, J. Higher Educ. **71**, 17 (2000).

[12] S. A. Basow and N. T. Silberg, Student evaluations of college professors: Are female and male professors rated differently?, J. Educ. Psychol. **79**, 308 (1987).

[13] J. Miller and M. Chamberlin, Women are teachers, men are professors: A study of student perceptions, Teaching Sociol. **28**, 283 (2000).

[14] J. Sidanius and M. Crane, Job evaluation and gender: The case of university faculty, J. Appl. Soc. Psychol. **19**, 174 (1989).

[15] W. E. Cashin, Student ratings of teaching: The research revisited, *IDEA paper No. 32, Center for Faculty Evaluation and Development* (Kansas State University, Manhattan, KS, 1995).

[16] K. A. Feldman, College students' views of male and female college teachers: Part I—Evidence from the social laboratory and experiments, Res. High. Educ. **33**, 317 (1992).

[17] K. A. Feldman, College students' views of male and female college teachers: Part II—Evidence from students' evaluations of their classroom teachers, Res. High. Educ. **34**, 151 (1993).

[18] K. Anderson and E. D. Miller, Gender and student evaluations of teaching, Political Sci. Politics **30**, 216 (1997).

[19] A. R. Linse (2003). Student ratings of women faculty: Data and strategies. Retrieved July 31, 2008, from https://advance.washington.edu/apps/resources/docs/20030513-student_ratings_ds.pdf.

[20] S.-J. Leslie, A. Cimpian, M. Meyer, and E. Freeland, Expectations of brilliance underlie gender distributions across academic disciplines, Science **347**, 262 (2015).

[21] E. Wenger, *Communities of Practice: Learning, Meaning, and Identity* (Cambridge University Press, Cambridge, England, 1998).

[22] H. B. Carlone, The cultural production of science in reform-based physics: Girls' access, participation, and resistance, J. Res. Sci. Teach. **41**, 392 (2004).

[23] J. Cribbs, Z. Hazari, G. Sonnert, and P. M. Sadler, Establishing an explanatory model for mathematics identity, child development, Child Development, **86**, 1048 (2015).

[24] Z. Hazari, A. P. Cass, and C. Beattie, Obscuring Power Structures in the Physics Class: Linking Teacher Positioning, Student Engagement, and Physics Identity Development, J. Res. Sci. Teach. **52**, 735 (2015).

[25] A. Godwin, G. Potvin, Z. Hazari, and R. M. Lock, The development of critical engineering agency, identity, and impact on engineering career choices, *Proceedings of the American Society for Engineering Education (ASEE) Annual International Conference, Atlanta, GA, 2013* (American Society for Engineering Education, Washington, DC, 2014).

[26] S. White and J. Tyler, Who Teaches High School Physics? Results from the 2012-13 Nationwide Survey of High School Physics Teacher, *American Institute of Physics Statistical Research Center* (AIP, College Park, MD, 2014), http://aip.org/statistics.

[27] R Core, Team R: A language and environment for statistical computing, *R Foundation for Statistical Computing* (R Foundation, Vienna, Austria, 2014), http://www.R-project.org/.

[28] T. D. Fletcher, QuantPsyc: Quantitative Psychology Tools. R package version 1.5. http://CRAN.R-project.org/package=QuantPsyc (2012).

[29] J. John Fox and S. Weisberg, *An R Companion to Applied Regression*, 2nd ed. (Sage Publishing, Thousand Oaks, CA, 2011), http://socserv.socsci.mcmaster.ca/jfox/Books/Companion.

[30] H. Wickham, *ggplot2: Elegant Graphics for Data Analysis* (Springer, New York, NY, 2009).

[31] R. Bivand and N. Nicholas Lewin-Koh, maptools: Tools for Reading and Handling Spatial Objects. R package version 0.8–34. http://CRAN.R-project.org/package=maptools (2015).

[32] G. Raiche, nFactors: an R package for parallel analysis and non-graphical solutions to the Cattell scree test. R package version 2.3.3. http://CRAN.R-project.org/package=nFactors (2010).

[33] J. Lemon, Plotrix: A package in the red light district of R, R-News **6**, 8 (2006).

[34] C. A. Moss-Racusin, J. F. Dovidio, V. L. Brescoll, M. J. Graham, and J. Handelsman, Science faculty's subtle gender biases favor male students, Proc. Natl. Acad. Sci. U.S.A. **109**, 16474 (2012).

[35] S. Freeman, S. L. Eddy, M. McDonough, M. K. Smith, N. Okoroafor, H. Jordt, and M. P. Wenderoth, Active learning increases student performance in science, engineering, and mathematics, Proc. Natl. Acad. Sci. U.S.A. **111**, 8410 (2014).

[36] A. J. Gonsalves, Physics and the girly girl—there is a contradiction somewhere: Doctoral students' positioning around discourses of gender and competence in physics, Cult. Stud. Sci. Educ. **9**, 503 (2014).

[37] M. Urry, Speeding up the long slow path to change, APS news **12**, 12 (2003).

[38] S. J. Farenga and B. A. Joyce, Intentions of young students to enroll in science courses in the future: An examination of gender differences, Sci. Educ. **83**, 55 (1999).

[39] U. Kessels, Fitting into the stereotype: How gender-stereotyped perceptions of prototypic peers relate to liking for school subjects., Eur. J. Psychol. Educ. **20**, 309 (2005).

[40] V. Sawtelle, E. Brewe, and L. H. Kramer, Exploring the relationship between self-efficacy and retention in introductory physics., J. Res. Sci. Teach. **49**, 1096 (2012).

[41] L. R. Kogan, R. Schoenfeld-Tacher, and P. W. Hellyer, Student evaluations of teaching: Perceptions of faculty based on gender, position, and rank, Teach. Higher Educ. **15**, 623 (2010).

[42] S. Beyer, Gender differences in causal attributions by college students of performance on course examinations, Current psychology **17**, 346 (1998).