Chapter 12

Are Student Teaching Evaluations Holding Back Women and Minorities?

The Perils of "Doing" Gender and Race in the Classroom

Sylvia R. Lazos

Teaching is important. Among the traditional three main responsibilities of the professoriate—teaching, scholarship, and service—teaching is probably the most important from the public perspective. The Association of American Colleges and Universities has recently challenged its members to focus more on student learning and develop better ways to measure it (National Association of American Colleges and Universities 2006), citing the well-reported statistics that the United States is slowly slipping behind other industrial countries in student performance in math, science, and writing.[1] In addition, demographics have changed the student population and its educational needs. Increasingly a greater proportion of the student population is less ready for college. Legislators, faced with shrinking state budgets, have become more prone to scrutinize what highly paid, tenured faculty members do with their time and routinely insist that they do more or better teaching. In sum, the current political climate where universities are operating demands that administrators carefully monitor faculty teaching effectiveness.

While there are many ways to evaluate teaching, universities have come to rely widely on student evaluations. According to a 1993 survey conducted by management expert Peter Seldin, 86 percent of universities used student evaluations of teaching in decisions about faculty retention, tenure, promotion, and merit pay

1 The National Assessment of Educational Progress (NAEP) reports on nationally representative samples of student work in reading, mathematics, science, writing, US history, civics, geography, and the arts. NAEP has recently stated that the achievement of US students in grades four, eight, and twelve has been slipping as compared to that of other industrialized countries. For details see http://nces.ed.gov/nationsreportcard.

(Seldin 1993). The use of student evaluations grew rapidly from 1970 to the 1980s. Since 1985, many institutions have used the SIR, the Student Instructional Report, which was developed by educational assessment expert John A. Centra at the Education Testing Service in Princeton, New Jersey, which also administers SATs. Although thoughtful commentators have persistently proposed additional methods of evaluating teaching (Arreola 2000; Seldin 2004; Braskamp and Ory 1994), no other method approaches the popularity of student evaluations (Seldin 1999). These evaluations have the advantage of providing a summary number that purports to assess teaching efficacy. This makes comparisons among colleagues easier. And the supposed objectiveness of numbers washes away any possible ambiguities and complexities. This is why many university administrators continue to be enamored with student evaluations (Seldin 1999).

The professoriate has produced an avalanche of articles critiquing and defending student evaluations.[2] The criticisms are startling. Methodological questions start with the most basic one: what do student evaluations actually measure? Professors Harry Tagamori and Laurence Bishop have concluded that the questions on student evaluations are so ambiguous that you can't even determine what they are asking. Tagamori and Bishop examined a random sample of student evaluation forms and found that more than 90 percent contained questions or items "that were ambiguous, unclear, or vague; 76 percent contained subjectively stated items, and over 90 percent contained evaluation items that did not correlate with classroom teaching behavior" (74–75).

Another statistician, Professor Valen Johnson, assembled a massive data set of student evaluations at Duke University and concluded that there was a significant statistical link between a professor's goal of receiving positive student teaching evaluations and grade inflation (Johnson 2003). Indeed a student's expectation of what grade he or she will get in a class is a strong predictor of how positive the instructor's evaluations will be for the instructor (Marsh and Dunkin 1992).

These studies make a point that upon reflection should be intuitively obvious. Evaluations may not be measuring teaching effectiveness as much as they are capturing students' subjective reactions at the moment that they are being polled, and their opinions reflect their feelings and thoughts about a range of things: whether they like the professor, whether their expectations about the course were met or they felt unsettled (perhaps because the professor deviated from the syllabus); and how well they imagined they were performing in school and in the class. Even student gossip becomes part of the picture (Feldman 1989a). Psychometric expert Mark Shevlin, comments that "students are not trained in rating or psychometrics"; rather he concludes that the main basis of their "global evaluation" of competency is "lecturer charisma" (Shevlin 2000 403). Statistician and assessment expert Professor Kenneth Feldman observes that student "ratings are designed not so much to obtain objective descriptions of teachers and courses but to measure the subjective reactions of students to them" (1989a, 257).

If student evaluations are subjective, are they then also subjective about the race and gender of the instructor? To what extent will a student's reaction to a professor's gender and race influence his or her evaluation? If student evaluations

2 According to Peter Seldin (1993), more than fifteen thousand articles have been published on the subject. Another set of researchers observes that this is one of the most common topics of research in higher education (Theall and Franklin 1999).

systematically produce lower ratings for women and minorities, then they may well be inhibiting these teachers' professional advancement. Consider that in most liberal arts colleges, a candidate's rating of satisfactory and most often excellent teaching is often the principal criterion for granting tenure. Empirical evidence irrefutably establishes that women and minorities gain tenure at lower rates than their majority counterparts and earn less in merit increases as well (Curtis 2005).[3] Student evaluations may well be holding them back.

Certainly, anecdotally many women and minorities blame student evaluations as a principal reason why they have not been able to get a foothold in academia. *New York Times* special reporter Mark Oppenheimer recently detailed the cases of two women, one teaching African American theatre at Wesleyan and the other teaching feminist studies. In the case involving the African American studies professor, she was told that her contract would not be renewed unless she received "the top two ratings (Outstanding and Good) by at least 85 percent of the students in both your courses." Her student evaluations were 76 percent outstanding and good (meaning a handful of students failed to rate her at the top two levels), and she lost her job. Although she received rave feedback from many students, she believed that the few students who were uncomfortable with her discussion of race and gender issues in class were very negative in their evaluations. The article also reports the case of a journalist professor at Antioch College who focused on race and gender issues in her classes and research. Her tenure case was derailed because her department chair focused on forty-three handwritten comments in her student evaluations that accused her of "having a political agenda" and "supporting gay rights." Because she failed to be rated excellent in teaching, she was denied tenure (Oppenheimer 2008, 24).

Part II discusses the psychology and sociology literature that establishes with robust empiricism that gender and race influence the way women and minorities are viewed in the classroom. Unconscious bias, stereotypes, and assumptions about role appropriateness are the subjective parameters that students unconsciously carry in their heads and use to shape the way they perceive their women and minority professors. These professors must walk a narrow pathway to manifest their gender and race and balance their teaching goals; they must maintain their individual authenticity in the classroom and yet avoid alienating students who—even at this late date—may not have encountered a minority authority figure in a professional setting. In sum, women and minority professors' performance in the classroom is fraught with potential land mines that they must navigate on the way to tenure.

Teasing out just how gender and race impact student evaluations from the empirical studies is a complex task. Part III of this chapter details that the empirical data as to whether subjective gender and race bias exists in student evaluations are equivocal. Some studies report that gender influences student evaluations in a positive way, and others show that gender and race have a negative effect. So if empiricism does not help resolve this issue, should we be satisfied to conclude that because we cannot detect large differences in the way women or minorities are

3 For as long as the American Association of University Professors' survey has collected data on tenure status—since the late 1970s—approximately 47 percent of women on the full time faculty have had tenure, while 70 percent of men have. The proportions of faculty with tenure have dropped slightly in recent years among both men and women, but the gap has remained consistent.

treated, university faculty should continue to rely on student evaluations in making personnel decisions?[4] Even though statisticians cannot yet resolve the tricky question of how gender and race influence student evaluations, that does not mean that gender and race are not present in the classroom and are not influencing the way students see their professors and react to them.

Part I:. Subjectivity in Student Evaluations: Like Me, Like My Teaching

There is robust and extensive literature that defends the use of student evaluations and finds student evaluations to be useful as both a formative feedback instrument for the professor and a reliable summative evaluation assessment tool for administrators (Centra 1979; W. E. Cashin 1995; Cohen 1981; Costin, Greenough, and Menges 1971). For example, in an exhaustive monograph published in 1987, Professor Herbert Marsh concluded that "student ratings are (a) multi dimensional (b) reliable and stable, (c) primarily a function of the instructor that teaches the course, rather than the course that is taught, (d) relatively valid against a set of indicators of effective teaching, (e) relatively unaffected by a set of indicators hypothesized as potential biases, and (f) seen to be useful to faculty as feedback" (1987, 255). Even so, there is also recognition that "student evaluations seldom make an optimal contribution to improving either teaching or [helping make accurate] personnel decisions" (McKeachie and Kaplan 1996). This dose of skepticism is justified by the substantial research on the subjectivity of student ratings.

Beauty and the Student

A 2003 empirical study by economists Daniel Hamermesh and Amy Parker examined a large sample of student instructional ratings at the University of Texas for a random group of professors.[5] Researchers assigned six independent measures of professors' beauty and found "that measures of perceived beauty have a substantial independent positive impact on instructional ratings by undergraduate students" (p. 373). Instructors who were judged better looking received higher student ratings, which moved them from the tenth to the ninetieth percentile. This impact exists within university departments and even particular courses and is larger for male than for female instructors.

Professor Hamermesh also found that perceptions of beauty among minorities have bigger effects—a bigger penalty in evaluations for ugly minorities, a bigger positive payoff for good-looking minority group members. Although the observations for minorities contain some "noise" (for example, since the evaluators were mostly white, they had a harder time judging the beauty of other races, and minorities were also disproportionately made up of non-English speakers, who—because of their accent—are generally penalized in student evaluations), it seems fair to conclude that bad looks negatively impact minorities more than whites, and good looks help them more than whites (email from Hamermesh, August 3, 2010).

4 This is the conclusion of John Centra, the father of the Student Evaluation Instructional Ratings (SIRs) form ("The differences in ratings, though statistically significant, are not large and should not make much difference in personnel decisions" (Centra and Gaubatz 2000, 32)).

5 The study covered a total of 463 courses and 94 professors (Hamermesh and Parker 2005, t. 1).

Seduction, Enthusiasm, and Charisma

No other experiment has received as much attention as the "Dr. Fox" ones. The original 1973 study by Donald Naftulin, a psychiatrist, and his coauthors asked an actor to give a lecture titled "Mathematical Game Theory as Applied to Physician Education" to three groups: grad students, practicing psychologists, and educators and administrators (Naftulin, Ware, and Donnelly 1973). Dr. Fox was a distinguished looking actor with a pleasant voice, who was entertaining, lively, and charismatic; conveyed warmth; and made jokes. The scripted content of his lectures was nonsense, full of "double talk, neologisms, non sequiturs, and contradictory statements" (631). He was rated highly by all three groups. As the authors note, not even the group of experts in the audience was able to resist the charm of Dr. Fox and detect that his lectures were "crap" (633). The phenomenon became known as the "Dr. Fox" effect or "educational seduction."[6]

Because the original study was criticized for methodological shortcomings, the Fox study spawned a series of follow-up ones. One of the original authors, professor of medical education John Ware, and his coauthor Reed Williams, set up another Dr. Fox experiment, where 207 students rated six lectures on substantive teaching points (1975). This time the groups were divided so that there was a control group (with no seduction effect), and students were randomly assigned. Each lecture varied in substantive content, from low (four substantive points) to high (twenty-four points), and students were subsequently tested by multiple-choice exams. The same professional actor gave all six lectures using various degrees (low and high) of educational seduction. The results showed that ratings did not reflect the substantive content of the lectures and were also unrelated to how well students did on the exam. The most important factor affecting student ratings remained the seduction effect. The researchers concluded that "faculty who master the Doctor Fox Effect may receive favorable student ratings regardless of how well they know their subject and regardless of how much their students learn" (Ware and Williams 1975, 155).

Almost ten years following the original Dr. Fox study, Professor Abrami and his coauthors conducted a quantitative review of the Dr. Fox literature (Abrami, Leventhal, and Perry 1982). They found that instructor expressiveness had a substantial impact on student ratings. They also found that lecture content had a substantial impact on student achievement but a small impact on ratings. This research and follow up studies concluded that good and effective teaching was a multidimensional skill and that students were rating specific features of teaching on the basis of their global evaluation (Abrami, d'Apollonia and Cohen 1990; Abrami, d'Apollonia and Rosenfield 1997).

Recent scholarship continues to reaffirm how much the Dr. Fox effect influences student evaluations. Something that can be called enthusiasm, charisma, or likeability, originally described in that effect, strongly impacts student ratings: what students believe that they are learning—but not necessarily what they actually learn (Williams and Ceci 1997). Note, however, that other research concludes that student ratings correlate significantly with the amount students learn (Abrami,

6 It should be noted that the original Dr. Fox study has been faulted for "serious methodological shortcomings"; specifically the research was a series of one-shot case studies with neither control groups nor objective measures of student learning (Marsh 1987, 331; Abrami, Leventhal, and Perry 1982).

d'Appolinia and Rosenfield, 1997; Abrami, d'Appolinia and Cohen 1990; Feldman 1989a, 1989b).

Twenty-five years after the original Fox experiments, an internationally known professor of psychology at Cornell, Stephen Ceci, attended a teaching skills workshop taught by a professional media consultant who trained faculty to improve their presentation skills (Williams and Ceci 1997). The media consultant provided hands-on coaching and suggested ways that faculty could improve individual presentation styles, for example by varying their voice pitch (in Dr. Ceci's case, he was encouraged to lower his voice), and using more hand gestures. The goal was for the teachers to be perceived as enthusiastic. Ceci proceeded to compare the results of his student evaluations pre- and postmedia training in his developmental psychology course. He tried to teach the course in the spring semester as much as possible in the identical way that he had taught it during the fall semester, which was before the training. He taught the course on the same weekdays and at the same time, had approximately the same number of undergraduates (more than two hundred), used the same syllabus and book, adopted the same lecture design, and did not vary his content.[7] In other words, the only difference was that Dr. Ceci had acquired seduction skills with which to charm his spring class.

His student evaluation scores increased significantly in every category. Ratings went up for instructor effectiveness—knowledge of the material; tolerance of diverse views; accessibility to students; organization of lectures; enthusiasm in the classroom. For example, on a question that had nothing to do with style or enthusiasm—"How knowledgeable is the instructor?"—Ceci's pretraining mean rating was 3.6 (out of a total of 5) for the fall semester and jumped after the training to 4.05—a highly significant statistical difference. Students' reception of his more enthusiastic delivery extended to what they believed they had learned in the course. For the question, "How much did you learn in this course?," before the training, Ceci received a mean score of 2.93; after the training the mean jumped to 4.05—a change highly significant change. He concluded that the students "*thought* they learned more, but in fact, they had not; the end-of-semester point totals for the identical sets of exams . . . were virtually identical" (Williams and Ceci 1997, 22; emphasis in original).

Some researchers argue that it is personality that the students are rating. The statistical impact of an instructor's observed personality is so large that evaluations "could most accurately be called a 'likeability' scale" (Clayson and Haley 1990). In various studies, being described on personality tests as an extrovert (McCroskey 2004; Murray et al. 1990; Feldman 1986), exhibiting "charming" behavior, and having charisma (Shevlin 2000; Ederle et al. 1985,) have been shown to statistically positively influence student evaluations positively. One researcher concluded that "This robust relationship between instructor extraversion and students' perceptions of teaching effectiveness could be interpreted to support the fear of some faculty that student evaluations are just personality contests and may not be valid measures of teaching effectiveness" (McCroskey et al. 2004, 206).

7 Specifically Ceci used (1) the same syllabus, textbook, and reserve readings; (2) the same overhead transparencies at the same relative points in each semester; (3) the same teaching aids (slides, videos, demonstrations) at the same points in each semester; (4) the identical exams and quizzes; (5) nearly identical lectures; (6) the same schedule and room (days of the week, time); and, finally, (7) the same ratio of teaching assistants to students each semester. (Williams and Ceci, 1997, p. 16).

These studies have not defined what the students mean by "enthusiasm" or "charisma." Certainly humor helps (Waters 2004). So does beauty (Hamermesh and Parker 2005). Reaching out to students so that they perceive that you care is also important (McCroskey et al. 2004). And not being boring but striving to be enthusiastic and engaging seems to be a major key to good student evaluations (Shevlin 2000; Ederle et al. 1985; Waters 2004). If you can't be charming or humorous, then telegraph to the students that you will give them all good grades (Marsh and Dunkin 1992; Johnson 2003; Oppenheimer 2008). If all else fails, give your students chocolate before handing out the evaluations[8] (Youmans and Jee 2007).

On the other hand, student evaluations have been found to relate negatively to deep student learning. A recent study by Professors Scott Carrell and James West (2010) used a longitudinal data set of 10,534 students who attended the US Air Force Academy from fall 2000 to spring 2007. At the academy, students are assigned randomly to all of their classes and have to follow a rigorous track of courses in mathematics, humanities, and the sciences after taking introductory classes so, for example, students will take Calculus I, Calculus II, and then more advanced math. For this experiment, instructors in the introductory courses used common syllabi, and all students took standardized exams that were graded by several professors. Using advanced statistical methods, Carrell and West created a value-added model for instructors, which allowed them to isolate each teacher's contribution to student achievement in the actual course taught (e.g., Calculus I) and the following courses (Calculus II). They found that instructors who produced higher student achievement in the courses they taught received better evaluations from their students. They were also more likely to be untenured and less experienced. On the other hand, instructors who received low student evaluations in the courses that they currently taught also increased student achievement in follow up courses. These instructors were more experienced and mostly tenured faculty. The study concluded that instructors who were highly rated by their students did not necessarily promote the deep learning that is necessary for students to do well in more rigorous course work. Rather, students' evaluations were negatively correlated with subsequent performance—"students appear to reward higher grades in the introductory course but punish professors who increase deep learning" (Carell and West 2010, 412).

These findings may appear paradoxical to some people. Professor Stanley Fish, without knowing about this study, blogged about what it takes to promote deep learning and the tension between this choice in pedagogy and positive student evaluations: "Sometimes (not always) effective teaching involves the deliberately inducing of confusion, the withholding of clarity, the refusal to provide answers; sometime a class or an entire semester is spent being taken down various garden paths leading to dead ends that require inquiry to begin all over again . . . . Needless to say that kind of teaching is unlikely to receive high marks on a questionnaire that

8 This experiment involved 98 undergraduates from the University of Illinois at Chicago from three different classes: two statistics classes (n = 34 and 29) and one research-methods class (n = 35). The same instructor taught all three classes. Each class required students to enroll in one of two Friday discussion sections of approximately equal size. The same teaching assistants led sections for each class (with different teaching assistants serving each of the three classes). Participants who were offered chocolate gave higher ratings on average (M = 4.07, SD = .88) than those who were not offered chocolate (M = 3.85, SD = .89) (Youmans and Jee 2007).

rewards the linear delivery of information and penalizes a pedagogy that probes, discomforts and fails to provide closure" (blog by Stanley Fish, June 23, 2010).

Thin-Slice Judgments: Instructors' Nonverbal Behavior and Student Evaluations

Malcolm Gladwell's book, *Blink,* brought into the popular mainstream the cognitive research on "thin-slice" judgments (2005).[9] As Gladwell describes in his opening chapter, unconscious, lightning-fast judgments that we make at a glance are very often correct and may be more accurate than if we stopped and reflected step by step on what is going on in our decision. People may not be able to articulate what is happening as they process information, but they are thinking very rapidly at an unconscious level. These "blink" or thin-slice judgments reflect what individuals have learned in their lives about the situation they are facing; more importantly, the mind is picking up on nonverbal behaviors that may be important clues about what they are observing. In lay terms, we refer to this kind of thinking as intuition or deciding from the gut, but the mind is thinking—it is just doing so very quickly and at an unconscious level.

To illustrate, Gladwell describes a research psychologist who has spent years studying the factors that keeps couples together in happy marriages or leads them to divorce. (It turns out that mutual respect and humor are key elements in a marriage's long-term survival.) This expert's thin-slice judgments are so accurate that he can observe forty-five seconds of a video of a couple's interaction—without sound—and accurately forecast whether husband and wife will stay together or break up (2005, 21–39).

Research of thin-slice judgments also has shown that our appraisal of others—even when based on very brief observations—can be remarkably accurate. As Harvard psychology professor Nalini Ambady and her coauthors describe, "many day to day judgments of people occur unwittingly and intuitively, . . . a fleeting glimpse or a mere glance can lead to an instantaneous evaluative judgment. Once made, such judgments provide the anchor from which subsequent judgments are realized" (2000, 20).

Thin-slice judges are surprisingly accurate about reading the nonverbal behaviors and emotions of the subject (by observing the face, the voice, and the body) (Waxer 1976, 1977); assessing interpersonal behavior (Bernieri et al. 1996); determining who is in charge or is dominant within the social group (Ambady, Koo, Lee and Rosenthal, et al. 1996); and assessing kinship or empathy (Constanzo and Archer, 1989). Thin-slice judgments are more likely to be accurate in assessing nonverbal behaviors and personality traits such as interpersonal skills (Ambady et al. 2000). As an example, Nalini Ambady's research has shown that observers' quick, thin-slice judgments of a surgeon's nonverbal gestures and tone-of-voice interactions with a patient are a better predictor of whether that doctor will get sued for malpractice than his or her education (Ambady et al 2000).

Women seem to be better at making thin-slice judgments (Hall et al 2000), perhaps because they are hardwired to read emotions more accurately on people's faces. Psychologists have found that the accuracy of thin-slice judgments is also subject to

9 As defined by Nalini Ambady and her coauthors, thin-sliced judgments are "brief excerpts of expressive behavior sampled from the behavioral stream . . . less than 5 minutes long . . . from any available channel of communication, including the face, the body, speech, the voice, transcripts or combination of the above" (Ambady et al. 2000, 203).

the observer's affective state. So for example, if the observer is depressed, he or she will project negativity about the subject's emotional state onto her thin-slice judgments (Forgas 1992; Ambady and Gray 2002). Thin sliced judgments are most accurate when made by well-adjusted and stable people (Ambady and Gray 2002).

Professors Ambady and Rosenthal (1992), as well as other researchers (Clayson and Sheffet 2006; Babad, Babad, and Rosenthal 2004), have found that observers' thin-slice judgments are highly predictive of student evaluations. Ambady and Rosenthal selected thirteen graduate teaching fellows (seven men and six women) who were instructing undergraduate courses at Harvard in the humanities, social sciences, and natural sciences. No section was larger than twenty students. They videotaped classes and then produced a thirty-second composite tape of ten-second snapshots taken from the beginning, middle, and end of the class. They instructed nine female judges to score the composite videotapes of the instructors—without sound—on observable characteristics (competence, confidence, professionalism, dominance, honesty, attentiveness, enthusiasm, likeable, optimism, supportiveness, anxiety, warmth) based on thin-slice judgments of nonverbal behavior—how often the instructors smiled, grimaced, bit their lips; how they held their hands (open hand, pointing, fists); their hand gestures; head shakes; and the positioning of their heads, legs, and torso (leaning forward or backward). The researchers then calculated a composite likeability rating (that excluded anxiety). They found their nonverbal composite variable correlated significantly with the instructors' student evaluations at the end of the course by a significant factor *(r =. 70)*. A follow up study in 2004, with a larger data set, found the same relationship, but the magnitude was lower (Babad, Babad, and Rosenthal 2004).

Law professor Deborah Merritt from Ohio State University summarizes this research and its relevance to the wisdom of using student evaluations for important personnel decisions, such as retention, promotion, and merit pay:

> Students . . . rapidly form an impression of a professor's personality. An image based almost entirely on nonverbal behavior gels within the first few minutes of the semester. The students may refine their impressions as the semester progresses, but the initial image remains telling. The significant correlation between assessments completed after just five minutes of class and those offered at semester's end is . . . troubling. It confirms not *some* connection between a professor's style and student evaluations, but an *overwhelming* link between those two factors. Nonverbal behaviors appear to matter much more than anything else in student ratings. Enthusiastic gestures and vocal tones can mask gobbledygook, smiles count more than sample exam questions, and impressions formed in thirty seconds accurately foretell end-of-semester evaluations. The strong connection between mere nonverbal behaviors and student evaluations creates a very narrow definition of good teaching. By relying on the current student evaluation system, law schools implicitly endorse an inflexible, largely stylistic, and homogeneous description of good teaching (Merritt 2008, 251–52).

Summing up: Concerns and Questions

Professor Merritt's critique of student evaluations, based on her thorough research of the way the brain works and much of the research discussed here, is

devastating and raises concerns of fairness. Should a professor's ability to contribute to the academy depend on how well he or she emotively presents him or herself to the students? For many defenders of student evaluations, the answer is yes—they argue that teachers need to unpack behaviors of warmth, rapport, and connection with the students and recreate themselves as enthusiastic Dr. Fox-like professors since this is a key strategy of good teaching (Matthews 1997).

Others, like Professor Merritt, argue that likeability, charisma, and warmth are rooted in an individual's "physiology, culture, personality and habit" (Clayson and Sheffet 2006, 158) which are difficult for a faculty member to change. Certainly no instructor can alter physical things about himself or herself like beauty, tone of voice, or whether a face seems warm. Professors Clayson and Sheffet, who replicated Ambady and Rosenthal's research of thin-slice judgments and student evaluations, concur and argue that "if . . . student perceptions are even marginally related to relatively long-lasting traits [in instructors], it may be true that some teachers never will receive consistently high evaluations in certain environments, irrespective of anything they do or possibly could do" (2006, 158). Instructors can "game" the system by manipulating the students' affective state and giving them chocolates just before administering evaluations and get better student ratings or telegraphing to the students that they will get good grades in the course. Regretfully, the recent Air Force Academy study argues that evaluations are negatively correlated to students' deep learning. It is no wonder that so much research exists on student evaluations. Yet these paradoxical findings indicate that there is still more research to be done.

Part II: Manifesting Gender and Race in the Classroom

Navigating the goal of getting good student evaluations is difficult for any professor seeking a foothold and advancement in the academy. In spite of the words of caution from proponents of student evaluations, they often are given too much weight in tenure, promotion, and pay decisions. The system is "crazy for everybody" (Grillo 1997, 748), but it is particularly dangerous for minorities and women who must also contend with the unconscious biases of their students who have role expectations that are anchored in gender and race stereotypes.

Unconscious Bias and Stereotyping

As Malcolm Gladwell notes, thin-slice judgments can be very accurate, and in many instances, such as the emergency room, wizened, experienced professionals should let go of step-by-step analysis and trust them. That saves time, and doctors should trust that their unconscious mind is making quick, accurate judgments that will be hard to replicate in step-by-step analysis (Gladwell 2005). But there is a dark side to thin-slicing, as Gladwell points out (71). The learned concepts in our unconscious cognition reflect stereotypes and unconscious biases of which we are unaware, and in fact, in many cases, our conscious values may be incompatible with our unconscious attitudes. Because we are all unconsciously impacted by stereotypes that we have learned from the culture around us, most of us unconsciously discriminate in one area or another. Intuitive thinking can be very accurate in certain situations, but thin-slice judgments shaped by stereotypes that we carry in our head about blacks, women, the disabled, overweight people, mothers who work, and "out" gay men and lesbians can be very wrong and lead us to make unconsciously biased judgments.

The Implicit Association Test (IAT) developed by Yale and the University of Washington in the mid-1990s is now widely recognized as a computer test that measures unconscious or latent cognitive associations and biases (Lane et al. 2007). The IAT works by asking participants by pressing a computer key to classify words into familiar categories. As the test progresses, participants are asked to sort words into categories that reveal associational bias, for example, cockroach/bad, flower/good. Then the computer flashes images, and the ordering becomes more confusing when the sorting does not follow a pattern like cockroach/bad, flower/good, and it takes longer to do it. The test now asks the person to sort categories that require him or her to resist stereotypical associational bias—male/career, male/science, female/family, black/crime, fat/ugly. The longer that it takes the test taker to sort out categories by going against stereotype, the stronger the link to prior cognitive stereotypical associations (Greenwald, McGhee, and Schwartz 1998; Lane et al. 2007). Based on reaction times, the test reports back to the individual, for example, "Your data suggest a moderate automatic preference for thin people compared to fat people."

Over the course of the last fifteen years, the IAT has been shown to be consistent and reliable. As a group, the millions tested "demonstrated, on average, greater positivity for white over black, non-Arab Muslims over Arab Muslims, abled over disabled, young over old, and straight over gay" (Lane et al. 2007, 66). It has also measured stereotype-consistent associations between white/American, males/science, females/liberal arts, males/career, females/family, and blacks/weapons.

This is evidence of the hold that stereotypes have on unconscious cognitive processes. A person may not be aware of automatic negative reactions to a racial group and may even regard them as objectionable. Most test takers report themselves as unbiased and not holding prejudiced beliefs (Lane et al. 2007). However, most participants also possess automatic, unconscious negative feelings—in the case of whites, 97 percent have negative unconscious attitudes toward blacks; and in the case of blacks, 45 percent hold negative unconscious attitudes toward blacks (Dasgupta et al. 2000).

Stereotypes are "overgeneralizations and are either inaccurate or do not apply to the individual group member in question" (Heilman 1983, 270). Certainly many times stereotypes have some truth to them, but they also grossly overgeneralize (e.g., Latinos or persons with accents are illegals, blacks are associated with crime) and lead to distortions. As Professor Heilman notes, "Once an individual is classified as a member of a social group, perceptions of that group's average or reputed characteristics, and perceptions of behavior based on those characteristics, are readily relied on by those doing the classifying. It then becomes more difficult for the classifier to respond to the other person's own particular characteristics, making accurate, differentiated, and unique impressions less likely" (1983, 272).

Further research links unconscious negative attitudes based on stereotypes to discriminatory behavior. In a well-known study, for example, police officers whose IAT scores demonstrated strong associations between black and weapons were more likely in a video game to shoot at ambiguous black figures who popped up from behind buildings but had no weapons (Correll et al. 2007).

The prevalence of unconscious biases is connected to social structure, power, and the distribution of resources and opportunities. Specifically feminists and critical race theorists have argued that the academy is gendered and raced, meaning

that power, resources, and opportunities are not distributed equally but are based on gender and class. Men and whites are privileged and have an easier time navigating obstacles to get through the door and then rise through the ranks of the professoriate (Valian 1998; Basow 1986; McGinley 2009; Maranville 2006; Delgado and Bell, 1989). Whites and men start from a presumption of competence; minorities and women do not and have to deal with a multitude of unconscious biases that put them at a disadvantage. The playing field is not level.

Empirical research backs this claim. Researchers have shown that unconscious bias impacts who makes it through the door of academic institutions. Professors Steinpreis, Anders, and Ritzke presented real-life curriculum vitas of successful academic psychologists to a panel made up of 238 male and female academic psychologists who were to review them and make hiring recommendations. The names on the vitas were changed to male and female at random. Both men and women judges were more likely to hire male job applicants over female candidates with an identical record. The panel also rated the teaching, research, and service records of male job applicants over those of women candidates even when they were identical (Steinpreis, Anders, and Ritzke 1999).

Recent research reflects just how hard it is for blacks to get that first opportunity. In a well-known study by economists Bertrand and Mullaithan (2000), researchers sent out 5,000 resumes in response to 1,250 Boston and Chicago employers' help-wanted ads. They used made-up identical resumes; one set had "white-sounding" names (e.g., Emily) and the other "black-sounding" names (e.g., Lakisha). Every employer was mailed four resumes: (1) average qualifications with a white-sounding name, (2) average qualifications with a black-sounding name, (3) highly skilled with a white-sounding name, and (4) highly skilled with a black-sounding name. The results were that resumes with white-sounding names got 50 percent more callbacks than those with black-sounding names. The high-quality resumes with black-sounding names attracted no more interest than the average black ones, and lower-skilled candidates with white names got more callbacks than highly skilled ones with black names.

The Dynamics of Gender and Race in the Classroom

As the curriculum vita experiments show, unconscious stereotypical beliefs create expectations about someone before that person walks in the door. When women and minorities enter their classrooms, their students, too, have expectations about them. Their majority counterparts do not face this obstacle. As women and minority instructors labor to make their classrooms friendly and warm (so that they can get decent student evaluations), they must ponder how their conduct will be perceived by their students in the context of their gendered and raced role expectations. From the get-go, the task is daunting.

Gender stereotypes place women in a double bind (Eagly, Makhijani, and Klonsky 1992). When women labor in roles and jobs that are viewed as male, they must fight the stereotypical presumptions that they are not competent, authoritative, or charismatic leaders (Valian 1998; Eagly, Makhijani, and Klonsky 1992). However, when women try to compensate for those perceived shortfalls, they can come across as more incompetent (because she lectures too much), insecure (because she keeps referring to her credentials), or self-promoting (because she tries to put herself in a leadership position). Further, if she does not fulfill the stereotypical expectations of

being nurturing and caring and polite, she will experience backlash (Valian 1998). Women professors who behave counter to stereotypes and exhibit "non-lady-like" behavior receive lower evaluations than men (Basow 1998), and many see their careers placed in jeopardy. Moreover, if she is part of only a handful of women within her institution, below critical mass, she will stick out as a token, which will amplify stereotypical expectations (Kanter 1977).

Minorities also experience double bind stereotypic expectations. The presumption when a minority professor walks in the door is that he or she is not well credentialed (Harlow 2003). Showing irritation or anger backfires. Creating a comfortable learning atmosphere requires that the minority teacher put white students at ease in relation to issues of race. Yet African Americans must labor under the additional burden that white students have a harder time sorting out the emotions on their faces because generally whites cannot read black faces very well.

Numerous studies have attempted to determine how gender and race impact student evaluations. Surprisingly the results have been equivocal. Studies with large data sets across campus for a single large institution, like Dr. Hamermesh's University of Texas study (Hamermesh and Parker 2005) and the Air Force Academy's study (Carrell and Scott 2010), report that women are rated slightly lower (but still statistically significant) than their male counterparts. The Texas data set also showed that minorities fare slightly worse than white instructors (Hamermesh and Parker 2005).

Another set of studies reached the opposite conclusion. Professors John Centra and Noreen Gaubatz (2000) assembled a large data set made up of 741 classes in the humanities and sciences in twenty-one colleges and found no differences in the ratings of male and female instructors. Only in one area—course organization and planning—was there a slightly significant ratings difference in favor of male instructors. During the 1990s, Professor John Feldman published a two-part meta-analysis, and concluded that the differences that existed were slight, and not sufficiently significant to show gender bias (Feldman 1992, 1993). Peter Seldin's (1993) brief review as well concluded that gender has little or no effect on student evaluations. Yet other studies have had such mixed results that the authors hesitated to conclude whether their data showed gender bias (Hancock, Shannon, and Terntham 1993).

The sociological perspective offered by feminist researchers is helpful in resolving this apparent quandary. Gender and race affect student evaluations in more subtle ways than statistics reveal. Professor Anne Statham is unequivocal that students bring gender expectations into the classroom (Statham et al. 1991, 117), even though the statistics are deceptive in other studies (Basow 1998; Laube, Massoni, and Sprague 2007). Although in overall ratings, women appear to be, or are close to being, on a par with male professors, a more careful examination shows that women have to labor harder to satisfy student expectations (Basow 1998; Laube, Massoni, and Sprague 2007). Things can go wrong very quickly for women and minority instructors.

That is because two sets of expectations are in conflict: one is based on the social/role expectations that come from being a woman (warm, welcoming, nurturing), and the other relates to being a competent professor (knowledgeable, enthusiastic, and interesting) (Basow 1998; Valian 1998). In addition, expectations themselves are variable and are shaped by the discipline the woman is teaching—humanities and nursing, for example, are considered more female as compared to science, engineering,

law, or medicine (Basow 1995, 1998). Finally, some institutions may have a history of being more friendly and welcoming of women and minorities than others (McGinley 2009; Basow 1986). For example, Professor McGinley (2009) discusses possible bullying that may occur in some institutions with a history of male dominance.

Women in academia are comparable to women managers in leadership positions. In the course of a semester, they must lead their students through a syllabus, somehow convince them that the materials are fun and accessible, and challenge them to challenge themselves through difficult passages. Women's leadership, both inside and outside the academy, is expected to embody both stereotypically feminine qualities of nurturing and relationship building as well as the stereotypically masculine qualities associated with competence and leadership (Valian 1998). In workplace settings, women in leadership positions are decidedly at a disadvantage. In a startling experiment with trained actors who pretended to be managers, women managers who took the lead in workplace discussions were unfavorably received by both women and men listeners as measured by their nonverbal facial expressions. Male managers, on the other hand, were always well received (Valian 1998, 130). In a meta-analysis, Professors Eagly, Makhijani, and Klonsky (1992) found that women in leadership positions were evaluated least favorably when they deviated from prescribed gender roles or acted in a masculine (or strict) manner.

Women have to navigate within narrow boundaries set by cultural stenotopic expectations. In workplace leadership settings, they must be sufficiently assertive to be listened to and taken seriously, and yet not be viewed too assertive or overly masculine. Professors Eagly, Makhijani, and Klonsky (1992) found that having a style that is too assertive or perceived as autocratic is especially costly for a woman. In such situations, women receive especially negative evaluations. While a man may get away with being snippy, not consulting those who work for him, or not always saying please and thank you, when a woman commits such errors, the backlash is severe and may result in rejection by her peers and being fired by her superiors (Eagly, Makhijani, and Klonsky 1992; Valian 1998; APA brief, Price Waterhouse 1988). In the notorious case of *Hopkins v. Price Waterhouse* (1988), Anne Hopkins, a woman who was very competent and worked as hard as any male manager, was not promoted at the time of partnership, and was advised by her superiors to "act more feminine."

Interpreting Student Evaluations and Gender Dynamics in the Classroom

Many women and minorities report that deciphering their student evaluations is confusing (Grillo 1997; P. J. Smith 2000). What does the research show are the keys for women instructors to do well in their student evaluations?

Presumption of Incompetence

Research shows that both minorities and women are presumed to be incompetent as soon as they walk in the door. Professors Miller and Chamberlain (2000) conducted a survey of three hundred undergraduates taking sociology classes in a department that had a critical mass of women faculty (25 percent). They found that students consistently underestimated the educational credentials and academic rank of women and minority professors. In a study of a public midwestern university where there were few African American professors, Professor Harlow (2003) reported that her interviewees—minority professors—said they were challenged

frequently about their qualifications to teach in the classroom. Most black professors interviewed felt that their classes always contained at least some students who questioned their ability to be professors.

Different Strokes for Different Genders

An early field study by Professors Basow and Silberg matching male and female professors of similar rank from comparable disciplines in a liberal arts college showed that the gender of the student was a key variable in student evaluations. There was a consistent pattern that male students rated their female professors lower on all measurements on the student evaluations—scholarship, clarity, student interaction, and enthusiasm (Basow and Silberg 1987). Another recent study of evaluations gathered in 741 different courses taught at twenty-one different institutions showed that women faculty received significantly lower ratings from male students than from females (Merritt 2008). Researchers have also found that male students in disciplines considered masculine, such as economics, business, and engineering, are more likely to rate their women instructors negatively (Basow 1995). Professor Basow speculates that this may be because male students in these disciplines hold more traditional views.

In a later study where Basow (1995) reviewed four years of student evaluations at a liberal arts college, she found a strong pattern of student interaction with professors' gender. Women students consistently rated their women professors highest, and male students were consistently hard on them. So particularly for women professors, the adage "you can't please everyone all the time" is particularly fitting. The same lecture from the front of the class may be ringing all kinds of bells in a woman student's brain while a male brain may be hearing just "blah blah blah."

In the Air Force Academy database put together by Professors Scott Carrell and James West that mapped the progress of students over six years, researchers found that women instructors had a highly positive value-added effect on female cadets. Young women who were taught introductory courses in science or math by a female instructor performed substantially better in the following advanced courses than their counterparts who had male instructors. The researchers found that the female students who had very high scores on their SAT in science and math benefited (by performing at the highest level in the follow up courses) when they had woman instructors in their introductory courses (Carrell et al. 2009).

The Carrell and West gender study may indicate that a reason that female students rate their women instructors higher is that they get something from interacting with a female teacher—inspiration, confidence building, female role modeling, or a teaching style particularly tuned to female sensibilities—that they don't receive from male instruction. This is a valuable kind of learning and goes beyond mere in-group preferences.

Likeability and Warmth

For women instructors to be well recommended in student evaluations, they must live up to the female-stereotyped expectations that they should be warm, friendly, and supportive inside and outside the classroom and have good interpersonal skills (Kierstead et al. 1988). One study found that women who smiled were rated much more favorably than unsmiling ones. Men also gain standing by smiling frequently—although not as much as women—and are not as heavily penalized when they do not smile (Kierstead et al. 1988).

As Professors Sprague and Massoni point out, women function under a different scaling system than men. Stereotypes can influence the evaluators' understanding of a trait. Stereotypes shift not only their balance in expecting things from teachers but also their perceptions about what it entails to achieve those qualities. Students expect women to engage in a different set of behaviors to satisfy a particular trait (Sprague and Massoni 2005). To be considered caring, women had to spend more time meeting students outside of class and being accessible during office hours (Bennett 1982; Statham et al. 1991). Students were more harshly critical if their women instructors were not available (Bernstein 1995). In another study that looked at the way students described their best and worst male and female teachers, the best women teachers were called caring, helpful, and kind (that is, nurturing); in contrast, the best male professors were funny and friendly (that is, entertaining) (Sprague and Massoni 2005). In a study that actually observed women's interactions in the classroom, likeability increased more when they interacted with students, generated laughter, acknowledged contributions, and allowed students to interrupt comfortably for clarification and input (Statham et al. 1991). Feedback and correction from women were well received only when they were gentle and affirming (Statham et al. 1991).

Women who conform to stereotypical expectations of approachability, caring, and warmth are rewarded with good evaluations. Projecting warmth and putting in the time to be considered caring and kind and relieving tension by frequently smiling or keeping things light are traditional female behaviors; at the same time, these are class-management techniques that produce good teaching in general. However, women "outliers," whose personality is male oriented and who are not smiley or giggly, are more likely to be disliked by students because they do not exhibit these stereotypical behaviors (Statham et al. 1991). Professors Sprague and Massoni's (2005) study of the best/worst teachers found that students were particularly vitriolic against women who disappointed them by not seeming nurturing. The worst women teachers were chastised as cold, mean, and unfair; students sometimes used terms such as "bitch" and "witch." By contrast, these kinds of gender-specific phrases were not used to describe the worst male teachers. Indicating just how emotional students can get with female teachers who fail to be nurturing, disappointment can be so extreme that it results in death threats. Professor Pam Smith's ethnographic study of what can go wrong when an African American female is viewed by her students as overly demanding and harsh showed a divided student body where the teacher, not the material, became the focus of the class and student incivilities were extreme, ranging from personal comments on her dress and hairstyle to death threats and hate mail (Smith 1999).

Managing Authority

Being authoritative represents a particular challenge for women and minority professors. Recall that because of stereotypes, students assume women and minorities are under qualified to teach (Miller and Chamberlain 2000). Students have less fear of and respect for their female and minority instructors and are more likely to challenge their authority. Professor Statham and her coauthors (1991) observed the interactions of women and male instructors with students over the course of a year at a liberal arts college and found that women were challenged in class at least 10 percent more often than men. Challenges were more frequent when women professors were at the assistant and associate levels.

Both male and female instructors find maintaining authority in the classroom and at the same time keeping the atmosphere warm (necessary to get positive student evaluations) to be challenging. Instructor corrective strategies that students will accept from women are limited (Statham et al. 1991). Women must avoid being considered "mean" and having "no sense of humor"—descriptive terms that students reserve for their worst women professors (Basow 1998).

Professor Ann Statham's field study found that women instructors handled authority differently from men, namely, with a light touch and by seldom directly confronting the student. When students did not directly challenge the professor's authority (such as by talking in class or arriving late), women professors, particularly at the lower ranks, dealt with the problem by approaching them indirectly after class or ignoring the problem. Male assistant professors felt that they could confront the offending behavior directly, such as, for example, taking a newspaper away from a student who was reading during their lectures.

Professor Statham and her coresearchers found that women professors handled verbal challenges to their authority in class with a "considerable amount of patience, even when they thought that the students were wrong" (1991, 77). In one case, a woman associate professor patiently endured the objections to her presentation of the class materials by one student for three weeks. On the other hand, male assistant and associate professors felt that they could directly confront challenges by explaining to the student why he or she was wrong. Only when a woman professor had reached the rank of full professor did she feel that she could publicly stop a student's challenging behavior with a reprimand (Statham et al. 1991).

Another study examined student evaluations to determine the way students reacted to negative grades from women and minority instructors. In an empirical study of more than two hundred students and seven hundred course evaluations, students judged the quality of their instructors after they received their grades. Female instructors were evaluated much more harshly than males, and minority teachers were judged more severely than their white counterparts (Sinclair and Kunda 2000). The researches call this dynamic *motivated stereotyping,* which they say occurs because stereotypes allow students to be more dismissive of a disappointing grade from a female or minority instructor. Motivated stereotyping puts the blame for a student's disappointing performance directly on the female or minority instructor, who was judged incompetent to begin with (Sinclair and Kunda 2000).

In the Statham field study, women instructors often adopted positive reinforcement strategies in the classroom by, for example, pointing out what the students were doing well and correcting them by suggesting ways they could do better. Professor Statham and her coauthors call this feedback *modified control* that is partially positive. This soft student-professor interaction correlates with positive student evaluations for women professors in both competence and likeability. (No comparable correlation was measured for male professors.) By contrast, women professors' unspoken positive reinforcement in the classroom, for example, just nodding or smiling and not expressly saying, "I like the way that you are approaching that question," correlated negatively with the way the students rated their women professors' competence and likeability (Statham et al. 1991).

Professor Statham and her coresearchers concluded that women deal with the stereotype double bind by redefining the way they exercise authority in the classroom. The corrective strategies that many women professors use with mostly good

effect does not stray too far from students' stereotypic expectations. Statham and her coauthors argue that women professors are remolding what students see as role-appropriate behaviors for women professors. However, exercising authority so subtly, where students are rarely directly reprimanded or embarrassed, can hardly be said to reshape students' gender-role expectations. However, Statham and others retort that women professors are changing role expectations because the interactive teaching style that many women adopt in the classroom is less hierarchical and more informal, and students are more directly involved in the process of learning. Statham argues that such a feminine approach to teaching "abolishes" women's power and authority. However, these transformation claims may be overstated because, again, a nonauthoritarian teaching approach is the kind of style that students come into the classroom expecting from a woman professor (Statham et al. 1991).

Feminist analysts like Statham, who write glowingly about a generally soft feminine approach in managing authority and teaching, are also rightly concerned about women instructors whose personality and style do not match these stereotypical feminine strategies (Basow 1998; Sprague et al 2007). It is clear that direct exercises of authority by a woman professor engender student backlash (see e. g., Maranville 2007; Smith 1999).

Competence and knowledge

To receive favorable evaluations from students, women professors are expected to act more experienced and professional, have a highly structured instructional approach, demonstrate more effort preparing for class, spend more time with students, provide a reduced workload, and give higher grades than men professors (Bennett 1982). However, when women overwork being competent or capable, they can receive student backlash. Professor Statham and her coauthors found that women instructors who spent more time presenting material in the classroom and going over substantive points got higher competency ratings but lower likeability ones. When women instructors checked on students' understanding and solicited input, they got higher likeability ratings, but their competency marks fell. As Professor Statham and her coauthors point out, this represents a particular double bind for women since likeability and warmth are key elements for women professors to get good student evaluations. Students may look down on women who labor to clarify difficult points (why is she trying so hard to teach me?) (Statham et al. 1991).

Perception as a "Partisan Hack" by Being a Woman Teaching Women's Studies

Recent research documents that a negative relationship exists between students' perceptions that their professors are ideologically driven and their evaluations. Professors Woessner and Woessner surveyed at random thirty political science instructors teaching undergraduate classes to 1,385 students. They found that students are more critical of a course when it is taught by an instructor that they view as highly partisan. The more that the professor's political views differ from the student's, the more likely students are to think that their professor is not competent and does not care about them. Students report not being comfortable in classrooms where the general ideological viewpoint differs from their own. The greater the differences between a professor's and student's ideological positions, the lower the student evaluations are (Woessner and Woessner 2006).

The case studies reported by journalist Oppenheimer (2008) of professors who lost tenure-track jobs because of student evaluations involved feminists who were presenting material from a feminist or an outsider viewpoint. In another study, minority faculty reported that they were more likely to be challenged by their students when they discussed issues of race in the classroom (Harlow 2003).

Several studies indicate that stereotypes predispose students to view their minority and women professors as ideological partisans when they are teaching controversial subject matter. Professors Moore and Trahan tested students' attitudes by asking students to rate a syllabus for a proposed sociology of gender course to be taught by a hypothetical woman professor. The students were asked to project what they anticipated the course experience would be like. The majority predicted that the professor would be biased and more than likely would have a political agenda. When the hypothetical teacher was a male professor, students did not believe that he would have an ideological agenda (Moore and Trahan 1997). Another study found similar results with a Racism and Sexism in American Society class when the instructor was African American (as opposed to white) (Ludwig and Meacham 1997). And a third study found this attitudinal bias when a hypothetical Latino professor was proposed to teach a course called Race, Gender, and Inequality (Smith and Anderson).

Some students react negatively to professors who challenge their ideological beliefs. Psychological research has shown that reviewers asked to read articles on the death penalty rated the authors most harshly when they differed from the reader's ideological belief (Lord, Ross and Leeper 1979). When the article contained divergent ideological views, the reader easily identified flaws in it and was more likely to question the credentials and authority of the author. These readers were much more likely to disparage the sources of information. When students sit in a classroom and have to hear a viewpoint from a feminist teacher or a critical race theorist that clashes with their worldview, he or she cannot escape so the most convenient way to deal with this unpleasant classroom experience is to disparage the professor, his or her abilities, and the teaching approach. The student evaluation provides a handy complaint form. As observers and researchers have noted, the vitriol that students express in forms that take aim at feminist, multicultural, or any outsider subject matter sounds extreme and highly emotional. Yet these outlier evaluations are averaged in with the other, more temperate student evaluations.

Part III: Minority Professors and Student Evaluations

This chapter has already discussed the kind of dynamics that minority and women professors share. In addition, critical race theorists Richard Delgado and Derrick Bell reported the following as part of their survey of minority law professors:

> Minority law professors' teaching evaluations, as reported, are generally at or near the institutional median. Substantial numbers reported that their evaluations vary greatly from subject to subject, are sometimes both positive and negative for a single course, or are best in technical subjects that do not call for much normative analysis. Some said that while they are treated politely by majority race students in class and around the law building, they are regularly "trashed" on evaluations. Some report increasing numbers of "bullets": students who give the professor the lowest rating

> in all categories, thereby lowering his or her average as much as possible. (p. 355)

A study by Professor Harlow interviewing African American faculty found that they were highly aware of what they believed were students' (unconscious) biased perceptions:

> 83 percent believed that students immediately reacted to their race;
>
> 76 percent believed that students questioned their intellectual authority;
>
> 55 percent believed that they had to prove their competence and intelligence; and
>
> 34 percent experienced inappropriate intellectual challenges (2003, 352)

As a consequence, minority law professors must face racial performance burdens in the classroom that white professors do not encounter :

White Faculty	African American Faculty
Do not have to worry about race (instead hold white privilege)	50 percent believe race will have a negative effect on students' evaluations.
Do not have to worry about students questioning competence	Because minority professors fear that their competence will be questioned, 69 percent of black women and 44 percent of black men choose an authoritative demeanor, which in turn, may turn off students who reward likeable professors.
Do not have to deal with students' stereotypes that make negative assumptions about the professors' competence, knowledge, and qualifications	To be effective, black faculty must manage their own perceptions of students' behavior that is influenced by negative stereotypes, and not overreact by becoming unfriendly, sullen or angry.
The more white male professors interact with students, the more likely it is that they will be rewarded with positive evaluations.	White students are not able to accurately perceive the emotions behind the facial expressions of minorities, so misunderstandings about a minority professor's intentions, their emotional warmth, are very likely occur. As well, white students perceive professors with African American features as less attractive, which in turn negatively impacts student evaluations.
More range of choice as to the selection and emphasis of the subject matter	When minority professors talk about race in their classroom, students are more likely to say they are biased or "spend too much time" doing it. A minority professor can safely address controversial race issues, only if she positions herself as a "nonpartisan."

In sum, minority professors must negotiate many more burdens than non-minority professors from the first moment that they walk into the classroom. These

additional burdens and potential risks are difficult to navigate even for the most experienced professor, but the risks are higher and the penalties even heavier for newly minted assistant professors who must also master new material, learn to teach effectively, and get a productive research agenda on track. New minority professors start their careers with a significant handicap not of their own making.

Professor Harlow reports that one in two minority professors anticipate that they will receive less favorable evaluations solely because of their race (Harlow, 2003). The classroom is filled with positive and negative emotions. Students enter the classroom with unconscious stereotypes about the professor's race, ethnicity or accent, which in turn informs how the student perceives, listens, and reacts to the professor. The student may well perceive herself as fair-minded and racially enlightened, yet these unconscious stereotypes influence cognition and emotions at an unconscious level; it is part of the brain's system of "blink" or automatic thinking.

On the other hand, a minority professor who has had to deal with a lifetime of racial slights might well react negatively to what he or she perceives as conscious or unconscious biased treatment from his or her students. The situation can quickly deteriorate. Students misread their minority professor, and the minority professor reacts with irritation. Now the students have reason to perceive the minority professors as not being as "warm," or, more damaging, as the "angry" minority who is overly sensitive about her race and eager to push a partisan ideological racial agenda. The professor reacts to what she perceives to be unjustified hostility by deploying more authority in the classroom and becoming even more formal and emotionally unapproachable. More students, in turn, become disconcerted, alienated or angry, and these reactions will be recorded in harsh and emotional evaluations.

Hence, minority professors must be able to closely monitor and manage their emotions, conscious and unconscious, from the first day they walk into the classroom. Minority professors cannot get caught up in anticipating that their students will be hostile, because the classroom atmospherics will deteriorate and become tense. Neither can a minority professor make an issue of students' lack of racial knowledge or sophistication, because it will sound like preaching, "talking down to," or partisan politics, all of which create a high likelihood of student backlash in her evaluations (Smith 1999).

In one of the early works of critical race theory, *The Alchemy of Race and Rights*, Professor Patricia Williams described ordinary blacks as consciously over-dressing to do routine tasks, such as going shopping, because they did not want to trigger white shopkeepers' and white security guards' negative unconscious stereotypes; for example, not being buzzed into a Benetton store or being shadowed by security when shopping for shoes (Williams 1992). In a similar manner, minority professors can choose to perform their race in the classroom so that they are not tapping into the most negative stereotypes about minorities (angry political racial partisan) but rather more benign or neutral stereotypes (such as middle class black professional). Not all discrimination is the same. According to survey data, the most acute kind of discrimination is aimed towards blacks who are viewed as militants, while middle class blacks are viewed more neutrally (more respect about their competence, and more warmth towards them as a social group) (Fiske 2010). Accordingly, the strategic minority professor will "perform" her race in the classroom so that students think about her more as a middle class professional, and stay away from the more negative stereotypes associated with minorities. So at one level, this means doing the

basics well: being prepared, being knowledgeable, listening to students, and thinking about how to communicate well difficult concepts to students; in sum, teaching well (Bain 2004). However, as well, performing as a middle class professional minority may well mean staying away from racial hot button topics, until and unless the professor can figure out how to handle such volatile topics without seeming to be a racial partisan. Sadly, this final observation undermines the most compelling justification for diversifying faculties—the assumption that minority professors will be able to teach students more empathy and sensitivity about the racial issues that divide American society.

Conclusion

Individual minority and women professors can do a great deal to negotiate the stereotypes in the classroom that will influence how students see them and judge them. Many individual minority professors, including women, are able to manage the complex process of overcoming stereotypes, adopting effective teaching techniques, and making material accessible. Thus, they become highly successful teachers.

However, academia needs to make systemic changes to account for the factors that systemically negatively impact both women and minority professors. The question of what student evaluations measure should be framed productively. This is a systemic problem, not an individual (female, minority female) one. For example, the American Bar Association took responsibility (at least in its 2006 report) for the lack of black women's success in private law firms. The Association of American Law Schools should take such leadership. Only if academia adopts responsibility as an institution will the playing field become level for minority and women professors.

First, at a minimum, macroanalyses of bias in student evaluations are clearly needed by gender, race, and sexual orientation. In spite of decades of research, controversy continues to exist as to whether there are systemic biases that impact women and minority professors whose careers can too easily be negatively impacted by student evaluations.

Second, institutions should think about teaching and the evaluative process more creatively. Suggestions from Professor Merritt include: Use focus groups mediated by outsiders. Do evaluations less often but more deeply. Get students to think, not react intuitively. Each teacher should get feedback at least once during the semester and react to it. Think of teaching as an ongoing process, not an end product (Merritt 2008).

If decision makers do not take the time or care to fully understand the candidate's teaching file, including evaluations, and permit important personnel decisions to proceed on the basis of potentially misleading or biased data, then they ethically fail the professoriate, students, and the institution.