

# Unraveling Bias From Student Evaluations of Their High School Science Teachers

GEOFF POTVIN, ZAHRA HAZARI

*Department of Engineering & Science Education and Department of Mathematical Sciences, Clemson University, Clemson, SC 29634, USA*

ROBERT H. TAI

*Curry School of Education, University of Virginia, Charlottesville, VA 22904, USA*

PHILIP M. SADLER

*Science Education Department, Harvard Smithsonian Center for Astrophysics, Cambridge, MA 02138, USA*

*Received 27 March 2008; revised 29 October 2008, 4 November 2008; accepted 6 November 2008*

*DOI 10.1002/sce.20332*

*Published online in Wiley InterScience (www.interscience.wiley.com).*

**ABSTRACT:** In this study, the evaluation of high school biology, chemistry, and physics teachers by their students is examined according to the gender of the student and the gender of the teacher. Female teachers are rated significantly lower than male teachers by male students in all three disciplines, whereas female students underrate female teachers only in physics. Interestingly, physics is also the field that suffers the greatest lack of females and has been criticized most for its androcentric culture. The gender bias in teacher ratings persists even when accounting for academic performance, classroom experiences, and family support. Furthermore, male and female teachers in each discipline appear equally effective at preparing their students for future science study in college, suggesting that students have a discipline-specific gender bias. Such a bias may negatively impact female students and contribute to the loss of females in science, technology, engineering, and mathematics fields. © 2009 Wiley Periodicals, Inc. *Sci Ed* 1–19, 2009

*Correspondence to:* Geoff Potvin; e-mail: gpotvin@clemson.edu

© 2009 Wiley Periodicals, Inc.

## INTRODUCTION

We are almost all prejudiced in the sense that we have absorbed the gender and race stereotypes that prevail in our society.

—Urry (2003, p. 12)

Beliefs regarding appropriate roles for males and females are as pervasive in our society as they are subtle. Science fields are no exception to the assigning of these stereotypic roles (Barman, 1997; Hughes, 2001; Kahle & Meece, 1994; Lederman, 2003). There is evidence that the inculcation of gender stereotypes begins at a young age and that young students quickly learn which sciences are “appropriate” for them. At the kindergarten level, gifted girls show equal interest in biological and physical science activities (Johnson, 1999). However, a few years later, in grades 4–6, Farenga and Joyce (1999) found that both boys and girls perceive physical science and technology-related courses as appropriate subjects for boys to study and life sciences as appropriate for girls. Several other studies also show that by late elementary school, gender differences in interests have developed with boys preferring physical science topics and girls preferring life science topics (Baker & Leary, 2003; Dawson, 2000; Jones, Howe, & Rua, 2000).

The adoption of gender-role stereotypes has considerable implications for both educational (performance, learning, and curricular) and systemic (cultural, climatic, and representational) outcomes. It can impact students’ perceptions of their own science abilities (Andre et al., 1999; Meece et al., 2006; Selimbegovic, Chatard, & Mugny, 2007), performance (Aronson et al., 1999; Dar-Nimrod & Heine, 2006; Spencer, Steele, & Quinn, 1999; Steele, 1997), interests/choices regarding fields of study (DeBacker & Nelson, 1999; Cleaves, 2005; Farenga & Joyce, 1999), and their views/judgments about others (Kessels, 2005). Andre et al. found that both girls and boys in grades 4–6 perceived physical science as being dominated by males and that parents perceived science as being more important for boys. Not surprisingly, then, they also found that girls perceived themselves to have significantly lower self-competence in physical science than did males. However, they did not draw any theoretical or quantitative connections between the two results; the purpose of their paper was to provide a descriptive database of attitudinal differences. Drawing such a connection in high school, Selimbegovic et al. (2007) found that “female students who believed in gender stereotype[s] evaluated themselves more negatively in math/science than those who refuted these beliefs” (p. 281). Furthermore, Hollinger (1991) writes that “of all the existing barriers, sex-role socialization’s impact on the child’s developing self-belief system is the most pervasive and limiting” (p. 136) and Meece et al. (2006) concludes that small gender differences in competency beliefs emerge in elementary school, which “follow gender norms and stereotypes” (p. 356).

Coupled with self-competency beliefs, stereotypes can also influence actual performance. A substantial body of literature exists on quasi-experimental studies where introducing a stereotype threat (e.g., a test that differentiates between genders) before testing significantly lowers the performance of the negatively stereotyped group in comparison with control groups where the threat is not introduced. This result has been found for females as compared with males on mathematics tests (Spencer et al., 1999; Steele, 1997), African Americans as compared with Whites on verbal tests (Steele, 1997), and White males as compared with Asians on mathematics tests (Aronson et al., 1999). However, there are two caveats: for a stereotype threat to manifest itself in performance outcomes, as in many of these studies, the subjects must care about identifying with the domain and have sufficient performance anxiety or frustration (e.g., when administered a difficult test) (Aronson et al., 1999; Steele,

1997). More recently, Dar-Nimrod and Heine (2006) compared the performance of four treatment groups of women reading a different essay before taking a mathematics test. One essay intimated that there was no gender difference in mathematical performance, the second primed gender (a general essay about women), the third intimated that males performed better on mathematics tests because of genetic differences, and the last that males performed better on mathematics test because of biased teaching experiences faced in the formative years of early schooling. Women in the first (no gender difference) and last (experience-based gender difference) groups did equally well and significantly better than women in the second (primed gender) and third (gene-based gender difference) groups. Clearly, stereotypes, both general and specific, can influence performance.

Although not entirely independent of the constructs of self-competency beliefs and achievement, interests and choices regarding field of study are also influenced by stereotypes. Cleaves (2005) found that stereotypic views of science and scientists were more often held by high school students who did not choose post-compulsory science courses. Similarly, Selimbegovic et al. (2007) found large gender differences in the intention to pursue mathematics/science studies between males and females who believed in the stereotype that boys were better in mathematics and science. However, there were no gender differences in the intention to pursue mathematics/science studies among male and female students who did not believe in the stereotype. In support of this, DeBacker and Nelson (1999) also found that perceiving science as a male domain was correlated negatively with persistence as well as achievement for high school girls. Lastly, Farenga and Joyce (1999) found in grades 4–6 that both boys and girls largely adhered to the stereotype that physical science and technology-related subjects are appropriate for boys, whereas life sciences are appropriate for girls and that these numbers surprisingly resembled the enrollment data for master's and doctoral candidates in the sciences.

Also relevant to the current study is the influence that stereotypic views have on individuals' judgments about others. This has strong implications for the cultural climates in different fields. It has been found that students' adherence to stereotypic gender roles in science is greater when students selected options for the opposite gender than when they selected for themselves (Farenga & Joyce, 1999). Kessels (2005) found that both male and female high school students perceived peers who preferred physics as possessing more masculine and fewer feminine traits. In addition, boys in Kessel's study were more likely to dislike girls who were the best in physics and girls who did well in physics did report that they felt unpopular with boys. Finally, Andre et al. (1999) found that parents of elementary school children adhered to the stereotype that boys are more competent in science and, likewise, perceived science as more important for their boys and expected higher performance of boys. In either case, whether students judge students or parents judge children, it is clear that stereotypes play into these judgments. In this study, we turn our attention to students' judgments about their teachers and ask the question, "do their evaluations reflect stereotypic views that perpetuate gender bias?"

## STUDENT EVALUATION OF TEACHERS

One situation in which students have the opportunity to externalize gender bias on others is in the case of student evaluations of their teachers. There is extensive literature on the subject of gender bias in teacher evaluations (e.g., Anderson & Miller, 1997; Basow & Silberg, 1987; Burns-Glover & Veith, 1995; Cashin, 1995; Centra & Gaubatz, 2000; Feldman 1992, 1993; Linse, 2003; Miller & Chamberlin, 2000; Sidanius & Crane, 1989; Statham, Richardson, & Cook, 1991); the collective conclusions are rich but also complex. Given this complexity, it is instructive to consider these works in detail.

#### 4 POTVIN ET AL.

Centra and Gaubatz (2000) used data from 741 college classes in a wide variety of subjects to examine potential bias in student evaluations of their teachers. Aggregating across all classes and all subjects, they found that female students rated their female professors somewhat higher than their male professors and the ratings of male students also had a (smaller) same-sex bias. The authors left open the possibility that gender differences in teaching styles might account for this differential in evaluations. The authors also considered the natural sciences separately and found slightly different results: both female and male students rated female instructors somewhat higher than male instructors. However, they did not go further to distinguish between the various physical and life sciences. In addition, they presented partial evidence that the differences in student ratings might be due to differences in teaching style—female teachers were rated to be more helpful and approachable and also used discussion more frequently than their male counterparts, who more often relied upon pure lecturing.

Basow and Silberg (1987) compared the “teaching effectiveness” rating of 16 pairs of professors as evaluated by more than 1000 students. Each pair of professors included one male and one female, who had been matched according to course division, tenure status, and teaching experience. The authors found that male students underrated female professors on four (of six) measures and female students also underrated female professors on three measures. These gender effects were only partly explained by the degree to which professors exhibited certain behaviors/attitudes as well as students’ major and class standing.

Sidanius and Crane (1989) found similar results. In a large survey that included the evaluation of 401 instructors by nearly 9000 students, the authors found that male professors were rated significantly higher than female faculty with respect to their overall teacher effectiveness and academic competence by both male and female students. Furthermore, academic competence was seen as more important in student ratings of male instructors’ overall performance than that of female instructors.

Burns-Glover and Veith (1995), in a study tracking students’ preferences for the traits and behaviors of a “great professor,” as measured by students’ evaluations of fictional applicants for a university teaching position, found a bias toward “masculine” traits over “feminine” traits. Also, this bias was exhibited more clearly by male students than by female students and there was an interaction between the students’ preferences for these traits and the professor’s gender.

Lastly, Miller and Chamberlin (2000) examined sociology students’ perceptions of their instructors’ level of educational achievement. In particular, students were asked to attribute the highest level of education (i.e., the highest degree) attained by various male and female instructors, which included graduate students, assistant, associate, and full professors. On average, students were more likely to overestimate their male instructors’ actual level of education while underestimating that of their female instructors. The authors felt that these results showed a pattern of gender bias and recommended proactive measures to counter it, namely, required courses in gender issues for students.

There have been several attempts to synthesize the results of the various individual studies of student evaluations (e.g., Anderson & Miller, 1997; Cashin, 1995; Feldman, 1992, 1993; Linse, 2003). Cashin summarized the existing literature on student evaluations and highlighted, among other issues, the complexities of multidimensionality, reliability, and validity of such measures. He claimed that gender (as well as several variables including age, teaching experience, race, and personality) “tend[s] to show *little or no* relationship to student ratings” (p. 5), citing Feldman (1992, 1993) for his conclusions on gender bias.

Feldman’s two papers (1992, 1993) summarized the findings of 14 experimental studies and 28 studies that analyzed actual teaching evaluations. In the earlier work, Feldman found that few experimental studies uncovered gender bias, although some showed a bias

in favor of male teachers. In the later work, his meta-analysis found a small bias in favor of female teachers but he also pointed out that some studies suggest that female teachers might have to perform at a higher level in order to receive the same ratings as male teachers. In particular, he also summarized a few discipline-specific studies in the social sciences that found varying degrees of bias: education students tend to rate female instructors higher than males, whereas those in anthropology, communication studies, and psychology rate them lower. On these findings, Miller and Chamberlin (2000) comment as follows:

Feldman's [1992 and 1993] gender findings, reportedly small and insignificant, highlight a problem with meta analytic studies . . . First, the mathematics used [to analyze the collective results] disguise large differences in empirical findings . . . Second, the studies summarized ranged from evaluations of single courses to evaluations across college campuses. They summarized evaluations from undergraduate and graduate students enrolled in community colleges, private liberal arts colleges, and research universities. Third, variations in evaluation scores are apparently associated with field of study . . . All told, the meta analytic approach misaggregates data across studies, resulting in a masking of the differences, including gender biases, in students' perceptions of teaching effectiveness. (pp. 284–285)

Evidently, the varied results of the earlier literature, as analyzed by Feldman, led Cashin (1995) to the conclusion that these contrasting findings nullified each other, at least with respect to questions of gender bias.

Anderson and Miller (1997), in a summary of previous literature covering both “laboratory” studies (i.e., experimental-design studies) and “real-life” studies, also found ambiguous results with respect to gender bias but pointed out that “one important consistency that emerges from these studies is the following: student expectations of the instructor, including expectations based on gender-role beliefs, play a significant role in student evaluations” (p. 217). Referring particularly to the study by Statham et al. (1991), they report that instructors tend to be rewarded if they fit students’ gender-role expectations in their teaching style but are punished if they do not fit these expectations, concluding that “students appear to evaluate ‘likability’ and ‘competence’ for men and women on somewhat different bases” (p. 218).

Linse (2003) provided a useful and concise outline of the literature on gender bias in teacher evaluations as it relates specifically to female academics. This work summarizes the complexities of the earlier literature and the collectively ambiguous findings. However, it also pointed out that one shortcoming of many quantitative studies that compared student ratings of male and female faculty is that the evaluations themselves did not account for *student* gender, assuming that there would be no differences between male and female student ratings. Similar to Centra and Gaubatz (2000), Linse also pointed out that “[s]tudents in certain fields may require women faculty to meet more stringent credibility criteria by virtue of their sex” (p. 4) and concluded that faculty should ensure that student evaluations are never used as the sole measure of teacher effectiveness.

In summary, then, the literature examining student evaluations of their instructors has found mixed results with respect to gender bias. Although some have found same-sex biases (e.g., the general analysis of Centra & Gaubatz, 2000), some have uncovered uniform bias against female teachers (e.g., Basow & Silberg, 1987; Burns-Glover & Veith, 1995; Sidanius & Crane, 1989), whereas others, primarily the papers that have attempted to provide meta-analyses, have concluded that teacher gender does not play a relevant role in teacher evaluations (e.g., Cashin, 1995; Feldman, 1992, 1993). Furthermore, in some studies, gender biasing occurs for certain evaluation items but not for others. So it remains unclear whether gender differences measured in various studies can be accounted for by

differences in teaching style or effectiveness (Anderson & Miller, 1997; Centra & Gaubatz, 2000; Linse, 2003).

Another question that has not been sufficiently answered is that of a discipline-specific effect with respect to gender bias in teacher evaluations. This is a particularly relevant question for the natural sciences in which it is well known that, across science disciplines, disparities in gender representation are not consistent: The life sciences have a greater representation of females, whereas the physical sciences sorely lack females. Some works have looked at specific fields: For example, as discussed above, part of the analysis in Centra and Gaubatz (2000) devoted itself to natural science courses, Miller and Chamberlin (2000) considered sociology students, and some works have looked at various social science fields (see Feldman, 1993). But lacking is a consideration of student evaluations of teachers in physical and life science disciplines. Many more studies have tended to aggregate teacher evaluations across disciplines and focus instead on separate evaluation measures and associated gender effects. Consequently, they have found varying degrees of gender bias depending on the questions asked and the measures considered but have little to say about discipline-specific student bias.

## RESEARCH QUESTIONS

In the current paper, we would like to answer the following questions:

1. Are there gender differences in student evaluations of high school teachers in three different science fields: biology, chemistry, and physics?
- 2a. If such differences are found, can teaching style account for the gender differences in teacher evaluations?
- 2b. Again, if such differences are found, can they be explained by a genuine disparity in the effectiveness of male and female teachers to prepare their students?

In the absence of gender bias, male and female students evaluating male and female high school teachers will produce ratings that are, on average, not significantly different (after controlling for any differences in instructors' teaching style and effectiveness).

The current study explores student perceptions of gender-based efficacy through their evaluations of the people in science with whom they have had daily contact: their high school science teachers. Our framework of analysis is similar to the work on younger students' perceptions of gender appropriateness in science, such as Farenga and Joyce (1999); it is our argument that the gender-based differentials that are expressed through the evaluations of teachers in this analysis offer insight into the personal views of the participants regarding gender-based competency in these fields. That is, to say, the external process of evaluating teachers reflects, in part, the evaluator's internal attitudes toward the appropriateness and competency of males and females in science.

The importance of bias in student evaluations of teachers is that it reflects much larger issues in science regarding a discipline's culture, climate, and gender stereotypes. It is obvious that any bias in teaching evaluations is important at the college level, where student evaluations are considered in tenure and promotion decisions, or in any situation in which student evaluations are used as indicators of teacher performance (including the internal effects on teachers regarding their competency). But science educators must also consider the effect these biases will have on female students (Kessels, 2005). As found by Selimbegovic et al. (2007), Andre et al. (1999), Meece et al. (2006), Cleaves (2005), and DeBacker and Nelson (1999), negative stereotypes about gender may influence female students to lose interest in science study and/or lower their self-competence in science.

As gender-based performance differentials have largely disappeared in high school subjects (Hyde, Lindberg, Linn, Ellis, & Williams, 2008), researchers have been forced to consider other reasons for the lingering gap in female persistence in the physical sciences. The attitudes uncovered in the current analysis may highlight another source of this problem.

## STUDY AND DATA

The data used in this study came from the “Factors Influencing College Science Success” (FICSS) study, a large-scale survey of students in introductory biology, chemistry, and physics courses at randomly selected colleges across the United States.<sup>1</sup> The three versions of this survey focused on students’ high school biology, chemistry, or physics class experiences, and the primary intent of the survey was to determine how these experiences influence success in college science courses (Hazari, Tai, & Sadler, 2007). The questionnaires were based on a previous pilot survey of 2000 college physics students conducted in 1994 (Sadler & Tai, 2001) plus a series of interviews with 20 high school science teachers and 22 introductory college science professors that focused on the factors that influence student success in college science. The survey’s validity was determined through focus group interviews with college science students as well as further consultation with high school science teachers and college professors. This feedback led to the rewriting of some survey items. In order to assess the degree of reliability in participant responses to the survey questions, we performed a test–retest study in which the same 113 college chemistry students completed the survey 2 weeks apart and their responses were compared. In an analysis of the results, on any given individual survey item, at least 90.7% of the responses were within one choice of the original response with 60% of the responses identical. This result translates into reliability coefficients ranging between 0.5 and 0.7 for the various items in the survey. According to Thorndike (1997), in an analysis of groups of 100 participants, a reliability coefficient of 0.5 corresponds to a 0.04% likelihood of a reversal in the direction of difference. This means that the survey instrument has a high degree of reliability. As discussed in Thorndike (1997), the reason for such a strong reliability stems from the sample size: Although the responses of any given individual may vary, overall trends found in large groups tend to be quite stable.

The employed methodology applies a cross-sectional approach relying upon the natural variation in the experiences and background of the sample students. The FICSS project used a representative, stratified, random sample taken from a comprehensive list of 4-year colleges and universities in the United States. The stratification accounted for the size of the institution and prevented an over-sampling of the smaller, but numerous, liberal arts colleges in the United States. In total, 63 of these colleges agreed to participate in the survey. From the participating institutions, the courses that were surveyed consisted of biology, chemistry, and physics courses that satisfied first-year requirements for science majors. The survey was administered to students during the fall semester. At the end of the term, professors reported each student’s final grade.

As mentioned, the primary intent of the survey was to understand how various high school science experiences influence college science success. The final version of the

<sup>1</sup> Project FICSS was supported by the Interagency Educational Research Initiative (NSF-REC 0115649). Any opinions, findings and conclusions, or recommendations expressed do not necessarily reflect the views of the NSF, the U.S. Department of Education or the NIH.

## 8 POTVIN ET AL.

survey included 66 questions about student demographics, earlier mathematics and science academic achievement, the pedagogies employed in students' last high school science course (in the same discipline as their college enrollment), and a section for the college professor to enter the student's final grade. (For a complete sample survey, please visit <http://www.ficss.org> and click on "About the Research.") In addition to these topics, the survey included a nine-item evaluation of the quality of students' last high school science teacher in the science discipline of their survey. These items took the following form:

How would you rate your high school physics teacher on the following characteristics?

- Knowledge of [biology/chemistry/physics]
- Enthusiasm for [biology/chemistry/physics]
- Fairness
- Pleasantness
- Ability to organize lessons and class activities
- Ability to explain problems in several different ways
- Ability to handle discipline and manage the classroom
- Ability to keep students on task during a lesson
- Ability to maintain students' interest during a lesson.

Note that these items were listed with a Likert-scale response running from 1 (low) to 5 (high). Following these, students were asked to indicate their teacher's gender.

Accounting for nonresponses on all of the variables included in the analysis, a sample of 6994 students were represented in the final models; these included 1405 females and 953 males in biology (60% female), 1652 females and 1378 males in chemistry (55% female), and 612 females and 994 males in physics (38% female). To estimate the number of high school teachers that were being evaluated, we employed a conservative method using students' home zip code as a proxy. For each discipline, if two or more students reported that they came from the same home zip code, we assumed that they had the *same* last high school teacher (unless they reported different teacher genders). This likely results in an underestimate of the number of teachers being evaluated in our data, but provides a sensible baseline. Applying this formula to the data, we estimate that the number of teachers being evaluated in our analysis is, at least, 1640 in biology (47% of which are female), 1847 in chemistry (48% female), and 1106 in physics (28% female).

Although all of the classes that participated in the survey were introductory-level, the sample of students consisted of 49.4% freshmen, 27.8% sophomores, 16.5% juniors, and 6.3% seniors. Considering the fact that students were asked to respond to the survey based on the last high school science course in the discipline of their college enrollment, one might worry whether the variable gap of time between their last high school course and the FICSS survey might have an effect on responses. However, throughout all of the analysis that follows, student year of enrollment was considered and is not significant with respect to the variables under consideration. In other words, the responses by sophomores, juniors, and seniors are not significantly different from those of freshman. This indicates that student responses are not significantly different if the time increases between a student's last high school science course and their introductory college course. In addition to this, in the questionnaire, students who had not taken a high school science course in their college-enrolled discipline (such students appeared almost entirely in the physics sample) were instructed to skip over the sections of the FICSS survey that dealt with their high school classroom experiences (including teacher evaluations). In the current analysis, these respondents are completely excluded.

## ANALYSIS AND FINDINGS

A principal component analysis performed on the above-mentioned teacher evaluation items showed that, in each discipline, these items loaded together. Specifically, it was found that 69%, 68%, and 68% of the variance in the nine teacher characteristics was subsumed by a single factor for each of the biology, chemistry, and physics data sets, respectively. Adding additional factors explained less than an additional 10% of the variance in each case. Thus, a one-factor solution was the only possibility, which represents a measure of the overall quality of the high school teacher from the perspective of the student.

Using these items, a 100-point “teacher evaluation score” outcome variable was constructed that weighed each of the nine points equally. The average score and standard deviation in each discipline were  $71.3 \pm 24.6$  in biology,  $70.5 \pm 25.1$  in chemistry, and  $72.0 \pm 24.4$  in physics. Then, for each of the three subject areas, multiple linear regression models of the various student-level predictors of the teacher evaluation score were constructed. Hierarchical linear models were first considered for these data; however, in all three disciplines, institutional-level effects were non-significant. This is an intuitive finding: Since the teacher evaluations are based on students’ experiences in high school, we would expect that each institution contains a similar sample of students in terms of their high school teacher experience; the variability is found at the individual (student/teacher) level. Note also that, in this analysis, all dichotomous and other categorical responses such as student gender, teacher gender, etc., were dummy coded to properly account for non-responders to these questions. Table 1 displays the results of the multiple regression models, and the estimated coefficients produced from these analyses are summarized in Figure 1, separately for each field. Note that all effect sizes are in units of the teacher evaluation score (out of 100) and that the minimum significance level that is reported throughout this study is  $\alpha = .05$ .

First, note that, in all cases, male and female students’ ratings of male teachers are statistically equivalent (each pair of ratings is within standard errors of each other), which suggests that the current analysis is consistent across disciplines. Also, as might be expected, the students’ grade in their high school course is a significant predictor ( $p < .001$ ) of their evaluation of their teacher in all three datasets; hence, it was controlled for in all cases. Third, the factors measuring students’ racial/ethnic and socioeconomic background were non-significant predictors of the teacher evaluation score and were thus excluded from these models.

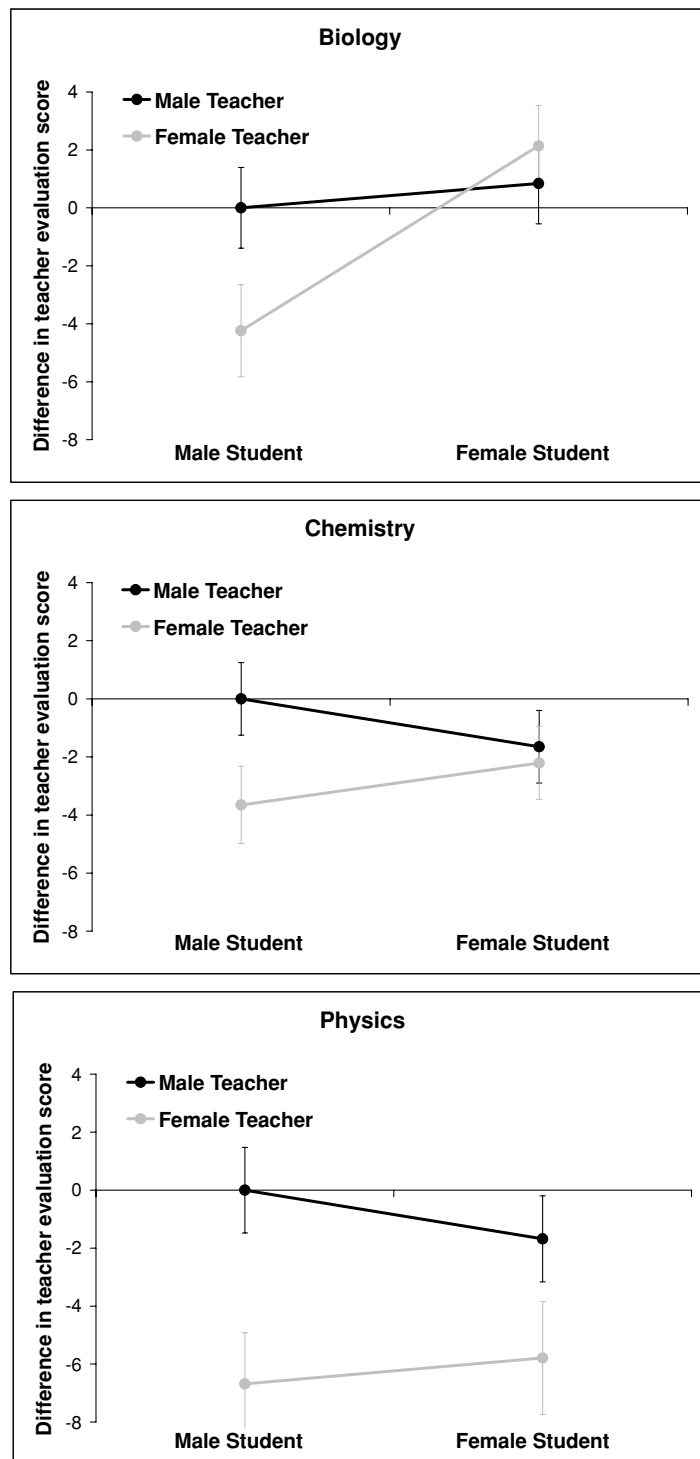
Overall, Table 1 shows that the teacher’s gender is a significant predictor of evaluation scores in these models, at the  $p < .01$  level in biology ( $B = -4.24$ ) and chemistry ( $B = -3.65$ ) and at the  $p < .001$  level in physics ( $B = -6.68$ ). Also, there is a significant interaction between the gender of the teacher and the gender of the student in the case of biology ( $B = 5.54$ ,  $p < .01$ ), although the same interaction has borderline significance in chemistry as well ( $p = .08$ ); the models of biology and chemistry are qualitatively similar in structure.

To make the meaning of these models clear, consider the evaluation of female teachers separately from that of male teachers in each discipline, as in Figure 1. In Figure 1, we have indicated separately the average teacher evaluation score assigned to male and female teachers by male and female students; the scores assigned by male students appear to the left, whereas those assigned by female students appear to the right. The lines connecting each point are a graphical reference for comparing teacher groups. In the absence of any bias, we would expect each line to have zero slope (indicating that both male and female students give the same score to each teacher group) and would completely overlap (indicating that

**TABLE 1**  
**Multiple Regression Models of Teacher Evaluation Score as Predicted by Teacher and Student Gender**

Predictors	Biology (N = 2358)			Chemistry (N = 3030)			Physics (N = 1606)		
	B	SE	Significance	B	SE	Significance	B	SE	Significance
Intercept	48.64	3.63	***	43.38	2.76	***	57.53	4.18	***
Final grade	5.00	0.78	***	6.62	0.60	***	3.69	0.91	***
Student gender	0.84	1.39	ns	-1.65	1.25	ns	-1.68	1.48	ns
Teacher gender	-4.24	1.59	**	-3.65	1.33	**	-6.68	1.76	***
Student × teacher gender	5.54	2.05	**	3.09	1.80	ns	2.57	2.74	ns

Note. ns, not significant. \*  $p < .05$ . \*\*  $p < .01$ . \*\*\*  $p < .001$ .



**Figure 1.** Average teacher evaluation score, partitioned by teacher and student gender. The lines connecting each point are a graphical reference for comparing teacher groups: in the absence of any bias, we would expect that each line would have zero slope and that each pair of lines would completely overlap. Note that the location of the horizontal baseline is arbitrary: We chose to set it to the percentile score given to male teachers by male students.

both teacher groups receive the same average score). The degree to which the lines diverge from this ideal is a graphical indicator of the level of bias.

Male students consistently rate female teachers significantly lower than male teachers in all cases (4.24 points lower in biology, 3.65 in chemistry), with the largest differential of 6.68 occurring in physics. On the other hand, female students in biology rate female biology teachers with a small positive coefficient (+2.14) that is statistically equivalent to their rating of male biology teachers (+0.84). Similarly, in chemistry, female students rate female teachers with a small negative coefficient (−2.21) that is, again, statistically equivalent with their rating of male chemistry teachers (−1.65). Most interestingly, the physics data show that female physics students rate their female teachers with a larger negative coefficient (−5.79), which is statistically equivalent to male student ratings of female physics teachers (−6.68), but *not* equivalent to the evaluation of male teachers by female students (−1.68), as was the case in biology and chemistry.

In summary, the models indicate that male students report significantly lower ratings for female science teachers in all three disciplines, whereas female students report significantly lower ratings for female science teachers only in physics. In all cases, male and female students agree on their rating of male science teachers. This suggests that there is some biasing that occurs in each discipline, although the nature of this bias appears to be different in biology and chemistry as opposed to the bias in physics: In the former two cases, only male students exhibit this differential behavior whereas, in the latter, all students exhibit it.

### DOES TEACHING STYLE ACCOUNT FOR THE GENDER DIFFERENCES IN TEACHER EVALUATIONS?

One possible alternative explanation of the gender effects that appear in these models is that, on average, male and female teachers may have different classroom teaching styles, which have different levels of popularity among students, leading to differences in their evaluations. By considering the numerous questions regarding student classroom experiences that also appeared in the FICSS survey, this possibility can be tested. This section will examine more comprehensive models that account for a wide range of factors that predict the teacher evaluation score from students' high school science experiences in addition to gender, academic performance, and outside classroom (e.g., family attitude, socioeconomic status (SES), etc.) predictors. Note that all nonsignificant predictors were removed from the comprehensive models. Only the predictors that are relevant to the current discussion (e.g., student gender and/or student–teacher interaction) are listed as nonsignificant where appropriate.

In the case of the biology data, even controlling for many pedagogical factors, both the teacher gender and student–teacher interaction predictors continue to be significant at the  $p < .01$  level, and the entire model accounts for 33.2% of the variance (measured by the adjusted  $R^2$ ). Table 2 shows that there are a number of significant predictors of teacher evaluations beyond the student's final grade (which continues to be significant,  $p < .001$ ) that have been taken into account in this more complete model. In terms of classroom experiences, the significant predictors are the frequency of the teacher lecturing to the whole class ( $p < .001$ ), the frequency of whole class discussions ( $p < .001$ ), and the course focus being “understanding” (rather than “memorization”) ( $p < .001$ ). Students who report that they had no coverage of cell biology ( $p < .001$ ) or ecology ( $p < .05$ ) rated their teachers significantly lower; similarly, students who experienced “recurring” coverage of the history of science rated their teachers higher ( $p < .01$ ) as well as those students who experienced more frequent use of real-world examples in their courses ( $p < .001$ ). Overall, students who reported more frequent use of demonstrations rate their

**TABLE 2**  
**Comprehensive Model for Biology Data ( $N = 2358$ )**

Predictors	$B$	SE	$\beta$	Significance
Intercept	19.76	3.68	–	***
Final grade	2.82	0.66	0.07	***
Student gender	–	–	–	ns
Teacher gender	–3.60	1.15	–0.07	**
Student $\times$ teacher gender	3.68	1.26	0.07	**
Frequency of teacher lecturing the class	0.06	0.02	0.07	***
Frequency of whole class discussions	0.07	0.01	0.09	***
Course focus: memorization or understanding?	5.03	0.42	0.22	***
No coverage of cell biology	–10.4	2.58	–0.07	***
No coverage of ecology	–3.12	1.28	–0.04	*
Recurrent coverage of the history of science	3.85	1.37	0.05	**
Frequency of everyday world examples	0.15	0.01	0.20	***
Demonstrations: number per week	1.85	0.37	0.09	***
Laboratories: built upon previous laboratories	2.39	0.42	0.11	***
Laboratories, preparation: read and discussed in prior class	3.17	1.02	0.06	**
Laboratories, preparation: discussed laboratory in class	4.30	0.87	0.09	***
Test questions referred to laboratories	3.13	0.91	0.06	**
Encouragement to study science from science teacher	3.98	1.07	0.07	***
Encouragement to study science from other teacher	–7.51	1.98	–0.07	***
Family: highest level of parental education	–1.41	0.40	–0.06	***
Family: English as primary language	5.72	1.62	0.06	***
Family attitudes: science is for a better career	–2.42	1.04	–0.04	*

Note. ns, not significant. \* $p < .05$ . \*\* $p < .01$ . \*\*\* $p < .001$ .

teachers higher ( $p < .001$ ). The predictors from the students' laboratory experiences that influenced their evaluation of their teachers are as follows: laboratories that were built upon previous laboratories ( $p < .001$ ); students read and discussed the laboratory in a class before carrying it out ( $p < .01$ ); students read and discussed the laboratory in the same class that they carried it out ( $p < .001$ ); and having test questions that referred to laboratories ( $p < .01$ ). Students who reported that their biology teacher encouraged them to study science rated those teachers significantly higher ( $p < .001$ ), whereas students who reported that *another* teacher encouraged them rated their biology teachers significantly lower ( $p < .001$ ). Finally, the highest level of education of students' parents was the only SES control that was significant ( $p < .001$ ); those students who speak English as a first language rated their teachers more highly ( $p < .001$ ), whereas those students who report that their family's attitudes toward science was that "science is for a better career" is a significant predictor of their teacher's evaluation ( $p < .05$ ).

Table 3 shows that, for chemistry, teacher gender continues to be a significant predictor, albeit at the  $p < .05$  level, the weakest gender effect in any of the models, as well as student final grade ( $p < .001$ ). The complete model accounts for 33.6% of the variance, as measured by the adjusted  $R^2$ . In this case, there are no family, SES, or encouragement predictors that are significant, but a number of classroom/pedagogical predictors are relevant. As in

**TABLE 3**  
**Comprehensive Model for Chemistry Data ( $N = 3030$ )**

Predictors	$B$	SE	$\beta$	Significance
Intercept	22.84	3.78	–	***
Final grade	4.00	0.53	0.12	***
SAT verbal score	–0.01	0.004	–0.05	**
Student gender	–	–	–	ns
Teacher gender	–1.66	0.75	–0.03	*
Student*teacher gender	–	–	–	ns
Frequency of teacher lecturing the class	0.04	0.02	0.04	*
Frequency of whole class discussions	0.06	0.01	0.07	***
Number of topics covered: depth or breadth?	–2.05	0.43	–0.07	***
Course focus: memorization or understanding?	5.30	0.38	0.23	***
No coverage of solutions	–8.65	2.32	–0.06	***
No coverage of stoichiometry	–3.65	1.22	–0.05	**
Frequency of individual student work in class	0.04	0.01	0.04	**
Importance of the textbook	0.97	0.33	0.05	**
Frequency of everyday world examples	0.11	0.02	0.12	***
Problem solving: problems with calculations assigned	0.15	0.03	0.07	***
Demonstrations: number per week	2.08	0.34	0.10	***
Demonstrations: length of discussion afterwards	0.15	0.05	0.05	**
Laboratories: frequency per month	1.17	0.28	0.07	***
Laboratories: laboratories addressed real world beliefs	1.51	0.40	0.06	***
Laboratories: length of discussion afterwards	0.12	0.03	0.06	***
Laboratories, preparation: discussed laboratory in class	3.44	0.77	0.07	***

Note. ns, not significant. \* $p < .05$ . \*\* $p < .01$ . \*\*\* $p < .001$ .

biology, the frequency of the teacher lecturing the whole class ( $p < .05$ ), the frequency of whole-class discussions ( $p < .001$ ), and the course focus on understanding (as opposed to memorization) are significant ( $p < .001$ ). In addition, the amount of content topics in the course (e.g., depth of coverage vs. breadth of coverage) is a significant predictor ( $p < .001$ ), as are the importance of the textbook ( $p < .01$ ) and the use of problems (in homework and class work) that required calculations ( $p < .001$ ). Those students who report that they spent no time on the subjects of solutions ( $p < .001$ ) or stoichiometry ( $p < .01$ ) rated their teachers significantly lower. As before, the use of real-world examples is a significant predictor ( $p < .001$ ), but in this case, the amount of individual student work in class is also significant ( $p < .01$ ). Again, the number of demonstrations per week has an influence ( $p < .001$ ) over teacher evaluations as does the length of discussion after demonstrations ( $p < .01$ ). A number of factors related to laboratory experiences have a positive effect on teacher evaluations: a greater number of laboratories per month ( $p < .001$ ), reading and discussing laboratories in the same class as they were carried out ( $p < .001$ ), more laboratories that explored real-world beliefs ( $p < .001$ ), and a longer period of discussion afterward ( $p < .001$ ).

The complete model in Table 4 for the physics data accounts for 36.0% of the variance, and teacher gender continues to be a significant predictor of the teacher evaluation score, now at the  $p < .01$  level. Furthermore, of the three comprehensive models, this one accounts

**TABLE 4**  
**Comprehensive Model for Physics Data ( $N = 1606$ )**

Predictors	$B$	SE	$\beta$	Significance
Intercept	31.23	4.35	–	***
Final grade	2.05	0.75	0.06	**
One AP science course taken	3.33	1.43	0.05	*
Student gender	–	–	–	ns
Teacher gender	–3.22	1.09	–0.06	**
Student $\times$ teacher gender	–	–	–	ns
Number of topics covered: depth or breadth?	–1.97	0.55	–0.08	***
Course focus: memorization or understanding?	4.92	0.49	0.22	***
No coverage of heat and kinetic theory	–6.36	1.80	–0.07	***
Frequency of everyday world examples	0.16	0.02	0.21	***
Tests: questions could be solved without mathematics	3.83	0.99	0.08	***
Problem solving: problems with calculations assigned	0.16	0.05	0.08	***
Demonstrations: number per week	2.84	0.40	0.16	***
Demonstrations: length of discussion beforehand	0.24	0.07	0.07	***
Laboratories: built upon previous laboratories	2.30	0.50	0.10	***
Laboratories, preparation: read directions while doing laboratory	–2.48	1.01	–0.05	*
Laboratories, preparation: discussed laboratory in class	4.11	1.03	0.08	***
Family attitudes: science is a series of courses to pass	–3.54	1.29	–0.06	**

Note. ns, not significant. \* $p < .05$ . \*\* $p < .01$ . \*\*\* $p < .001$ .

for the most variance, although it has the fewest number of predictors appearing in it. In addition to final grade ( $p < .01$ ), if students reported that they took any advanced placement (AP) science course, they evaluated their teachers higher ( $p < .05$ ). Several significant predictors are held commonly with the chemistry data: the number of topics covered (depth vs. breadth) ( $p < .001$ ), the course focus on understanding (rather than memorization) ( $p < .001$ ), the frequent use of everyday-world examples ( $p < .001$ ), the use of problems that required calculations ( $p < .001$ ), and the number of demonstrations used in class ( $p < .001$ ). In addition, the length of discussion before demonstrations is significant in this case ( $p < .001$ ), as well as the use of test questions that could be solved without mathematics ( $p < .001$ ). Students who reported that they had no coverage of heat and kinetic theory rated their teachers significantly lower ( $p < .001$ ). In terms of laboratory experiences, laboratories that built upon previous experiences ( $p < .001$ ) and having read and discussed the laboratories in the class in which they were carried out are positive predictors of teacher evaluations ( $p < .001$ ), but students who reported that they “read the directions while doing the laboratory” rated their teacher lower ( $p < .05$ ). Finally, one family attitude factor (“science is a series of courses to pass”) is a significant, negative predictor of teacher evaluations ( $p < .01$ ).

To repeat, it was found in each discipline that the gender effects persist even when controlling for many classroom experiences. Such factors that proved to be significant predictors of teacher evaluations are interesting in their own right; however, as they did not explain the gender gap in teacher evaluations, the basic gender results remain qualitatively the same. Namely, in each field, male students rate female teachers lower than their male counterparts, whereas female students’ ratings behave similarly only in physics.

**TABLE 5**  
**Multiple Regression Models Showing the Gender of Students' Last High School Teacher Does Not Predict Their College Science Grade**

Predictors	Biology ( <i>N</i> = 2358)			Chemistry ( <i>N</i> = 3030)			Physics ( <i>N</i> = 1606)		
	<i>B</i>	SE	Significance	<i>B</i>	SE	Significance	<i>B</i>	SE	Significance
Intercept	80.39	0.85	***	79.98	0.82	***	82.69	0.71	***
Teacher gender	-0.15	0.46	ns	-0.12	0.40	ns	-0.76	0.61	ns

Note. ns, not significant. \* $p < .05$ . \*\* $p < .01$ . \*\*\* $p < .001$ .

### ARE MALE AND FEMALE TEACHERS EQUALLY EFFECTIVE AT PREPARING THEIR STUDENTS?

Despite the preceding discussion that showed that differences in teaching style did not account for the differences in the evaluations of male and female teachers, one might continue to worry that the gap in teaching evaluations stems from a genuine disparity in the teaching *effectiveness* of male and female teachers. However, there is significant evidence in this data set that male and female teachers are equally effective at preparing students for college science study. First, in all three disciplines, the gender of the teacher does not predict later college performance in the subject area of study (see Table 5). This remains true whether or not one controls for student SES, their previous academic performance (e.g., SAT verbal and mathematics scores, previous grades in English and mathematics, etc.), and previous academic experiences (e.g., type of high school attended, higher-level mathematics, and science enrollment). So this suggests that the students of male and female high school teachers are equally well prepared for later college science study.

Second, there is further independent evidence that female teachers are equally effective at preparing students for college science: The students of female high school teachers persist in college science studies at the same rate as the students of male high school teachers. Hypothetically, if male teachers were more effectively preparing their students for, and generating interest in, college science (in ways that are not captured simply by students' college grades, as discussed above), we might expect the *fraction* of students who had male teachers in high school to increase in college enrollment numbers. That is to say, in such a situation, the students of female high school teachers might become underrepresented in college science courses, possibly due to a differential decline in their interest in college science or due to these students somehow failing to meet college entrance requirements (grades notwithstanding). However, the data show us that this is not the case: national estimates on the fraction of female high school science teachers are 50% in biology, 46% in chemistry, and 29% in physics (Blank & Langesen, 2001). In the current sample, the fraction of high school science teachers who are female, as reported by our (successfully enrolled) college students, is in excellent agreement with the national estimates: 47% in biology, 48% in chemistry, and 28% in physics, respectively. Thus, in our sample, students of female teachers continue studying science in college at the same rate as students of male teachers. This is another argument for the equal effectiveness of male and female science teachers, in both their ability to prepare students for college science (and any hurdles to entrance into college) and their ability to keep students *interested* enough to continue their studies in science when they get to college.

Thus, we have seen that (a) there is no significant difference in the college performance of students based on the gender of their high school science teacher and (b) there is no effect of teacher gender on rates of college science persistence. Since all of these results are based on reports of a nationally representative sample of nearly 7000 college science students on their high school science teachers, it follows that the general findings of a disparity in the evaluations of teachers are consistent with a claim of gender biasing against female teachers.

## DISCUSSION AND IMPLICATIONS

Gender equity [in education] is the elimination of sex role stereotyping and sex bias from the educational process, thus providing the opportunity and environment to validate and empower individuals as they make appropriate career and life choices. (Hilke & Conway-Gerhardt, 1994, p. 8)

Bias is a difficult construct to measure, particularly since openly stated bias is often taken as an affront. As a result, proxies are often employed, as in this case where introductory college science students' evaluations of their former high school science teachers were used. Our results offer evidence that gender biasing exists among male students (and some females) against female teachers in science at the end of high school and in disciplines that have seen enormous growth in the participation of females at all levels. The gender effects persisted when a variety of classroom experience predictors were added to try to explain differences between male and female teacher evaluations. These results are also internally consistent: A statistically equivalent rating of male teachers was replicated across three different samples in three different disciplines, which suggests that both male and female students agreed in their evaluation of male teachers. In terms of the nature of the bias uncovered, our results in physics are reminiscent of the work by Basow and Silberg (1987), Sidanius and Crane (1989), and Burns-Glover and Veith (1995), which found bias against female teachers among all students. Our results from biology and chemistry are also generally consistent with the finding from other studies that indicated that male students may hold gender-related stereotypes more strongly than female students (Andre et al., 1999; Burns-Glover & Veith, 1995).

Furthermore, our results have implications for some of the outstanding questions about student evaluations highlighted by the literature. In particular, we have seen that each field in our study seems to have a certain bias that is peculiar to that field; this implies that attempts to analyze bias (such as Cashin, 1995; Feldman, 1992, 1993) in teacher evaluations across disciplines will fail to understand the nature and depth of the problem. The fact that we did not uncover uniform bias across all three science disciplines suggests that students have internalized specific "cultural" gender stereotypes in each case (Centra & Gaubatz, 2000; Miller & Chamberlin, 2000). Our results, therefore, mean that generalizations across disciplines will be difficult to make sensibly, as argued by Miller and Chamberlin.

The biases we have uncovered are important because they have a significant impact on female students. The work of Andre et al. (1999), Meece et al. (2006), and Selimbegovic et al. (2007) implies that the internalization of these stereotypes will lead females to have depressed senses of self-competency and reduce their likelihood of choosing to study science further. This is most significant in physics, in which gender differentials were the largest and is also the field in our study that has the lowest representation of females. Using the connection made by DeBacker and Nelson (1999), Cleaves (2005), and Farenga and Joyce (1999), we also argue that these two facts are connected: gender bias at the high school level will contribute to females leaving physics study in college (and beyond).

Our findings are similar, then, to the results of the stereotype threat literature (Spencer et al., 1999; Steele, 1997). These papers found that student *performance* can be depressed if a stereotype is introduced into the classroom. Although there are no significant performance differences between male and female students in our data (see also Hazari et al., 2007), our results show that the underlying bias can impact other domains as well. In this case, female students internalize this bias, which impacts their attitude toward science study after high school (particularly for physics), leading to a decline in their later participation. Note also that the evidence we presented on teacher effectiveness strongly suggests that male and female teachers are equally competent, so there does not seem to be any performance drop of female *teachers* related to stereotype threat presented by students. This may be because students are not able to threaten teachers or that high school teachers do not have sufficient performance anxiety toward the outcome (i.e., teachers preparing students well) for the stereotype threat to manifest itself (Aronson et al., 1999; Steele, 1997).

The existence of these gender biases among our sample population presents a very real problem. The students in our sample are those who have successfully enrolled in college science programs and are likely to form a critical source for the scientific workforce of the future. The fact that these students have internalized gender stereotypes in their judgment of others has negative implications for their female peers and subordinates if and when they become successful scientists. In other words, these attitudes contribute to the chilly climate, as characterized by Urry (2003): “in physics departments around the country, women are feeling ill at ease, out of place, not at home” (p. 12).

As the United States begins to seek out means for enhancing scientific workforce development among its citizens, gender bias may be partially eroding the effectiveness of these efforts. More work should be done to parse out these gender biases and to understand more precisely how they are manifested in an individual’s perception of the competency of others and, indeed, herself or himself. Furthermore, efforts should be made to explicitly combat these biases in each field: our results suggest that each field should recognize its own gender bias and should consider specific ways to combat it in the training of future scientists and the general public.

We thank A. Trenga, M. Filisky, J. Loehr, B. Ward, J. Peritz, F. Deutsch, C. Crockett, the other members of the FICSS team, and all of the professor and student participants for making this study possible. We also thank the National Science Foundation (NSF) for their support and G. Sonnert, J. Miller, C. Bowman, and A. Baleisis for helpful discussions and feedback on this work.

## REFERENCES

- Anderson, K., & Miller, E. D. (1997). Gender and student evaluations of teaching. *Political Science & Politics*, 30, 216–219.
- Andre, T., Whigham, M., Hendrickson, M., & Chambers, S. (1999). Competency beliefs, positive affect, and gender stereotypes of elementary students and their parents about science versus other school subjects. *Journal of Research in Science Teaching*, 36(6), 719–747.
- Aronson, J., Lustina, M. J., Good, C., Keough, K., Steele, C. M., & Brown, J. (1999). When White men can’t do math: Necessary and sufficient factors in stereotype threat. *Journal of Experimental Social Psychology*, 35, 29–46.
- Baker, D., & Leary, R. (2003). Letting girls speak out about science. *Journal of Research in Science Teaching*, 40(Suppl.), S176–S200.
- Barman, C. (1997). Students’ views of scientists and science: Results from a national study. *Science and Children*, 35(1), 18–23.
- Basow, S. A., & Silberg, N. T. (1987). Student evaluations of college professors: Are female and male professors rated differently? *Journal of Educational Psychology*, 79(3), 308–314.
- Blank, R. K., & Langesen, D. (2001). State indicators of science and mathematics education 2001 (p. 142). Washington, DC: Council of Chief State School Officers.

- Burns-Glover, A. L., & Veith, D. J. (1995). Revisiting gender and teaching evaluations: Sex still makes a difference. *Journal of Social Behavior & Personality*, 10 (6, Special issue: Gender in the Workplace), 69–80.
- Cashin, W. E. (1995). Student ratings of teaching: The research revisited. IDEA paper No. 32, Center for Faculty Evaluation and Development, Kansas State University, Manhattan, KS, September 1995.
- Centra, J. A., & Gaubatz, N. B. (2000). Is there gender bias in student evaluations of teaching? *Journal of Higher Education*, 71(1), 17–33.
- Cleaves, A. (2005). The formation of science choices in secondary school. *International Journal of Science Education*, 27(4), 471–486.
- Dar-Nimrod, I., & Heine, S. J. (2006). Exposure to scientific theories affects women's math performance. *Science*, 314, 435.
- Dawson, C. (2000). Upper primary boys' and girls' interests in science: Have they changed since 1980? *International Journal of Science Education*, 22(6), 557–570.
- DeBacker, T. K., & Nelson, R. M. (1999). Variations on an expectancy-value model in science. *Contemporary Educational Psychology*, 24, 71–94.
- Farenga, S. J., & Joyce, B. A. (1999). Intentions of young students to enroll in science courses in the future: An examination of gender differences. *Science Education*, 83(1), 55–75.
- Feldman, K. A. (1992). College students' views of male and female college teachers: Part I—Evidence from the social laboratory and experiments. *Research in Higher Education*, 33(3), 317–375.
- Feldman, K. A. (1993). College students' views of male and female college teachers: Part II—Evidence from students' evaluations of their classroom teachers. *Research in Higher Education*, 34(2), 151–211.
- Hazari, Z., Tai, R., & Sadler, P. (2007). Gender differences in introductory university physics performance: The influence of H. S. physics preparation and affective factors. *Science Education*, 91(6), 847–876.
- Hilke, E. V., & Conway-Gerhardt, C. (1994). Gender equity in education (pp. 7–17). Bloomington, IN: Phi Delta Kappa Educational Foundation.
- Hollinger, C. L. (1991). Facilitating career development of young gifted women. *Roeper Review*, 13(3), 135–139.
- Hughes, G. (2001). Exploring the availability of student scientist identities within curriculum discourse. *Gender & Education*, 13(3), 275–290.
- Hyde, J. S., Lindberg, S. M., Linn, M. C., Ellis, A. B., & Williams, C. C. (2008). Gender similarities characterize math performance. *Science*, 321(5888), 494–495.
- Johnson, S. (1999). Discovering the potential of gifted girls: The biological and physical science interests of gifted kindergarten girls. *School Science and Mathematics*, 99(6), 302–312.
- Jones, G., Howe, A., & Rua, M. (2000). Gender difference in students' experiences, interests, and attitudes towards science and scientists. *Science Education*, 84(2), 180–192.
- Kahle, J., & Meece, J. (1994). Research on gender issues in the classroom. In D. Gable (Ed.), *Handbook of research on science teaching and learning* (pp. 542–557). New York: Macmillan.
- Kessels, U. (2005). Fitting into the stereotype: How gender-stereotyped perceptions of prototypic peers relate to liking for school subjects. *European Journal of Psychology of Education*, 20(3), 309–323.
- Lederman, M. (2003). Gender/InEquity in science education: A response. *Journal of Research in Science Teaching*, 40(6), 604–606.
- Linse, A. R. (2003). Student ratings of women faculty: Data and strategies. Retrieved July 31, 2008, from [http://www.engr.washington.edu/advance/resources/20030513-student\\_ratings\\_ds.pdf](http://www.engr.washington.edu/advance/resources/20030513-student_ratings_ds.pdf).
- Meece, J. L., Glienke, B. B., & Burg, S. (2006). Gender and motivation. *Journal of School Psychology*, 44, 351–373.
- Miller, J., & Chamberlin, M. (2000). Women are teachers, men are professors: A study of student perceptions. *Teaching Sociology*, 28(4), 283–298.
- Sadler, P. M., & Tai, R. H. (2001). Success in introductory college physics: The role of high school preparation. *Science Education*, 85(2), 111–136.
- Selimbegovic, L., Chatard, A., & Mugny, G. (2007). Can we encourage girls' mobility towards science-related careers? Disconfirming stereotype belief through expert influence. *European Journal of Psychology of Education*, 22(3), 275–290.
- Sidanius, J., & Crane, M. (1989). Job evaluation and gender: The case of university faculty. *Journal of Applied Social Psychology*, 19(2), 174–197.
- Spencer, S., Steele, C., & Quinn, D. (1999). Stereotype threat and women's math performance. *Journal of Experimental Social Psychology*, 35, 4–28.
- Statham, A., Richardson, I., & Cook, J. (1991). *Gender and University teaching: A negotiated difference*. Albany, NY: SUNY Press.
- Steele, C. (1997). A threat in the air. *American Psychologist*, 52(6), 613–629.
- Thorndike, R. M. (1997). *Measurement and evaluation in psychology and education* (6th ed., pp. 116–117). Upper Saddle River, NJ: Merrill.
- Urry, M. (2003). Speeding up the long slow path to change. *APS News*, 12(2), 12.