Sex Roles, Vol. 53, Nos. 11/12, December 2005 (© 2005)
DOI: 10.1007/s11199-005-8292-4

Student Evaluations and Gendered Expectations: What We Can't Count Can Hurt Us[1]

Joey Sprague[2,3] and Kelley Massoni[2]

Does teacher's gender impact students' evaluations? We critically evaluated the research literature and concluded that the form gender bias takes may not be easily detectible by quantitative scales. To explore this possibility, we did a qualitative analysis of the words that 288 college students at two campuses used to describe their best- and worst-ever teachers. Although we found considerable overlap in the ways that students talked about their male and female teachers, we also saw indications that students hold teachers accountable to certain gendered expectations. These expectations place burdens on all teachers, but the burdens on women are more labor-intensive. We also saw signs of much greater hostility toward women than toward men who do not meet students' gendered expectations.

KEY WORDS: teaching evaluations; student evaluations; gender discrimination.

Administrators place increasing emphasis on student evaluation of professors in making personnel decisions. For example, in a nationwide survey of administrators of accredited, 4-year, undergraduate, liberal arts colleges in 1998, 88.1% reported that they "always used" systematic student ratings of teaching in evaluating faculty; in earlier surveys, fewer administrators gave this response: 80.3% in 1988 and 54.8% in 1978 (Seldin, 1999).

Reviews of the research literature on the student evaluation of college teaching could easily give the impression that gender is not a major factor in the process (e.g., Aleamoni, 1999; Feldman, 1992, 1993; Fernandez & Mateo, 1997; Freeman, 1994; Wheeless & Potorti, 1989). In fact, a respected guide for administrators on how to evaluate teaching specifically recommends that they not take gender into account in interpreting student ratings; the authors concluded that research has revealed that, if anything, students rate women higher than men (Cashin, 1999). The only citation to support this claim is two papers by Feldman (1992, 1993) that report separate meta-analyses of experimental studies and of studies based on students' evaluations of actual professors. Feldman concluded that direct effects of gender on evaluation were minor, trivial in size, and, in the case of evaluations of actual professors, favored women over men.

We argue that a more careful reading of the research literature reveals that the evidence is mixed. Meta-analytic strategies such as Feldman's (1992, 1993) may obscure more than they reveal. Miller and Chamberlin (2000) argued that Feldman's methodology may have depressed findings of gender effects because he combined information from qualitatively different sources. For example, he averaged correlations between gender and student rating of the professor that are quite disparate, and he aggregated ratings from studies that vary by discipline, type of institution, and unit of analysis.

In a review of the research, Aleamoni (1999) reported that "a majority of studies" (p. 156) find no relationship between either teacher gender or student gender and student ratings. But the claim of no gender difference was weakly supported with a reference to seven studies cited in an article published in 1971.

[1] An earlier version of this paper was presented at the annual meetings of the American Sociological Association, Chicago, IL, August, 2002.

[2] Department of Sociology, University of Kansas, Lawrence, Kansas.

[3] To whom correspondence should be addressed at Department of Sociology, University of Kansas, 716 Fraser Hall, Lawrence, KS 66045; e-mail: jsprague@ku.edu.

Further, it is important to note that "no difference" only means "no discrimination" if one assumes that women could not be better than men at teaching. If it is true that good teaching focuses on the learner and learning, and thus requires empathy, flexibility, and sensitivity to the needs of others, then these are skills to which women are more likely than men to be socialized.

Aleamoni also reported that although there is very little research on whether overall items are consistent with specific rating scales, the five studies on this question showed that global assessments are not strongly related to specific ones, and they are more highly correlated with gender, status, and other contextual factors than are specific measures. Given that personnel decisions tend to rely heavily on these overall assessment scores, Aleamoni was implicitly calling attention to gender biases in common practices.

Some researchers argue that it is not the sex of the professor but rather the degree to which the professor's personality matches (or departs from) traditional notions of gender that makes a difference in the kinds of ratings students give. Freeman (1994) found that students assumed a hypothetical professor was more effective when that professor was described in androgynous terms than in either "masculine" or "feminine" terms.[4] Wheeless and Potorti (1989) reported that students' descriptions of their professors as androgynous were correlated with more positive feelings about a course, the belief that they were learning more in it, and intentions to pursue similar courses in the future. There are indications, however, that the impact of personality on students' evaluations depends on whether the professor is male or female. Bray and Howard (1980) found that men college instructors who were more androgynous also had higher ratings on indicators of student satisfaction and student progress than did their same-sex peers. On the other hand, androgynous women instructors outranked other women instructors only on student satisfaction; they were tied with "masculine" women instructors on reports of student progress (see also Martin, 1984).

We can see several sources of confusion in this research literature. For one, it seems reasonable to expect that contextual factors influence students' responses to their professors, yet few researchers take contextual factors into account. As Tatro (1995) found, one contextual factor that is highly salient to most students is grades, and there is reason to suspect that expected grade is related to students' evaluations in gendered ways. Sinclair and Kunda (2000) reported that students who received better grades also gave their college instructors higher evaluations, whereas low grades disproportionately reduced the ratings of women instructors in comparison to men instructors.

Another source of variation in findings seems to be whether evaluations of professors are assessed in quantitative or qualitative terms. The survey findings of Bachen, McLoughlin, and Garcia (1999) are particularly striking in this regard. When asked to rate their experience with a male or female professor, male students' ratings on quantitative scales did not vary by gender of professor. However, when later given the opportunity to answer an open-ended question about gender differences in teaching, one-half of the male students volunteered that women were not as professional or as challenging as men. This kind of qualitative investigation can identify gendered relational dynamics that quantitative assessments are unable to detect.

Frankly, as sociologists who specialize in gender, we are puzzled by conclusions that gender has no impact on teaching evaluations. Three decades of scholarship has shown that gender is a significant factor in shaping interactions, practices, and outcomes in every major realm of human social life: family, work, science, medicine, religion, sports, and popular culture—to mention just a few (see, for example, the reviews of research in Chafetz, 1999; Ferree, Lorber, & Hess, 1999). Why would the classroom be any different? As Rakow (1991, p. 10) noted, "we cannot set aside the social relationships of the larger world—a world in which classifications of gender, race, and class are among the most paramount—as we take up the more temporary relationship of professor and student." We believe that taking into account what researchers have learned about how gender operates in other contexts provides a methodological explanation for why many studies on gender and the evaluation of teaching indicate that there is no gender difference.

The Broader Literature on Gender and Evaluation

West and Zimmerman (1987) first coined the term "doing gender" to emphasize that gender is less

[4] Although we are aware of the debates about the concept of "androgyny," much of the research that we review use this terminology and so we use it to describe their studies accurately.

an attribute than a performance, something we do according to culturally defined scripts for masculinity and femininity. Many have demonstrated since then that people hold one another accountable for cultural assumptions about the gender-appropriateness of our performance, and people do gender because they know they will be held accountable to those standards (e.g., Biernat & Kobrynowicz, 1997; Connell, 1995; Kobrynowicz & Biernat, 1997; West & Fenstermaker, 1993).

One particularly relevant area where gendered evaluation processes have been demonstrated is in the evaluation of work and workers. Comparable worth studies have revealed that perceptions of the demands of a job, the assessment of the performance, and the promise of a worker are all highly gendered (e.g., England, Herbert, Kilbourne, Reid, & Megdal, 1994; Ferree & McQuillan, 1998; Martin, 1996; Steinberg & Haignere, 1987). Studies of decision-making groups show that people discount the contributions of, and are less willing to be influenced by, women, particularly women who do not conform to traditional gender expectations (Carli, 1990; Ridgeway, 1987). The broad and consistent message across all of the studies in this area is that people discount women's skills and effort, are not comfortable with women in positions of power, and respond poorly to women who overstep their culturally assigned bounds.

If we extrapolate from that literature to the college classroom, we see that men professors need to live up to race- and class-specific expectations of "man" and "professor," whereas women teachers need to live up to similarly specific expectations of "woman" and "professor." The overlap between expectations for the category of "man" and "professor" is considerably greater than for "woman" and "professor" (cf., Martin, 1984; Rakow, 1991). When Burns-Glover and Veith (1995) asked college students to rate the desirable traits for a hypothetical professor, referred to as either "Sam," "Sarah," or "Dr. Lawson," students selected similar traits for "Sam," ostensibly a male professor, and the gender-neutral "Dr.," but diverged from this pattern when they selected desirable traits for "Sarah," the woman professor. For these students, "man" and "Dr." were synonymous. Similarly, Bachen et al. (1999) found that college students most often contrasted women professors to men professors—using men professors as their referent or standard from which women deviate (cf., Martin, 1984).

Psychological research on social cognition has explored gender-specific evaluation processes under the rubric of "shifting standards." Whenever people are called on to make a judgment, they do so in relation to some point of reference. When an evaluation concerns a behavior or attribute that resonates with race or gender stereotypes, these stereotypes influence the standard or context used to judge a particular member of the group (Biernat, 1995; Biernat & Kobrynowicz, 1997; Biernat & Manis, 1994; Biernat, Manis, & Nelson, 1991; Kobrynowicz & Biernat, 1997, 1998). Bennett (1982) pointed to a gendered shift in the standard that is applied to evaluate college instructors. She surveyed undergraduate students on how much personal attention they both expected and received from their instructors, as well as how they would rate their instructors on availability outside of class. Students expected and reported getting more personal time from women than from men, and yet were more likely to rate women instructors as not available enough. These students' reference point for "enough" availability clearly shifted to a higher order for women teachers (see also Burns-Glover & Veith, 1995).

Further, both the "doing gender" and the "shifting standards" models suggest that changes in standards of evaluation may be qualitative as well as quantitative. That is, the stereotypes that people hold for a particular group can influence their understanding of the meaning of a trait in members of that group. In the case of teacher evaluations, students' gender stereotypes are apt to "shift" not only their baseline expectations for their teachers' traits, but also their perceptions of what those traits entail. Thus, students may expect a woman professor to engage in a different set of behaviors to satisfy a particular standard than they would expect of a man professor. For example, students may expect a woman professor to spend office hours walking them through a task when they might expect a man professor to only give brief directions in class.

In summary, a wide range of research on how gender shapes other evaluation processes leads us to suspect that the evaluation of teaching is also gendered. We suspect that the frequent failure to find gender differences in student evaluation of professors is an artifact of the method used to look for them. In the typical approach, students are asked to rate a real or hypothetical professor or instructor on some range of attributes using a Likert-type scale that usually ranges from 1 to 5 or 1 to 6. Then the researchers compare the mean ratings for men

teachers to those for women teachers. It is not surprising that most research has taken this approach. It parallels the way students are usually asked to evaluate college professors and instructors in the real world. They are given a list of traits and behaviors and asked to rate their teacher on each using a numerical scale. The mean scores on each item are then taken as an indication of teaching performance.

However, this approach to evaluating teaching or to studying whether and how gender enters into the evaluation process is based on two assumptions that the research literature suggests are untenable. First, it assumes a universal metric: that a "3" is a "3" and a "5" is a "5," no matter who the teacher is. Second, it assumes that a specific rating corresponds to equivalent behaviors or abilities across professors and instructors. But if, as the research suggests, students use different baselines for men and women, or, in some cases, they draw on totally different behaviors to evaluate a trait, quantitative studies are not able to detect these gender differences. This would also explain why the research that does indicate gender differences tends to be qualitative (e.g., Basow, 2000; Bachen et al., 1999; Siskind & Kearns, 1997; Sprague & Kobrynowicz, 1999).

Before we can reliably address the question of whether student rating scales produce gendered outcomes for teachers, we must first resolve the question of whether the process of student ratings might itself be reproducing gender. We hope to take that first step by asking: are students seeing teachers in gendered ways, holding them accountable to gendered criteria? Our hypothesis is that the frameworks students use to perceive and evaluate teachers vary to some extent with the gender of the teacher.

METHOD

Participants

Students enrolled at two public universities, one on the East Coast ($n = 66$) and the other in the Midwest ($n = 223$), provided the data for this study during the Spring and Fall semesters of 1998. In an effort to maximize the diversity of the sample, we recruited students through lower division sociology and psychology courses, the vast majority of which satisfied general education requirements at their institutions. The 289 respondents included 198 women and 90 men (one person did not indicate gender). A majority (69.9%) of the respondents identified themselves as European American, another 5.5% as Asian American, 4.2% as Latino(a), 3.1% African American, 0.3% as Native American. A relatively large number, 15.9%, chose not to indicate race/ethnicity. We do not have data on year in school but the courses where we recruited participants mainly serve students in their first 2 years of college.

Procedures

Our review of the literatures led us to reject standard approaches to measurement of students' evaluation criteria. Standard rating scales are unreliable, given the research reviewed above. Direct questions that ask students if they employ gender-specific standards would seriously underestimate such practices, as few students probably *consciously* hold teachers accountable to gender-specific standards. Indeed, many shifts in the standards people apply take place without conscious awareness, and some cannot be stopped, even with effort (Biernat et al., 1991).

Instead we drew on the research literature to develop an alternative measurement strategy. Research in the area of social cognition reveals that people perceive the world through categories that are then used to store information in memory (Fiske & Taylor, 1991; Howard & Rothbart, 1980; Rothbart, Evans, & Fulero, 1979; Zadny & Gerard, 1974). Thus, the categories through which people recall phenomena are the categories through which they were perceived initially (cf., Bem, 1983).[5] Gender scholars who have closely examined language have found that the words that individuals choose for communication can sometimes reveal their deep-seated stereotypical catagorizations (e.g., Eitzen & Zinn, 1989; Kramer, Thorne, & Henley, 1978; Lakoff, 1975; Messner, Duncan, & Jensen, 1993).

Findings such as these suggest that we can address the question of whether students are holding teachers accountable for their performance in gender-specific ways by seeing whether students recall teachers through gendered categories. Thus, we looked to the words that students used in describing their teachers as indicators of the perceptual categories on which they rely in the assessment process.

[5]These categories may not accurately describe the phenomenon but that is not relevant to this question. Stereotypes are often inaccurate and yet they direct perception.

After being told that participation was completely voluntary and unrelated to class grades and that they were free simply to hand back a blank instrument if they decided not to participate, students completed a survey individually during class time. The questionnaire consisted of a range of attitude items, including opinions on social issues, confidence in social institutions, life goals, satisfaction, and standard demographic questions. Near the end of the survey, students were asked to "Think of the worst teacher you ever had. Now print below up to four adjectives to describe that teacher." Next they were asked to do the same thing with the "best teacher" they had ever had.[6] In the next section they reported their gender, race/ethnicity, and the year in school in which they had had each teacher. Our data are the words these students used in recalling their best- and worst-ever teachers.

Analytic Method

Ours is a qualitative analysis, a semantic investigation of students' evaluation considerations. In organizing our data, we adopted a grounded theory approach (Charmaz, 1983/2004; Strauss & Corbin 1994). We worked to be as inductive as possible, to stay close to our perceptions of students' ways of talking and making meaning. Words that referred to the best-ever teacher were coded in a separate wave from those that referred to the worst-ever teacher and, in each phase, words used to refer to men were coded separately from those used to refer to women so that any perceptual categories commonly applied to one gender would not distort the coding of the other. Two researchers initially coded the words independently and resolved disagreements through discussion. Later a third researcher reviewed all of the codes and made several suggestions for revision.

We analyzed these data in two stages. First, we looked at the actual *words* and *synonym clusters* that students used to describe their best and worst teachers. Then, we organized the clusters into *dimensions* of evaluation.

[6]Of course, there is a possibility that the categories through which the students responded about the first teacher would prime them for using similar categories for the second. In deciding how to order these items we chose to protect the descriptions of the worst-ever teachers from priming effects because negative evaluations can have more serious consequences for teachers than positive ones.

RESULTS

Descriptions of Best-Ever Teachers

Of the 288 students sampled, 135 identified the best teacher they had ever had as a man, and 153 identified a woman. Students were somewhat more likely to name a teacher of the same sex as their best-ever, $\chi^2 = 14.30$, $1df$, $p = .000$. What we were interested in for this analysis were the perceptual categories through which these students recalled their teachers.

In total, 1062 words were used by the 288 students to describe their best-ever teacher. Six of the top eight words used are the same for both genders: caring, understanding, intelligent, helpful, interesting, and fair. But where the distributions differ is also interesting. The words students used most often to describe their best men teachers were: caring (5.4%), understanding (4.6%), and funny (4.4%). Best women teachers were most often described as caring (7.1%), helpful (4.6%), and kind (4.3%). Men are funny (4.4%), women are fun (3.4%); men are friendly (2.2%), women are kind (4.3%).

The top mention for both genders, caring, is a word that can have different meanings, can refer to someone who "gives a darn" about the student (i.e., an attitude), but it can also describe someone who actively does the hands-on work of taking care of a student (i.e., a behavioral pattern). Which do students mean? The gendered contrast in the other most frequently occurring adjectives suggests that it may depend on the gender of the teacher. The other two most frequently used descriptors for men teachers, understanding and funny, point to things a person *is*. In the case of women teachers the second and third most common adjectives, helpful and kind, point more to things a person *does*.

What explains the seemingly odd gender link between funny and fun?[7] We think it too may signal gendered expectations. "Funny" is a performative word which represents the result of an entertaining performance and indicates a performer and an audience. "Funny" places the focus on the performer rather than the audience. "Fun," on the other hand, is a connective word—people talk of having fun *with* someone or of someone who is fun to be *with*. Unlike "funny," the word "fun" places the focus on the

[7]A comparison of word usage by gender of student ruled out the possibility that students' gender influenced whether they used the word funny versus the word fun.

relationship between or among participants (in this case, teacher and students) and on their common experience. From a semantic standpoint, then, there seems to be a relational difference between a funny male teacher and his student *audience* and a fun female teacher and her student *participants*.

Of course, word usage is a function of an individual's vocabulary, so word choice is subject to, and limited by, individual experiences and academic opportunities. Thus, we coded the actual words students used into groups of words that share essentially the same meaning, groupings that we call synonym clusters.

Synonym Clusters

The goal in creating clusters was to combine words that meant the same thing or very nearly the same thing. We labeled each cluster with the word that seemed best to convey what we interpreted as the kernel of shared meaning among the words; we favored words that were used most often by students as we labeled the clusters. For example, the cluster *Caring* includes the following words: caring, made me feel he cared, truly cared, cared about students, concern, concerned, and genuinely concerned. The cluster *Kind* includes: kind, kind-hearted, considerate, thoughtful, and gentle (we use italics to signify the names of synonym clusters).

Table I lists the most common clusters of meanings used to describe best teachers by the gender of the teacher. Men and women were described as *Smart*, *Helpful*, and *Enthusiastic* at about the same rates. The best men teachers were more commonly described as *Funny*, *Personable*, and *Understanding*. The best women teachers were more commonly described as *Caring*, *Kind*, and *Tough*. Because *Kind* and *Caring* describe common feminine gender-stereotyped attributes, it is easy to understand why they might be used more often to describe best women teachers. However, the association between *Understanding* and best men teachers, and *Tough* and best women teachers seems counter-stereotypical. Perhaps these attributes were more salient to students precisely because they conflict with the students' expectations as they brought gendered schema to bear in remembering their teachers. A similar effect has been reported previously in social cognition research: stereotype-inconsistent information or behaviors are especially memorable to individual observers (Kashima, 2000).

Table I. Most Commonly Mentioned Attributes (Synonym Clusters) of Best Teachers by Gender of Teacher

Cluster	*f*	%
Best teacher is a man, [*n* (total words) = 500]		
Caring	36	7.2
Funny	36	7.2
Smart	34	6.8
Helpful	31	6.2
Personable	29	5.8
Understanding	25	5.0
Enthusiastic	23	4.6
:		
Encouraging	15	3.0
Fair	14	2.8
:		
Kind	10	2.0
:		
Spontaneous[a]	6	1.2
:		
Tough	3	0.6
Best teacher is a woman, [*n* (total words) = 562]		
Caring	50	8.9
Smart	35	6.2
Helpful	33	5.9
Kind	29	5.1
Enthusiastic	26	4.6
Fair	22	3.9
Encouraging	22	3.9
Understanding	21	3.7
:		
Compassionate[a]	11	2.0
Tough	11	2.0
:		
Sensitive[a]	7	1.2
Funny	7	1.2
Personable	7	1.2
:		
Giving[a]	5	0.9
:		
Attractive[a]	3	0.5

Note. In cases where an attribute was much less frequently mentioned for one gender, we included it for comparison purposes and indicate breaks in the frequency sequences with a colon (:).

[a]Gender-specific attribute (used to describe one gender only).

Some synonym clusters were completely gender-specific, i.e., some kinds of meaning were used only to describe either men or women best teachers. A semantic analysis of these gender-specific attributes reveals a difference in entitlement and in labor. Words that meant *Spontaneous* were only used to describe men teachers. Spontaneity does not take any work outside of class—to be spontaneous, a teacher needs only to show up and "go with the flow." Further,

spontaneity indicates a level of confidence, both in oneself and in one's position, the kind of confidence that is bred among White men, whose social category is valued in the prevailing culture (cf., Ridgeway, 1987).

On the other hand, words that meant *Compassionate*, *Sensitive*, *Giving*, and *Attractive* were used only to describe best women teachers. To earn descriptions like compassionate and giving, one must engage in a significant amount of emotional labor. And, whereas spontaneity most probably is limited to classroom behavior, "compassion," "sensitivity," and "giving" reflect emotional qualities and behaviors that logically extend beyond the classroom. The achievement of attractiveness might well take considerable home-work for the teachers it is used to describe (cf., Bartky, 1988; Bordo, 1993). The use of the word attractive as a descriptor for best women teachers only, although relatively infrequent, points to the ongoing use of beauty as an indicator of value for women in our society—even for teachers by their students.

The words and synonym clusters offer revealing insights into the gendered dynamics of the evaluation process for students, but the real story lies in how these words and attributes converge into dimensions of gendered expectations. We turn now to the factor structure that we believe organizes these dimensions.

Dimensions of Evaluation

In the next stages of analysis, the 47 synonym clusters were sorted into 31 factors, which represent dimensions on which teachers were being evaluated. Each factor was labeled with the adjective that seemed best to represent the meanings contained within it (we signify factor names by putting them in upper case). For example, the clusters *Brilliant, Smart*, and *Knowledgeable* were combined into a factor called INTELLIGENT. The clusters *Loving*, *Encouraging*, *Giving*, and *Compassionate* were combined into the factor NURTURING. Factor structures were constructed for each gender separately. Then we worked to make the two compatible without violating the integrity of the within-gender sorting.

In two-thirds of the cases (22/31), the resulting factors were fairly comparable in both the number of words and the kinds of words used to describe men and women teachers. We labeled these "gender-comparable" dimensions. INTELLIGENT is an example of a dimension in which the frequency, type, and range of words were relatively equivalent across teacher gender. The most common words in this dimension are intelligent and knowledgeable, although students did use a somewhat greater diversity of words to describe intelligent male teachers (including extremely intelligent, insightful, and academically minded). A total of 57 (11%) times, students described their male teachers in terms of their intelligence, using 17 different words to do so. Students referred to this attribute in describing a female teacher 58 (10%) times, using 11 different words.

In addition to categories that simply refer to the teacher's gender, nonspecific statements that the teacher was good or descriptions that had no evaluative component, other gender-comparable dimensions include CLEAR, HARD WORKER, CHALLENGING, COMMITTED, STRICT, FAIR, OPEN-MINDED, RESPECTFUL, APPROACHABLE, CARING, PERSONABLE, GOOD PERSON, POSITIVE, INTERESTED, HONEST, AVAILABLE, and AGGRESSIVE.

In some cases, however, we could not create parallel structures to organize the synonym clusters that refer to men and women teachers. We call these factors "gender-loaded" because there is a lack of equivalence in either the number of words or the kinds of words used to describe men or women teachers. In all, we identified nine distinct dimensions of evaluation in recalling best teachers that were gender-loaded: FUNNY/FUN, NURTURING, NONAUTHORITARIAN, INTERACTIVE, ENERGETIC, UNDERSTANDING, ENGAGING, ATTRACTIVE, and EASY. Table II illustrates two ways in which dimensions could be gender-loaded using the examples of FUNNY/FUN and NURTURING.

A dimension can be gender-loaded because of differences in the number, type, and range of words students use under the rubric depending on whether the description is of a man or a woman. FUNNY/FUN is one such dimension. Compare the clusters that make up this factor. Students were far more apt to describe best men teachers than best women teachers as *Funny*, and they used a greater variety of words to do so. Both men and women were "funny," "humorous," and "entertaining." Only men, however, were described as "hilarious," "witty," having a "great sense of humor," and "fun-luvin" (vs. the more formal "fun-loving" used to describe a woman). It is worth noting that "sarcastic" is considered funny when used to describe best men teachers. When this word is used to describe a woman teacher, however,

Table II. Gender-Loaded Dimensions of Evaluation Describing Best Teacher: Funny/Fun and Nurturing

Men		Women	
Cluster[a]	*f*	Cluster	*f*
Factor: Funny/Fun			
Funny	22	**Funny**	5
Humorous	6	Humorous	1
Witty	3	Fun loving	1
Entertaining	3	Entertaining	1
Sarcastic	2		
Great sense of humor	1		
Fun luvin	1		
Hilarious	1		
Fun	9	**Fun**	**19**
Enjoyable	1		
Made learning fun	1		
Factor: Nurturing			
Loving	1	**Loving**	3
Sharing	1	Personal	1
Personal	1	Sharing	1
Encouraging	8	**Encouraging**	7
Patient	5	Patient	9
Supportive	1	Nurturing	2
Assuring	1	Supportive	3
		Complimentary	1
		Giving	2
		Student oriented	1
		Always there for you	1
		Selfless	1
		Compassionate	4
		Sympathetic	3
		Empathetic	1
		Empathic	1
		Believing in me	1
		Look out for my best	1

[a]Bolded cluster words represent cluster names.

it is only when she is the worst teacher the student ever had.

We found four other dimensions in which the degree of differentiation in meaning and/or emphasis implied by the words used differed depending on the gender of the teacher. The words used to indicate a teacher who is NONAUTHORITARIAN are more detailed, include specific applications, and express higher expectations of teacher/student equality when used to describe women teachers. The kinds of words used to describe ENERGETIC men teachers lean toward enthusiasm and spontaneity, whereas the words used to describe women lean toward creative and exciting. In terms of how UNDERSTANDING a teacher is, there were more kinds of words that refer to men's listening skills and perception, whereas only women were described with words that refer to sensitivity. In the case of teachers whom students recalled as ENGAGING, the words that describe men include more intensity of feeling (inspiring, loved) than do those that describe women.

If the number of different words used, the intensity of the kinds of words used, and the degree of detail in the kinds of words used are related to students' interest in a dimension, then when these patterns vary by gender there is reason to wonder if students hold teachers accountable on that dimension to different degrees depending on their gender. In fact, we found some dimensions that were almost gender-specific.

An example is the factor NURTURING (see Table II). Whereas FUNNY/FUN expressed a masculine/feminine patterning of the content, NURTURING is predominantly a feminine dimension. Students in our study used this category of meaning much less often and used many fewer words to describe men than to describe women. Further, though *Loving* and *Encouraging* were meanings applied to both men and women, words that reflect *Giving* and *Compassionate* were only used to describe best female teachers. We found three other dimensions that were exclusively or nearly exclusively used to describe one gender. Students only described women in terms of how ATTRACTIVE they were. Only men were described as EASY. There were five different words that students used to refer to a man teacher's INTERACTIVE approach in the classroom, but only one mention of interaction was used to refer to a best woman teacher.

In summary, our analysis thus far suggests that, although there are many dimensions on which students seem to perceive their teachers comparably, how students perceive their best teachers differs somewhat with the teachers' gender. We now turn to students' descriptions of their worst-ever teachers where these contrasts sharpen a bit.

Descriptions of Worst Ever Teachers

The 288 students in our sample used a total of 1043 words to describe their worst teacher ever, 141 of whom were men and 147 of whom were women. Although students used a smaller total number of words to describe their worst-ever than their best-ever teachers (1043 vs. 1062), they used a greater diversity of words (250 vs. 167). This might be the result of a greater emotional response to bad teachers or it may be an indication that, whereas good teachers have much in common, there are many ways to be a bad teacher. The word students used most often, regardless of the gender of the teacher, was boring. The next most common words to describe men teachers

were rude and arrogant, whereas for women teachers they were mean and unfair.

Synonym Clusters

Here again we combined the actual words students used into clusters of essentially the same meaning. For example, the cluster *Mean* includes mean, sarcastically mean, malevolent, nasty, nasty attitude, wants to give bad grades, harsh, ridiculed students, ill-willed, only interested in punishing students, out to fail, vindictive, threatening, and inimical. The cluster *Boring* includes boring, dull, dry, dead, uninteresting, not interesting, mundane, and comatose.

Table III. Most Commonly Mentioned Attributes (Synonym Clusters) of Worst Teachers by Gender of Teacher

Cluster	*f*	%
Worst teacher is a man		
[*n* (total words) = 514]		
Boring	50	9.7
Rude	35	6.8
Arrogant	33	6.4
Uncaring	23	4.5
Mean	18	3.5
Insensitive	18	3.5
Ignorant	17	3.3
Unfair	16	3.1
Rigid	16	3.1
:		
Out of touch[a]	6	1.2
Cold	5	1.0
Pretentious[a]	4	0.8
Worst teacher is a woman		
[*n* (total words) = 529]		
Boring	41	7.8
Mean	33	6.2
Unfair	31	5.9
Rude	27	5.1
Rigid	26	4.9
Ignorant	19	3.6
Uncaring	18	3.4
Cold	17	3.2
Insensitive	16	3.0
Arrogant	15	2.8
:		
Bitch[a]	7	1.3
Psychotic[a]	6	1.1
:		
Unhappy[a]	4	0.8

Note. In cases where an attribute was much less frequently mentioned for one gender, we included it for comparison purposes and indicate breaks in the frequency sequences with a colon (:).

[a]Gender-specific attribute (used to describe one gender only).

Table III lists the most common clusters of meaning used to describe worst teachers by gender of the teacher. The worst men teachers were most commonly described by words that mean *Boring*, *Rude*, *Arrogant*, *Uncaring*, *Mean*, and *Insensitive*. The worst women teachers were most commonly described by words that mean *Boring*, *Mean*, *Unfair*, *Rude*, *Rigid*, and *Ignorant*, *Boring*, *Rude*, and *Mean* were among the most frequently occurring descriptive categories for worst teachers for both men and women.

Some salient attribution categories were more common with one gender than the other. For example, words that mean *Arrogant*, *Uncaring*, and *Insensitive* were used more often to describe men, whereas words that mean *Unfair*, *Rigid*, and *Ignorant* were used more often to describe women. Some synonym clusters were completely gender-specific, that is, some kinds of meaning were only used to describe either men or women worst teachers. Words that mean *Out-of-touch* and *Pretentious* were used to describe only men teachers; words that mean *Bitch*, *Psychotic*, and *Unhappy* were used to describe only women teachers.

Dimensions of Evaluation

Next, the 68 synonym clusters (which include the 1043 words used by the students to describe their worst teachers) were grouped by commonality into 27 different dimensions of evaluation. As in the case of the best teachers, we then assessed and categorized these dimensions as gender-comparable or gender-loaded.

In about two-thirds of the cases (18/27), we found the dimensions roughly gender-comparable in both the number of words and kinds of words used to describe worst male and female teachers. The dimension IGNORANT is an example. Students used about the same number of different words to describe men and women, and in each case there were 25 mentions in this category. Among the references that had an evaluative component, other dimensions that seem to be comparable across gender of the teacher are CONFUSING, DEMANDING, EASY, BORING, LAZY, RIGID, WEAK, UNCARING, UNHELPFUL, INTIMIDATING, UNCONFIDENT, UGLY, AGE, and UNCOOL.

Nine of the factors that distinguished dimensions of evaluation for worst teachers are gender-loaded. RUDE (Table IV) is an example of a gender-loaded dimension in which the number, type, and range of words vary markedly with the gender

Table IV. Gender-Loaded Dimensions of Evaluation Describing Worst Teacher: Rude and Mean

Men		Women	
Cluster	*f*	Cluster	*f*
Factor: Rude			
Angry	4	**Angry**	6
Grouchy	1	Temper	2
		Bad temper	1
		Temperamental	1
		Ill-tempered	1
		Irritable	1
		Cranky	1
		Snappy	1
Loud	2	**Loud**	4
Obnoxious	2	Smart ass	1
Smart alec	1		
Sarcastic	1		
Loud mouth	1		
Juvenile	1		
Rude	28	**Rude**	17
Disrespectful	3	Inconsiderate	3
Inconsiderate	4	Direspectful	2
		Not respecting of students	1
		Thoughtless	1
		Insolent	1
		Insulting	1
		Lack of respect	1
Hypocritical	2	**Hypocritical**	2
Insincere	2	Manipulative	2
Dishonest	1	Phony	1
Manipulative	1	Fake	1
		Insincere	1
		Deceiving	1
Factor: Mean			
Mean	14	**Mean**	23
Malevolent	1	Harsh	1
Wants to give bad grades	1	Sarcastically mean	1
Nasty attitude	1	Ridiculed students	1
Harsh	1	Ill-willed	1
		Only interest punish students	1
		Nasty	1
		Out to fail	1
		Vindictive	1
		Threatening	1
		Inimical	1
		Bitch	3
		Bitchy	1
		Bitch toward male students	1
		Witch	1
		Feminazi	1
Cruel	2	**Cruel**	1
		Abusive	1
		Hateful	1
Aggressive	2	**Aggressive**	1
		Overly aggressive	1
		Argumentative	1
		Overbearing	1
		Provocative	1

Note. Bolded cluster words represent cluster names.

of the teacher. For example, although students used the same *number* of words to describe RUDE male and female worst teachers (54 each), they utilized a greater *diversity* of words to describe RUDE women than to describe RUDE men (24 vs. 15).

A comparison of the connotations of the clusters within the factor RUDE also helps to illuminate the gender differences. Although the cluster *Hypocritical* is somewhat gender-equivalent, the remaining three clusters of *Angry*, *Loud*, and *Rude* reveal gender disparities. Women teachers were described as *Angry* more often and with a greater range of words than were men teachers. Whereas men teachers were described as angry and grouchy, women teachers were described as angry, irritable, cranky, ill-tempered, snappy, temperamental, and as having a temper or an ill-temper. Conversely, men teachers were more likely to be described as *Loud* and related synonyms. Still, although one *Loud* man teacher was called a "smart alec," his *Loud* female counterpart was called a "smart ass," which perhaps again indicates the greater disrespect or even hostility leveled at bad women teachers. Finally, whereas men teachers were called *Rude* more often than women teachers, a greater diversity of words was used to describe *Rude* women, including "insolent," a word usually used to describe the behavior of a subordinate toward a superior.

On the other hand, the dimension MEAN (see Table IV) is an illustration of a gender-loaded dimension that is nearly exclusively applied to one gender, based upon gender differences in the frequency, type, and range of words that make it up. MEAN is a dimension of evaluation that is much more elaborated in the case of women teachers. To describe worst men teachers as MEAN, students used 7 different words 22 times. To describe worst female teachers as MEAN, students used 24 different words 48 times. Another way to put this is that students described women teachers as MEAN twice as often as men teachers, and used three times as many words to do so.

A comparison of the clusters within this factor elucidates the differences further. The factor MEAN is made up of four different clusters: *Mean*, *Bitch*, *Cruel*, and *Aggressive*. Although the cluster of *Cruel* is somewhat gender-equivalent, the other three clusters are gender-loaded. The least loaded of these clusters is *Aggressive*, for which a greater diversity of words are used to describe women teachers than men teachers. We see more gender-loading in the cluster *Mean*, where women are more likely to

be called "mean" (literally) and in which a much greater diversity of words are used to describe *Mean* women teachers. Although both worst men and women teachers were called mean, harsh, and nasty, only women teachers were called sarcastically mean, ill-willed, vindictive, threatening, inimical, and described as out to ridicule, fail, and punish students. Finally, perhaps the most obvious gender-loaded cluster in this factor is *Bitch*, which occurred for only women teachers and has no equivalent counterpart for men. The words used by students to describe their women teachers that are represented by this cluster—bitch, bitchy, bitch toward male students, witch, and feminazi—seem particularly angry and reveal very specific attacks on women teachers *as women*, rather than merely as bad teachers. No equivalently insulting and gender-specific slang terms were used to describe men teachers.

The remaining gender-loaded factors are PSYCHOTIC, COLD, NEGATIVE, TOO INTELLIGENT, UNFAIR, DISENGAGED, and ARROGANT. PSYCHOTIC is a purely feminine-loaded factor in that students described worst women (but not men) teachers with words that communicated that the teacher was crazy and out of control. The factors COLD and NEGATIVE are also strongly feminine-loaded. These factors include words that reflect teachers' perceived lack of emotional labor and/or emotional performance. For instance, the words that make up the factor NEGATIVE include negative, moody, emotional, and unhappy. In the case of both factors, students used far more words—both in number and diversity—to describe worst women teachers in these kinds of terms.

The factors UNFAIR and TOO INTELLIGENT skew toward women, too, but in a more nuanced way. For example, although students were almost twice as likely to use words that mean unfair to describe worst women teachers than worst men teachers (42 vs. 27), the diversity of the kinds of words used was fairly similar (11 different words to describe men, 13 to describe women). This was, in part, the result of women being described with more words that reflect generalized unfairness, whereas men were specifically described as being gender-biased (e.g., sexist, male chauvinist). The factor TOO INTELLIGENT is interesting in that, although both men and women worst teachers were described as being intelligent, only women teachers were described as being too much so.

Finally, in a departure from the trend of feminine-loaded dimensions, the remaining gender-loaded factors, ARROGANT and DISENGAGED, are strongly loaded for men. In the case of both factors, students used more words and a broader diversity of words to describe worst men teachers in these ways. In addition, both factors include synonym clusters that were present only for worst men teachers. Specifically, only DISENGAGED men were described with words that mean out of touch, and only ARROGANT men were described with words that mean pretentious.

In summary, again there is reason to ask whether students hold their teachers accountable to gendered expectations. In some cases it seems that the substance of a dimension changes depending on the gender of the teacher, and in other cases teachers' gender seems to influence whether the dimension is even applicable. Now we turn to the final questions: what are the most common categories of recall of these teachers, and do these differ by gender of the teacher?

Most Salient Dimensions of Evaluation of Best and Worst Teachers

Table V reports the most common dimensions of evaluation students used in recalling their best and worst ever teachers. In the top panel we report the data for students' best-ever teachers. Here, the kinds of dimensions students used seem to be comparable across gender. Students recalled teachers as CARING, INTELLIGENT, ENERGETIC, UNDERSTANDING, and ENGAGING. In the cases of ENERGETIC, UNDERSTANDING, and ENGAGING, the frequency of mentions of a dimension are nearly identical for both genders. However, as we have previously shown, each of these are gender-loaded dimensions, i.e., the kinds of words students used convey differences in 25 emphases (men are much more engaging; women are more intensely energetic).

The distribution of other dimensions also seems to indicate gender differences in emphasis. CARING and INTELLIGENT, the two most common dimensions, are the same for both genders, but "caring" was used more often to describe women than men (19.9% vs. 15.3% of mentions), and "intelligent" was used slightly more often to describe men than women (11.4% vs. 9.8%). Overall, the dimensions with the greatest percentage difference in word frequencies between genders were FUNNY/FUN, INTELLIGENT, PERSONABLE, CARING, NURTURING, and FAIR. We found earlier that the first three of these are gender-loaded

Table V. Most Common Dimensions of Evaluation in Recalling Best and Worst Teachers

Factor	*f*	%	Factor	*f*	%
Best teacher is a man [*n* (total words) = 500]			Best teacher is a woman [*n* (total words) = 562]		
Caring	77	15.3	Caring	112	19.9
Intelligent	57	11.4	Intelligent	55	9.8
Funny/fun	50	10.0	Energetic	51	9.0
Energetic	45	9.0	Nurturing	43	7.6
Understanding	33	6.6	Understanding	34	6.0
Personable	29	5.8	Engaging	29	5.1
Engaging	28	5.6	:		
:			Funny/fun	27	4.8
Nurturing	18	3.6	:		
:			Fair	25	4.4
Fair	15	3.0	Personable	24	4.3
:			:		
Easy[a]	2	0.4	Attractive[a]	3	0.5
Worst teacher is a man [*n* (total words) = 514]			Worst teacher is a woman [*n* (total words) = 529]		
Boring	81	15.8	Uncaring	72	13.6
Uncaring	70	13.6	Boring	63	11.9
Rude	54	10.5	Rigid	54	10.2
Self-centered	50	9.7	Rude	53	10.0
Confusing	41	8.0	Mean	49	9.3
Rigid	40	7.8	Unfair	44	8.3
Unfair	29	5.6	Confusing	38	7.2
Disengaged	27	5.3	Self-centered	25	4.7
:			Cold	17	3.2
Mean	22	4.3	:		
:			Disengaged	14	2.6
Cold	5	1.0	:		
			Psychotic[a]	6	1.1

Note. In cases where there was a considerable gender difference in factor frequency, we included the factor for comparison purposes and indicate breaks in the frequency sequences with a colon (:).
[a]Gender-specific factor (occurred for one gender only).

masculine, and the latter three are gender-loaded feminine. The students thought that men teachers are FUNNY and PERSONABLE, whereas women are NURTURING and ENGAGING.

The most common dimensions of recall for students' worst-ever teachers are reported in the bottom panel of Table V. Both male and female worst teachers were most often described as BORING, RUDE, UNCARING, and MEAN, although men were most frequently accused of being BORING, whereas women were most frequently accused of being UNCARING. The factors where the percentage difference in word frequencies differed the most depending on the gender of the teacher are ARROGANT, BORING, DISENGAGED, all gender-loaded masculine, and MEAN, UNFAIR, RIGID, COLD, and PSYCHOTIC, all gender-loaded feminine.

Some of these students were describing professors or instructors whom they had had in college, whereas others were describing teachers they had had in high school or earlier. To see whether students seemed to rely on different kinds of expectations for their college teachers we cross-tabulated each of the dimensions of evaluation (31 for best teachers and 27 for worst teachers) by when the student had had the teacher. There was striking consistency: of a total of 58 tables, only two indicated significant differences, and in both cases students were describing their worst-ever teachers. Thirty percent of the students who described precollege teachers referred to their being MEAN, whereas only 15% of those who described college teachers used that dimension, $\chi^2 = 8.99$, $df = 1$, $p = .003$. None of those few students who recalled a teacher as UGLY were describing a college teacher $\chi^2 = 4.14$, $df = 1$, $p = 0.042$.

DISCUSSION

Although we found considerable overlap in the ways students talk about their men and women teachers, we also found some signs of gender divergences. Students recalled their best-ever teachers as

caring, intelligent, and energetic, but they remember more of their best men teachers as funny, whereas their best women teachers were more likely to be described as caring and nurturing. Students described their worst ever teachers as uncaring, boring, and rude, but their worst men teachers were more often recalled as boring and self-centered, whereas their worst women teachers were more often recalled as rigid, mean, and unfair.

One potential explanation for differences in students' recall of their men and women teachers is that there might actually be significant gender differences in teaching styles. Could it be that women teachers are just more nurturing and men teachers are just more entertaining? Certainly many people believe that men and women have innately different personality traits. In a Gallup Poll (February 21, 2001) roughly contemporaneous with our data collection, interviewers listed personality characteristics and asked respondents to indicate whether they were more likely to characterize men or women. Gallup respondents were far more likely to describe women than men as emotional, affectionate, talkative, patient, and creative. They were also far more likely to describe men than women as aggressive and courageous. Women and men were equally likely to be described as intelligent.

However, look again at the findings presented in the top and bottom panels of Table V. The best men and women teachers were caring; the worst were uncaring. The best women teachers were caring and nurturing; the worst women teachers were mean, that is, not nurturing. The best men teachers were intelligent, fun, and energetic; the worst men teachers were boring, that is, not entertaining. If women were one personality type and men were another, then why did we find both poles of a dimension represented within each gender? We believe that the polar relationships between the terms used to talk about the best and worst of a gender are evidence that students do, at least to some extent, use gender-specific expectations in evaluating their teachers.

Men teachers are more likely to be held to an entertainer standard: are they funny, or is their performance a failure because that are arrogant and bore their audiences? Women teachers are held to a nurturer standard: are they caring and nurturing, or are their relationships with the students a failure because they are mean, unwilling to negotiate (rigid, unfair), or hard to relate to (cold, psychotic)? Rubin (1981) reported that students considered nurturing to be an ideal trait for women teachers (more than for men teachers), and she interpreted this finding as showing students' desire for "academic momism" (p. 972).

Note that students' memories of their worst-ever teachers appear to be more emotionally charged than their memories of their best-ever teachers and that the most hostile words are saved for women teachers. The worst women teachers are sometimes explicitly indicted for being bad women through the use of words like bitch and witch. Students may not like their arrogant, boring and disengaged men teachers, but they may hate their mean, unfair, rigid, cold, and "psychotic" women teachers. These findings are substantiated by the observations of other feminist researchers who have reported incidents of student hostility toward women instructors who are perceived as not properly enacting their gender role or who present material that challenges gender inequality (e.g., Baker & Copp, 1997; Davis, 1992; Messner, 2000; Neitz, 1985).

If these findings are accurate, then students' gendered expectations place burdens on both men and women teachers, but the burdens on women are likely to be far more consuming of time and energy. Once the anecdotes have been written, the power point slides composed, and the jokes incorporated, the same lectures can be presented with little or no modification from one semester to the next. On the other hand, responsiveness to student X does not save any labor from the need to be responsive to student Y. Each relationship must be constructed anew and individually maintained, which requires additional time in the office or in other formats to be available to and work with individual students, listening to them, supporting them, and helping them. That is, women teachers may be called on to do more of what sociologists call emotional labor (England et al., 1994; Hochschild, 1983), labor that is frequently invisible and uncounted.

Thus, if teachers are being held accountable to, and are attempting to meet, gendered standards, then women and men may be putting out very different levels of effort to achieve comparable results. If it takes more for a women to get a 5 and she nearly kills herself to do it, that difference in effort will not be measurable on student rating scales. We call this the "Ginger Rogers effect," borrowing from the observation of Ann Richards, the former governor of Texas, who noted that Ginger Rogers, one-half of the famous dance-team of 1930s movies, had to do everything Fred Astaire did, only she had to do it backwards and in high heels. Yet, even if Ginger was

exerting more effort, Fred still got most of the credit for their performance.

It is important to emphasize that our research is exploratory. The samples were not randomly drawn and, like any qualitative analysis, the interpretations were shaped by the standpoint of the researchers who did them.[8] The point of this project was to identify an area of inquiry that needs to be continued in other contexts and elaborated on using other techniques. We are in the process of one such effort now as we conduct in-depth interviews with students in order to check the reliability of and extend our current semantic findings. We urge other researchers to do similar projects.

However, we believe these data are consistent with our reading of the implications of the literatures on the sociology of gender, on social cognition, and on the student evaluation of teaching. Together they raise concern that underlying the apparently equivalent evaluation procedures there is covert gender bias. At the very least, student ratings need to be interpreted by comparison with qualitative feedback from students, analyzed by controlling for gender of student, and even then interpreted cautiously, with a sensitivity to gender stereotyping.

Administrators' increasing reliance on student ratings of college teachers in making personnel decisions has generated concern about the fairness and the wisdom of this practice on the part of teachers and those who advocate for them (e.g., Trout, 2000). We see our research as contributing to the concern, but as also offering at least a partial corrective to the "common wisdom" of a singular quantitative assessment strategy. We echo Aleamoni's (1999) call that academics cease using global measures of the form "overall, s/he is an effective teacher," given the existing evidence that such measures invite (and obscure) perceptual biases. We urge those involved in evaluating teachers and teaching to find ways to assess teaching effectiveness that are focused on the goals and outcomes of the course, which would not be as vulnerable to the students' gendered expectations for the teacher.

ACKNOWLEDGMENT

We thank Diane Kobrynowicz for her help with data collection and analysis.

[8]We would argue the same limitations apply to quantitative research (see Sprague, 2005).

REFERENCES

Aleamoni, L. (1999). Student rating myths versus research facts from 1924 to 1998. *Journal of Personnel Evaluation in Education*, *13,* 153–166.

Bachen, C., McLoughlin, M., & Garcia, S. (1999). Assessing the role of gender in college students' evaluations of faculty. *Communication Education*, *48,* 193–210.

Baker, P., & Copp, M. (1997). Gender matters most: The interaction of gendered expectations, feminist course content, and pregnancy in student course evaluations. *Teaching Sociology*, *25,* 29–43.

Bartky, S. (1988). Foucault, femininity, and the modernization of patriarchal power. In L. Quinby & I. Diamond (Eds.), *Feminism and Foucault: Paths of resistance* (pp. 61–86). Boston, MA: Northeastern Univ. Press.

Basow, S. A. (2000). Best and worst professors: Gender patterns in students choices. *Sex Roles*, *43,* 401–417.

Bem, S. L. (1983). Gender schema theory and its implications for child development: Raising gender-aschematic children in a gender-schematic society. *Signs*, *8,* 598–616.

Bennett, S. K. (1982). Student perceptions of and expectations for male and female instructors: Evidence relating to the question of gender bias in teaching evaluation. *Journal of Educational Psychology*, *74,* 170–179.

Biernat, M. (1995). The shifting standards model: Implications of stereotype accuracy for social judgment. In Y. T. Lee, L. Jussim, & C. McCauley (Eds.), *Stereotypes: Perspectives on accuracy and inaccuracy* (pp. 87–114). Washington, DC: American Psychological Association.

Biernat, M., & Kobrynowicz, D. (1997). Gender- and race-based standards of competency: Lower minimum standards but higher ability standards for devalued groups. *Journal of Personality and Social Psychology*, *72,* 544–557.

Biernat, M., & Manis, M. (1994). Shifting standards and stereotype-based judgments. *Journal of Personality and Social Psychology*, *66,* 5–20.

Biernat, M., Manis, M., & Nelson, T. E. (1991). Stereotypes and standards of judgment. *Journal of Personality and Social Psychology*, *60,* 485–499.

Bordo, S. (1993). *Unbearable weight: Feminism, western culture, and the body*. Berkeley, CA: University of California Press.

Bray, J., & Howard, J. (1980). Interaction of teacher and student sex and sex role orientations and student evaluations of college instruction. *Contemporary Educational Psychology*, *5,* 241–248.

Burns-Glover, A., & Veith, D. (1995). Revisiting gender and teaching evaluations: Sex *still* makes a difference. *Journal of Social Behavior and Personality*, *10*(6), 69–80.

Carli, L. L. (1990). Gender, language, and influence. *Journal of Personality and Social Psychology, 59,* 941–951.

Cashin, W. E. (1999). Student ratings of teaching: Uses and misuses. In P. Seldin (Ed.), *Changing practices in evaluating teaching: A practical guide to improved faculty performance and promotion/tenure decisions* (pp. 25–40). Boston, MA: Anker.

Chafetz, J. (Ed.). (1999). *Handbook of the sociology of gender*. New York: Plenum.

Charmaz, K. (1983/2004). Grounded theory. In S. N. Hesse-Biber & P. Levy (Eds.), *Approaches to qualitative research: A reader on theory and practice* (pp. 496–521). Oxford: Oxford University Press.

Connell, R. W. (1995). *Masculinities*. Berkeley, CA: University of California Press.

Davis, N. (1992). Teaching about inequality: Student resistance, paralysis, and rage. *Teaching Sociology*, *20,* 232–238.

Eitzen, S., & Zinn, M. B. (1989). The de-athleticization of women: The naming and gender-marking of collegiate sport teams. *Sociology of Sport Journal*, *6,* 362–370.

England, P., Herbert, M. S., Kilbourne, B. S., Reid, L. L., & Megdal, L. M. (1994). The gendered valuation of occupations and skills: Earnings in the 1980 census occupations. *Social Forces*, *73*(1), 65–99.

Feldman, K. (1992). College students' views of male and female college teachers: Part I—Evidence from the social laboratory and experiments. *Research in Higher Education*, *33,* 317–373.

Feldman, K. (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,* 151–211.

Fernandez, J., & Mateo, M. (1997). Student and faculty gender in ratings of university teaching quality. *Sex Roles*, *37,* 997–1003.

Ferree, M. M., Lorber, J., & Hess, B. B. (Eds.). (1999). *Revisioning gender*. Thousand Oaks, CA: Sage.

Ferree, M. M., & McQuillan, J. (1998). Methodological and policy issues in university salary studies. *Gender and Society*, *12,* 7–39.

Fiske, S. T., & Taylor, S. E. (1991). *Social cognition*. New York: McGraw-Hill.

Freeman, H. (1994). Student evaluations of college instructors: Effects of type of course taught, instructor gender and gender role, and student gender. *Journal of Educational Psychology*, *86,* 627–630.

Gallup Poll online. (2001, February 21). *Americans see women as emotional and affectionate, men as more aggressive*. Retrieved from February 27, 2001 http://www.gallup.com/poll/releases/pr010221.asp

Hochschild, A. (1983). *The managed heart: Commercialization of human feeling*. Berkeley, CA: University of California Press.

Howard, J. W., & Rothbart, M. (1980). Social categorization and memory for in-group and out-group behavior. *Journal of Personality and Social Psychology*, *38,* 301–310.

Kashima, Y. (2000). Maintaining cultural stereotypes in the serial reproduction of narratives. *Personality and Social Psychology Bulletin*, *26,* 594–600.

Kobrynowicz, D., & Biernat, M. (1997). Decoding subjective evaluations: How stereotypes provide shifting standards. *Journal of Experimental Social Psychology*, *33,* 579–601.

Kobrynowicz, D., & Biernat, M. (1998). Considering correctness, contrast, and categorization in stereotyping phenomena. In R. S. Wyer Jr. (Ed.), *Stereotype activation and inhibition: Advances in social cognition* (pp. 109–126). Mahwah, NJ: Erlbaum.

Kramer, C., Thorne, B., & Henley, N. (1978). Perspectives on language and communication. *Signs*, *3,* 638–651.

Lakoff, R. (1975). *Language and women's place*. New York: Harper & Row.

Martin, E. (1984). Power and authority in the classroom: Sexist stereotypes in teaching evaluations. *Signs*, *9,* 482–492.

Martin, P. Y. (1996). Gendering and evaluating dynamics: Men, masculinities, and managements. In D. Collinson & J. Hearn (Eds.), *Men as managers, managers as men* (pp. 186–209). Thousand Oaks: Sage.

Messner, M. A. (2000). White guy habitus in the classroom: Challenging the reproduction of privilege. *Men and Masculinities*, *2,* 457–469.

Messner, M. A., Duncan, M., & Jensen, K. (1993). Separating the men from the girls: The gendered language of televised sports. *Gender and Society*, *7,* 121–137.

Miller, J., & Chamberlin, M. (2000). Women are teachers, men are professors: A study of student perceptions. *Teaching Sociology*, *28,* 283–298.

Neitz, M. J. (1985). Resistances to feminist analysis. *Teaching Sociology*, *12,* 339–353.

Rakow, L. F. (1991). Gender and race in the classroom: Teaching way out of line. *Feminist Teacher*, *6,* 10–13.

Ridgeway, C. L. (1987). Nonverbal behavior, dominance, and the basis of status in task groups. *American Sociological Review*, *52,* 683–694.

Rothbart, M., Evans, M., & Fulero, S. (1979). Recall for confirming events: Memory processes and the maintenance of social stereotyping. *Journal of Experimental Social Psychology*, *14,* 237–255.

Rubin, R. B. (1981). Ideal traits and terms of address for male and female college professors. *Journal of Personality and Social Psychology*, *41,* 966–974.

Seldin, P. (1999). Current practices-good and bad-nationally. In P. Seldin (Ed.), *Changing practices in evaluating teaching: A practical guide to improved faculty performance and promotion/tenure decisions* (pp. 1–24). Boston, MA: Anker.

Sinclair, L., & Kunda, Z. (2000). Motivated stereotyping of women: She's fine if she praised me, but incompetent if she criticized me. *Personality and Social Psychology Bulletin*, *26,* 1329–1342.

Siskind, T., & Kearns, S. (1997). Gender bias in the evaluation of female faculty at the citadel: A qualitative analysis. *Sex Roles*, *37,* 495–525.

Sprague, J. (2005). *Feminist methodologies for critical researchers: Bridging differences*. Walnut Creek, CA: Alta Mira/Rowman & Littlefield.

Sprague, J., & Kobrynowicz, D. (1999, August). *Gender and the evaluation of teachers*. Paper presented at the meetings of the American Sociological Association, Chicago, IL.

Steinberg, R., & Haignere, L. (1987). Equitable compensation: Methodological criteria for comparable worth. In C. Bose & G. Spitze (Eds.), *Ingredients for women's employment policy* (pp. 157–182). Albany, NY: State University of New York Press.

Strauss, A., & Corbin, J. (1994) Grounded theory methodology: An overview. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of qualitative research* (pp. 273–285). Thousand Oaks, CA: Sage.

Tatro, C. (1995). Gender effects on student evaluations of faculty. *Journal of Research and Development in Education*, *28,* 169–173.

Trout, P. (2000, July/August). Flunking the test: The dismal record of student evaluations. *Academe*, *86*, 58–61.

West, C., & Fenstermaker, S. (1993). Doing difference. *Gender and Society*, *9,* 8–37.

West, C., & Zimmerman, D. (1987). Doing gender. *Gender and Society*, *1,* 125–151.

Wheeless, V., & Potorti, P. (1989). Student assessment of teacher masculinity and femininity: A test of the sex role congruency hypothesis on student attitudes toward learning. *Journal of Educational Psychology*, *81,* 259–262.

Zadny, J., & Gerard, H. B. (1974). Attributed intentions and informational selectivity. *Journal of Experimental Social Psychology*, *10,* 34–52.