AN EXAMINATION OF RACE AND SEX-BASED BIASES IN PROFESSIONAL EMPLOYEE PERFORMANCE EVALUATIONS

<table>
<thead>
<tr>
<th>Journal:</th>
<th>Academy of Management Journal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manuscript ID:</td>
<td>draft</td>
</tr>
<tr>
<td>Manuscript Type:</td>
<td>Original Manuscript</td>
</tr>
</tbody>
</table>
AN EXAMINATION OF RACE AND SEX-BASED BIASES IN PROFESSIONAL
EMPLOYEE PERFORMANCE EVALUATIONS

ABSTRACT

This paper examines whether the relationship between objective and subjective performance is stronger for professional employees who belong to certain demographic groups. We hypothesize and find that customer ratings of employee performance quality are influenced by the combination of observable employee physical characteristics (e.g., race, sex) and the degree to which employees display behaviors that benefit customers. For white or male employees, the display of customer centered behaviors increased customers’ ratings of employee performance quality. However, for non-white or female employees, customer-centered behaviors exerted a negligible, or in some cases negative, influence on perceptions of employee performance quality.
As the U.S. economy matures, the manufacturing industry is shrinking and the service industry continues to expand. Changes in the labor force mirror these macroeconomic shifts. The fastest growing segment of the labor force is professional workers\(^1\) such as doctors, lawyers, and professors. Moreover, according to the Bureau of Labor Statistics, professional occupations will continue to be the fastest growing segment of the U.S. workforce for the foreseeable future (Hecker, 2005). Professionals already comprise over 20% of the workforce and are expected to increase by another 5.3 million by 2014 (Hecker, 2005).

The increase in professional jobs is also changing the nature of performance evaluation. In non-professional jobs, organizations typically rely on supervisors to assess employee performance because they are viewed as the most qualified and knowledgeable source for evaluating the employee. However, managers may have difficulty evaluating the skills of professional employees because their jobs involve complex skills that include pleasing customers. Pleasing customers is important for a service-oriented organization because it is assumed that positive service experiences should yield greater profitability. For this reason customer ratings have become an increasingly popular source of performance feedback for professionals (Hagan, Konopaske, Bernardin, & Tyler, 2006). For example between 1998 to 2003, the percentage of service organizations giving their employees feedback from customers increased from 10 to 30 percent (Heathfield, 2006).

The willingness of organizations to place greater weight on customer satisfaction as a way of assessing employee performance can be found in organizations ranging from airlines

---

\(^1\) We use the Bureau of Labor Statistics definition of professional to mean any person engaged in work predominantly intellectual and varied in character as opposed to routine mental, manual, mechanical, or physical work; involving the consistent exercise of discretion and judgment in its performance; of such a character that the output produced or the result accomplished cannot be standardized in relation to a given period of time; requiring knowledge of an advanced type in a field of science or learning customarily acquired by a prolonged course of specialized intellectual instruction and study in an institution of higher learning or a hospital, as distinguished from a general academic education or from an apprenticeship or from training in the performance of routine mental, manual, or physical processes.
(Jenkins, 1992) to universities (Davis & Davis, 1999). Some business leaders (Bracken, Church, & Timmreck, 2001) and researchers (Salam, Cox, & Sims, 1997) have applauded the use of customer ratings because they believe that aggregated evaluations are highly reliable measures of employee performance quality. Other observers have questioned their use, pointing to research showing that biases based on racial, gender, or other social categorization processes are unavoidable when using subjective evaluations of performance, particularly when they come from naïve and inexperienced raters (Pulakos, White, Oppler, & Borman, 1989; Wilkinson & Fontaine, 2002; Woehr & Roch, 1996).

The attempt to detect racial and gender biases in performance evaluation in general has proven to be a theoretical and methodological challenge because of one major problem: employees’ rated performance scores can be viewed as capturing employees’ true performance scores. As Rotundo and Sackett (1999) write, “There is no definitive way of determining whether the criterion used in a validity study is biased. Thus, there is no current method of establishing whether there is bias in performance ratings (Rotundo & Sackett, 1999: 816).” The results of the studies themselves have also been inconclusive. For example, in their comprehensive review of research on personnel selection Landy and colleagues concluded that: “No one seems to have found direct gender or race effects, at least in large-scale field studies (Landy, Shankster, & Kohler, 1994: 283).” Reviewing this same literature, Latham and Wexley (1994) reached a similar conclusion, stating that: “[W]hen the appraiser uses behaviorally based appraisal scales, ratee characteristics, such as age, race, and sex, have a negligible effect on the resulting performance appraisal” (p. 152). Indeed, researchers have found it difficult to reach definitive conclusions about how demographic characteristics might influence the performance
evaluation process (e.g., Barrett & Morris, 1993; Ellis, Ilgen, & Hollenbeck, 2006; Sackett & DuBois, 1991; Vecchio, 2002).

Despite the failure to demonstrate that racial and gender biases systematically influence performance evaluations, there are theoretical reasons to expect that in jobs where the assessment of performance has unavoidable subjective elements and where consensus on what constitutes excellent performance is difficult to reach – as is the case in most professional occupations – stereotypes and other unconscious attitudes and beliefs may sometimes be used by evaluators to “fill in the blanks.” As a result, we hypothesize that a significant amount of variance in performance evaluation may be attributable to these cognitive processes. The theoretical bases for this hypothesis are drawn from several converging streams of research showing that while overt expressions of prejudice and discrimination in U.S. society are now relatively rare, negative beliefs, stereotypes, and attitudes towards women and ethnic minorities persist in more subtle and often unconscious forms (e.g., Gaertner & Dovidio, 1986; Kinder & Sears, 1981). For example, studies show that many people hold a male leader prototype (Banaji & Greenwald, 1995; Eagly, 1987) and a generalized belief that men are superior and deserve to control and receive more resources than women (Jost & Banaji, 1994). Similarly, people who report low levels of overt prejudice against blacks have nevertheless been shown to recommend hiring a white compared to a black applicant when the two candidate’s qualifications were ambiguous (Dovidio & Gaertner, 2000), indicating the operation of what some scholars (e.g., Gaertner & Dovidio, 1986) refer to as “aversive racism.” What these studies suggest is that potential biases in the performance evaluation process may be motivated not by what people say they believe, but by stereotypic beliefs or attitudes that are not expressed or that may be unknown even to the raters.
If stereotypes and unconscious biases influence social cognition, as many studies over the years have shown, then the theoretical argument that the demographic characteristics of employees may influence how other people perceive, interpret, and evaluate their performance seems tenable. The purpose of our study was to test this possibility by examining whether customer service-oriented behaviors exhibited by highly skilled, professional employees and customer ratings of those same employees’ service quality vary, either in strength or direction, depending on the employee’s race and sex. We argue that if this relationship does in fact vary as function of the employee’s race and sex it could indicate a possible source of customer evaluation bias arising from negative stereotypes and beliefs they hold about certain demographic groups.

In our study, the professional service employees were physicians working for a large health maintenance organization; the customers were their patients. By shifting the focus of subjective performance evaluations to customer rather than supervisor ratings, our study overcomes a major methodological limitation of past research on performance evaluation because the customer-raters had no knowledge of what types of physician behaviors that organization valued and rewarded. Hence, customers’ evaluations should have been based solely on what they observed and experienced in their encounters with physicians. Moreover, our design is an improvement over earlier studies of performance evaluation which measured employee performance with a single subjective rating. In those types of studies, there is no standard by which to determine if the rating is biased (Rotundo & Sackett, 1999). Importantly, both the objective (service-oriented behaviors) and subjective (customer ratings of professional quality) performance measures were used by the focal organization to determine employee salaries which meant that any systematic biases in these measures could translate directly into
differences in pay and the generation of gender and racial inequality in other career outcomes.

The following sections present theoretical arguments for why we expect the same customer-oriented behaviors performed by physicians to be differentially related to patients’ evaluations of physician performance as a function of the latter’s race and sex.

THEORETICAL BACKGROUND AND HYPOTHESES

The classification of other people into categories based on ascribed characteristics like race or sex appears to be a universal human response to social information (Devine, 1989; Fiske & Neuberg, 1990; Hewstone, Hantzi, & Johnston, 1991). Social categorization, like other cognitive heuristics, is useful and adaptive for simplifying information processing and bringing order to the world. But categorization can also lead to biased or inaccurate judgments about those whom we assign to a given social category and, importantly, categorization often occurs with little conscious reflection (Hewstone et al., 1991). It has been shown that easily detectable physical characteristics like race and sex are among the most reliably observed bases for spontaneous social categorization (Stangor, Lynch, Duan, & Glass, 1992) and so it seems likely that judgments of employees by customer-raters can be at least partly influenced by these automatic categorizations.

One way the operation of social categorization biases in a performance evaluation process might be revealed is when identical behaviors exhibited by employees show a different relationship to customers’ subjective evaluations of service quality as a function of the rated employees’ race or sex. Although we do not expect the correlation between objective behaviors displayed by professionals and subjective ratings of service quality to approach unity, it seems reasonable to expect a significant, positive correlation, assuming the behaviors do in fact have a positive impact on customer welfare. The more crucial question though is whether this
relationship is stronger or weaker as a function of the race and sex of the professional being evaluated. Two theories of how group-based stereotypes influence social judgment provide an identical answer to this question. We review each of them beginning with status characteristics theory.

**Status characteristics theory**

Status characteristics theory asserts that individuals have preconceived, status-based notions about their own and each other’s competence (Berger, Conner, & Fisek, 1974; Berger, Fiske, Norman, & Zelditch, 1977). In the United States, whites and men are considered by most people to be members of a high status social group relative to women and ethnic minorities (see Ridgeway, 1991 for a review). This perception is not surprising given that members of these groups control most of the resources and occupy most of the powerful positions in U.S. society (Sidanius & Pratto, 1999). One of the benefits of belonging to a high status social group is that observers are more likely to make favorable inferences about one’s competence, normality, and legitimacy (Aquino & Bommer, 2003; Giannopoulos, Conway, & Mendelson, 2005; Sidanius & Pratto, 1999). In contrast, members of devalued groups are subject to negative stereotypes and attributions concerning their work-related competencies (Fernandez, 1981; O’Leary & Ickovics, 1992). One implication of these differences in inferential judgment is that behaviors performed by a high status group member may be interpreted more favorably by observers than the same behavior performed by someone from a low status group. Evidence for this possibility is provided by studies showing that women in leadership roles are disliked more than men in similar roles (Eagly, Makhijani, & Klonsky, 1992), that ethnic minorities and women are rewarded less than whites and males for exhibiting advice-giving or ingratiatory behaviors (Westphal & Stern, 2007), and that whites who exhibit high levels of organizational
citizenship behavior are less likely than blacks who exhibit the same level of citizenship to report being victimized by the harmful actions of co-workers (Aquino & Bommer, 2003).

We propose that ascribed characteristics like sex and race can exert the strongest influence on the performance evaluation process when subjective judgment is required. For example, one study found that when job applicant resume quality was ambiguous, applicants with African-American-sounding names (e.g. Aisha, Rasheed) were much less likely to be called for a job interview than applicants with white-sounding names (e.g. Kristin, Brad) (Bertrand & Mullaninathan, 2004). Likewise, when evaluators of orchestral positions could see the sex of the applicant they were more likely to select men. When the applicant’s sex could not be observed, the number of women hired significantly increased (Goldin & Rouse, 2000). People tend to have lower performance expectations for members of low-status groups and so it is also likely that racial minorities and women may have to achieve higher levels of performance than whites and men to be judged as having the same level of underlying ability (Biernat & Kobrnyowicz, 1997; Yarkin, Town, & Wallston, 1982). Compounding the adverse effects of status-based expectations is the fact that people appear to resist information that contradicts them (Biernat & Kobrnyowicz, 1997; Foschi, 2000). For example, individuals only acknowledge the competence of women if there is explicit evidence (Shackelford, Wood, & Worchel, 1996). When no explicit, objective evidence exists, people fit observable cues with their pre-conceived expectations. Based on status characteristics theory, we therefore expect customers to evaluate the behavior of professionals more favorably when that professional belongs to a high rather than low status demographic group. The exemplar-based model of social judgment (Smith & Zárate, 1992) supports this prediction.

The exemplar model of social judgment
According to the exemplar model, within any culture, particular person attributes come to be perceived as expected or default values. In Western society, white racial identity and male gender are treated as cultural expectations. Evidence for this “white male norm” hypothesis comes from experiments showing that membership in non-normative groups receives greater attention than membership in normative groups because of its incongruence (Smith & Zárate, 1990; Stroessner, 1996). Arguably, this normative expectation is exacerbated in the professional occupational category, which has historically been dominated by white males. For example, the 2006 Diversity Report by the Association of American Medical Colleges reports that only 24.5% of practicing physicians are women and 12.1% are non-whites.² If the exemplar model’s assumptions are valid, then customers of professionals should pay more scrupulous attention to the behavior of professionals who belong to groups that violate their normative expectation of the professional social category. As a result, customers might be more critical when forming subjective judgments of performance if the professional belongs to non-prototypical category. Some evidence for this possibility is provided by Westphal and Stern’s (2007) finding that women and ethnic minorities on corporate boards were more likely to be “punished” than their white male counterparts for carefully monitoring management initiatives and attempting to control management decision-making. In their study, being punished meant that they would have less chance of receiving a board appointment.

Combining the idea that the actions of persons who do not fit within the social category elicit heightened attentiveness with the propositions of status characteristics theory, we theorize that any perceived deficiency in professional’s behavior toward a customer elicits a more negative evaluative response, and may perhaps be more easily recalled by customers, when the

² However, the Association of American Medical Colleges also reports that these numbers are changing dramatically as 44% of American medical school graduates in 2004 were women and 34% were non-whites.
professional is female rather than male or an ethnic minority rather than a white. By extension, any positive behaviors directed toward the customer by the professional will elicit a more favorable evaluation when exhibited by a white or male than when exhibited by an ethnic minority or female. As a result of these differences in evaluative judgments, we expect the relationship between objective professional performance and subjectively measured performance to differ depending on a professional’s sex and race. Together, status characteristics theory and exemplar-based models of social judgment lead to the same conclusion, stated formally in the following hypotheses:

**Hypothesis 1:** The positive relationship between a professional employee’s objective performance, indicated by behavior directed toward the customer to improve customer welfare, and subjective evaluations of performance quality, indicated by customer ratings, is weaker for female than for male professionals.

**Hypothesis 2:** The positive relationship between a professional employee’s objective performance, as indicated by behaviors directed toward the customer to improve customer welfare, and subjective evaluations of performance quality, indicated by customer ratings, is weaker for ethnic minority than for white professionals.

**METHOD**

**Sample**

Our sample was drawn from a large health maintenance organization, hereafter referred to as Prohealth (a pseudonym). Prohealth was founded in 1947 and is non-profit and consumer-governed. Today, it provides coverage and healthcare for about 350,000 people in the Pacific Northwest region of the United States. It directly employs approximately 800 healthcare providers (both general practitioners and specialists) to care for its members. Our sample
consisted of all 113 primary-care physicians (i.e., family practitioners) at Prohealth. Within our sample, 38.4% were women, 11.5% were ethnic minorities, and all had a medical degree.

Measures

Prohealth routinely collected patient ratings to assess the quality of physician performance as well as objective behavioral indicators of physician performance that were assumed to have a direct, positive impact on patient health and well-being. This feature of our data represents a methodological improvement over many earlier studies of performance evaluation bias because in most of these studies the only measure of employee performance was a single subjective rating. In those types of studies, there is no standard by which to determine if the rating is biased (Rotundo & Sackett, 1999). The dependent variable in our study was physician performance quality as rated by patients. The independent variables were physician demographics (race and gender) and three types of patient-centered behaviors.

Customer rating of professional performance quality. Like other health plans, Prohealth measures customer satisfaction with the Consumer Assessment of Health Plans Survey. A survey was mailed to a percentage of each physician’s patients, following doctor visits, selecting the patients from a stratified sample based on the patients’ visit utilization rates (to avoid a bias toward those patients with frequent appointments). Patients completed and returned a total of 12,091 surveys for a response rate of 52%, so that each physician was rated by an average of 107 patients. Patients rated each of the three items targeting their physician on a 5-point Likert-scale (1=very poor; 5=excellent) “How would you rate…” (1) the attention the provider paid to you, (2) this provider’s thoroughness and competence, and (3) your opportunity to ask questions of this provider. The three items were highly correlated (average correlation is .93) so the organization combined them to create a composite patient satisfaction variable. The
organization did not provide us with access to raw patient-level surveys. Instead, they provided us with data indicating what percentage of each physician’s patients rated the physician as “excellent.” Thus, the range on this measure for each physician was from 0 to 100 percent. This measure was collected in the same quarter as all other variables.

**Professional race.** Prohealth identified each physician’s race, and we coded whites “0” and ethnic minorities as “1.” Of the 113 physicians in the sample, 10 were Asian or Pacific Islanders, two were black, and one was Native American. The percentage of ethnic minority physicians in our sample is consistent with the national average of 12.1%.

**Professional sex.** We also obtained this variable from Prohealth records and coded males “0” and females “1.” Forty-three of the physicians were female. Thirty-eight percent of physicians in our sample were women, which is slightly higher than the national average of 24.5%.

**Customer-centered behavior.** Patients’ welfare has always been the ultimate driving force in medicine; however, physicians have historically pursued patients’ welfare via a predominantly physician-centered perspective (Daghio et al., 2003; Laine & Davidoff, 1996). The patient’s point of view and concerns have historically been considered irrelevant or a hindrance to understanding the core issue—the patient’s disease. This physician-centered approach to medical care involves focusing on doctors’ convenience, rather than focusing on patients’ convenience. For example, patients historically have been expected to wait patiently for doctors, and to never keep the doctor waiting (Mishler & Waxler, 1963).

However, with the growing prevalence of health maintenance organizations (HMOs) and the increasing corporatization of medicine (Feinglass & Salmon, 1990), patients are increasingly being viewed by organizational administrators and physicians as customers. Therefore,
physicians are increasingly being rewarded for engaging in customer-centered, rather than physician-centered behaviors (Laine & Davidoff, 1996; Stewart et al., 2000). We used the customer-centered behaviors identified by Prohealth as our indicator of objective physician performance.

Prohealth measures customer-centered behaviors along three dimensions. The first is physician productivity, which is the number of health procedures performed and issues discussed in a given time period. The second is the physician’s accessibility to customers measured by the number of secure emails that doctors send to customers. The third is the physician’s level of quality measured by the standardized prescription rates of particular medications for customers that possess precise disease criteria. All three dimensions reflect behaviors that benefit customers by reducing the amount of time and money customers spend receiving medical care. For all metrics, physicians are shown how they compare to both the organizational goal and the organizational average. Physician compensation is tied to each of these customer-centered behaviors. Physicians who exceed the 40th percentile are given a bonus, while those below the 40th percentile are not given a bonus.

All three customer-centered behaviors are presumed to benefit customers. More productive physicians are able to treat more customer problems per visit, which saves customers time and additional trips to the doctor. Higher physician accessibility saves customers unnecessary visits to the doctor by allowing customers to simply email any medical questions to their physician. Higher physician quality benefits customer by preventing costly and deadly health events such as strokes and heart attacks.

**Physician Productivity.** The average number of patients seen, medical issues discussed, and medical procedures performed by each doctor in a standardized 8-hour day was recorded by
the organization’s scheduling software. Prohealth physicians have a great deal of control over the amount of work that they do in a day as they can control the intensity of each visit (e.g., the number of procedures performed and patient health issues addressed per visit). The number of patients that physicians see each day is controlled by organizational administrators.

This variable was the composite of average face-to-face visits and phone visits adjusted by the average intensity of each visit. Intensity was measured by Relative Value Units (RVUs), which are coded by physicians at the end of each visit according to national coding guidelines. RVUs capture the amount of time involved, the required physical and mental effort, the required judgment and technical skill, and the psychological stress entailed (Hsiao, Braun, Becker, & Thomas, 1988; Hsiao, Braun, Dunn, & Becker, 1988). According to quarterly audits by administrators, Prohealth physicians accurately record RVU’s in 90 percent of patient visits. Coding errors resulting from physicians coding too many or too few RVU’s are normally and equally distributed. The raw measure of productivity was standardized based on the full-time status of the physician and then multiplied by each physician’s average visit intensity to obtain the quarterly average RVU-adjusted patient encounters per day.

Physician Accessibility. The average daily number of emails that physicians sent to patients for the quarter was used to measure another type of patient-centered behavior. Physician emails benefit patients because patients value the ability to easily contact their physician. Indeed, a Harris poll has shown that 90% of Americans who are online want the ability to e-mail their physicians, and 37% are even willing to pay for it (Taylor, 2002). Prohealth patients can send unlimited emails to their physician at no cost and Prohealth physicians are expected to reply to each email within 24 hours. Patients primarily contact their physician via email to ask basic health-related questions, to request prescription refills and to schedule follow up appointments.
Prohealth administrators encourage Prohealth physicians to contact patients via email because they think that an email might save patients a trip to the doctor, thereby saving patients’ time and money. In general, physicians do not think that email improves the quality of patient care, but rather that email increases convenience for patients (Kleiner, Akers, Burke, & Werner, 2002).

The Prohealth computer server automatically recorded the number of emails that each medical professional sent to his or her patients. Medical professionals had a great deal of control over how many emails they sent for two reasons: (a) they could choose whether or not to inform their patients of their email address, and (b) they could choose whether or not to personally respond to their patients’ emails. We calculated the number of emails sent per day, taking into account the number of full working days that physicians were in clinic during the time-period of this study. To enhance the normality of the variable, we used an inverse transformation and then reflected these values such that higher values represented greater use (Tabachnick & Fidell, 2003).

**Physician Quality.** Every Prohealth primary care physician is responsible for a panel of member-patients. Of the thousands of possible treatments, prescriptions, and procedures that physicians can perform to benefit patients, one of the most important is each physician’s prescription rate of statins and angiotensin-converting-enzyme (ACE) inhibitors to patients with cardiovascular disease. Treatment of cardiovascular events such as strokes, clots, and heart attacks is the biggest healthcare cost for patients in the U.S. (Willerson & Cohn, 2000), and these drugs prevent cardiovascular events over patients’ lifetimes (Gerstein et al., 2000).

Nationally, only 50% of all cardiovascular disease patients that should be treated with statins and ACE inhibitors are currently taking such medication (Dubois et al., 2002). According to Prohealth guidelines, all patients with cardiovascular disease should be regularly taking ACE
inhibitors and some form of a statin. ACE inhibitors lower blood pressure, and statins lower cholesterol. These drugs significantly lower the immediate risk of a cardiovascular event (e.g., stroke, heart attack) for all individuals, regardless of sex or previous history of cardiovascular disease (LaRosa, He, & Vupputuri, 1999; Yusuf et al., 2000). To promote a higher prescription rate, Prohealth administrators send emails to physicians to remind doctors to prescribe such treatment. While prescription of these medications benefits patients by helping patients avoid death and reduce healthcare expenses, individual physicians do not generally see their utility on a daily basis. Indeed, physicians often forget to prescribe these medications (Isles, 2002).

This variable is the composite of the percent of cardiovascular disease patients 18 years and older who were dispensed the equivalent of a 90-day supply for ACE inhibitors and statins at any time within the quarterly reporting period. The component variables approached normality and were added together. The resulting variable was each physician’s overall prescription rate of statins and ACE inhibitors for cardiovascular disease patients. The average prescription rate at Prohealth is 50%, similar to the national average.

Control Variables

We controlled for several variables that were not of direct interest for testing our hypotheses but that could be theoretically related to the dependent variable and might provide plausible alternative explanations for our findings.

Average Practice Busyness. Patients who have to wait long periods of time to see their physician may be less satisfied than those who can see their physician immediately so we controlled for the busyness of each physician’s practice. This variable is automatically calculated by the Prohealth computer server at the end of each business day and represents the average time until each physician’s third available appointment for the quarter.
Physician full-time status. Patients might be more satisfied with physicians who work more hours because these patients might be more likely to schedule a convenient appointment time with these physicians. Thus, we controlled for the number of hours a physician worked. We collected this variable from the archival records of Prohealth. Physicians ranged from working 30 to 100 percent of a full-time position.

Number of patients in panel. Each full-time physician is assigned to care for a group (i.e. panel) of roughly 2,000 patients. Physicians with large panels may have high levels of patient demand, which could influence patient satisfaction. We controlled for the total number of patients in each physician’s panel divided by the full-time status of the physician over the most recent 12-month period. This data was obtained from Prohealth’s human resource department.

Average patient age. Physicians whose patients are older may be more or less demanding than physicians whose patients are younger. We used the average patient age for each physician’s panel for the most recent quarter.

Average chronic sickness of panel. Patient panels that are more chronically ill may be more or less likely to be satisfied than patient panels that are less chronically ill so we controlled from the chronic sickness of each physician’s panel. This variable was calculated by Prohealth for each physician and captures the chronic illness of each physician’s panel (e.g., diabetes, cardiovascular disease) so that higher values represent more chronically ill patient populations. This variable excludes acute episodic illness (e.g., broken bones, injuries due to accidents).

Physician Age and Tenure. Physicians who are older or who have been employed by Prohealth for more years may have more loyal, satisfied patients than younger or recently hired physicians.
RESULTS

Table 1 reports the means, standard deviations, and correlation coefficients between the dependent, independent, and control variables.

Insert Table 1 here

We used hierarchical moderated regression models to examine the hypothesized interaction effects (Aiken & West, 1991). We centered all variables involved in the interaction terms to minimize multicollinearity between the interaction terms and their individual components (Aiken & West, 1991). We entered all of the control and independent variables in Model 1, in Model 2 we entered all hypothesized two-way interactions involving sex and the three dimensions of objective performance, and in Model 3 we entered all two-way interactions involving race and the objective performance measures. The full model includes all interaction effects. Table 2 presents the results of this analysis.

Insert Table 2 here

Hypothesis 1 states that the positive relationship between professional performance measured by objective, customer-centered behaviors and subjective evaluations of professional performance will be weaker for female than for male professionals. The two-way sex X objective performance interactions as a set explained a significant amount of incremental variance in the dependent variable ($R^2 = .08, p < .01$) providing preliminary support for this hypothesis. Inspection of the individual regression weights showed that the accessibility X sex
and quality X sex interactions were significant \((p < .05)\). We probed the pattern of the interaction by examining the simple slope of the objective performance measures for male and female professionals (Aiken & West, 1991). The results of this analysis are shown graphically in Figures 1 and 2.

---

Both figures show a stronger positive relationship between customer-centered behaviors and performance ratings for men than for women. We calculated the significance of the simple slopes for interactions (Aiken & West, 1991). For male professionals the coefficient of the simple slopes of quality on subjective performance was significant and positive \((b = .03, p < .01)\) and accessibility behaviors on subjective performance was significant and marginally positive \((b = .13, p < .10)\). However, for female professionals the coefficient of the simple slope of quality behaviors on subjective evaluations of performance is negligible \((b = -.01, \text{n.s.})\) and for accessibility behaviors is marginally negative \((b = -.13, p < .10)\). These results support Hypothesis 1 for two of the objective-subjective performance relationships.

Hypothesis 2 states that the positive relationship between professional performance of customer-centered behaviors and subjective evaluations of professional performance is weaker for non-whites as compared to whites. The two-way race X objective performance interactions as a set explained a significant amount of incremental variance in the dependent variable \((R^2 = .06, p < .05)\) providing preliminary support for our hypothesis. Inspection of the individual regression weights showed that the productivity X race and quality X race interactions were
significant \((p < .05)\). The forms of the interactions are shown graphically in Figures 3 and 4.

---

Simple slope analysis reveals that for white professionals the coefficient of the simple slope of quality on subjective performance is significant and positive \((b = .02, p < .01)\) and productivity on subjective performance is marginally significant and positive \((b = .02, p < .10)\). However, for non-white professionals the simple slope of quality behaviors on subjective evaluations of performance is marginally significant and negative \((b = -.02, p < .10)\) and significant and negative for productivity behaviors \((b = -.03, p < .01)\). These results support Hypothesis 2 for two of the objective-subjective performance relationships.

**DISCUSSION**

We set out to explore whether professional’s demographic characteristics moderate the relationship between objectively measured customer-centered behaviors and subjective ratings of performance quality. We found that customer-centered behaviors were positively related to subjective evaluations of performance quality, but *only* for whites and males. Indeed, the direct relationships between customer-centered behaviors and subjective evaluations were not significant indicating that these relationships were fully qualified by professional employee demographics. We also found that certain types of customer-centered behaviors were *negatively* related to customer evaluations of performance for women and non-whites. This was an even stronger result than we anticipated because our prediction was that the relationship between customer-centered behaviors and subjective performance would be weaker, but still positive, for women and non-whites compared to men and whites.
We suggest that the observed pattern of relationships between objective, customer-centered behaviors and subjective performance ratings by customers may indicate the operation of a stereotypic bias that leads customers to ignore or discount positive behaviors when they are performed by professionals who do not fit within the exemplar category of physician. It is also possible, based on status characteristics theory, that beliefs about the competence and legitimacy of professionals who belong to a high status group (i.e., whites, males) can lead customers to elevate subjective performance ratings in response to customer-centered behaviors relative to when these same behaviors are performed by a professional from a lower status and socially devalued group.

One way that social categorization processes might influence customer ratings of professional service quality is by leading them to draw erroneous inferences of underlying traits and abilities based on an employee’s group membership (Biernat & Kobrynowicz, 1997). For example, Biernat and Kobrynowicz (1997) showed that when judging a fictional employee, participants required blacks and women to achieve higher levels of performance than whites and men in order to make the same inferences of underlying ability. Consequently, even if female or ethnic minority professionals behave the same way as their male and white counterparts, customers may not have the same level of confidence in their ability and competence because the standards for demonstrating the underlying traits are higher. This stereotype-driven bias is particularly likely to occur in cases where judgments are highly subjective and performance cannot be objectively assessed (Glick, Zion, & Neslon, 1988).

Traditionally, organizations have relied on supervisors to assess employee performance because they were viewed as the most qualified and knowledgeable source for evaluating the employee. More recently, many organizations have supplemented the standard supervisory
appraisals with feedback from other sources like customers (e.g., Miles & Snow, 1995). Our results suggest that if customer evaluations become widely and uncritically used to determine pay and promotion opportunities, the job outcomes of women and ethnic minorities could be adversely impacted.

Our method of testing for possible bias in performance evaluations was a significant improvement over past studies. First, we used an objective performance standard so that we could compare subjective ratings to this standard and therefore determine whether the later ratings might be influenced by race and sex. Second, our subjective ratings were based on aggregated judgments from a large number of customers rather than relying on the judgments of a single supervisor. This is important because the large number of raters provides a highly reliable subjective performance rating for each professional. Finally, we controlled for several variables that could provide alternative explanations for our results, such as the baseline busyness level of each professional’s practice, and each professional’s tenure with the organization.

Given these methodological strengths of our study, it is unsettling to find that even for highly skilled employees who possess similar levels of human capital and occupational prestige, positive, customer-centered behaviors exhibited by those who belong to low status demographic groups may not translate into increased customer satisfaction ratings relative to their prototypic, high status counterparts. At Prohealth, women and non-white physicians can in fact be economically harmed by subjective performance evaluations because physicians who fail to achieve the target level of performance quality (i.e. the 40th percentile) do not receive a salary bonus. The practical implications of our results become more apparent when we examine the effect sizes in our sample. The interactions between sex, race and objective performance
explained 14.5% of the variance in subjective evaluations of physician performance. Cohen (1988) provides ballpark descriptors of effect sizes based on R-squared values—"large" (R² = .25), "medium" (R² = .09), and "small" (R² = .01). Therefore, the observed effect size of the interactive effects of race, sex, and objective performance on subjective evaluations in our sample is between medium and large.

There is a potential alternative explanation for our findings that we must reckon with before drawing firm conclusions about what we have shown. Our predictions implicitly assume that customer-centered behaviors influence customers’ subjective ratings, although we are cautious to avoid making any causal statements because our design is cross-sectional. It is possible, though, that it is subjective ratings of performance that influence professionals’ customer-centered behaviors. The potential for incorrect conclusions based on the cross-sectional design is minimized in our study because we tested interaction rather than main effects (Bowen & Wiersema, 1999). Nevertheless, the reverse-causal argument could be made and so we wanted to test this possibility statistically even though we recognize that a stringent test requires a longitudinal design.

We tested an alternative model implying reverse causation by creating two interaction terms (i.e. sex by subjective performance and race by subjective performance) and regressing each of our three objective performance measures on these variables along with the main effects and controls. When we analyzed our data this way, none of the six interactions involving the subjective ratings and physician race and sex were significant (p > .10). This does not rule out the possibility of reverse causality, but it does show that the model we specified explains our data better than an alternative model with different interaction terms.
In many ways, physicians are an ideal sample for testing our research question of whether objective service behaviors are related to subjective customer evaluations of service quality. Physicians provide a service that can literally impact their customer’s very survival. Thus, patients should be highly motivated to attend to and scrutinize the behavior of their physicians. The work physicians do also involves many subtle, rich, and complex features that are not easily amenable to quantification. Physicians not only dispense medical knowledge, but may also provide socio-emotional and psychological support during times of illness and distress. And as anyone who has had a thorough medical examination knows, the level of physical contact they have with a patient’s body is generally surpassed only in the most intimate of social relationships. For these reasons, patients’ assessments of physician performance are likely to be influenced by cognitive and emotional processes that are intuitive, automatic, and perhaps largely unconscious. This does not mean that no aspects of physician performance can be assessed objectively or that patients do not deliberate about their physician’s performance. On the contrary, the focal organization identified several objective, behavioral criteria of performance that were assumed to directly benefit patients. If these behavioral criteria do indeed contribute to patient welfare, and if patients recognize this, then it is logical to expect that physicians who exhibit more of these behaviors should be evaluated more favorably by their patients than those who exhibit less of them. However, this was not the case for all physicians in the organization we studied.

We believe that the results we found among physicians can generalize to other highly skilled, professionals because the theoretical bases for our hypotheses are not specific to a health-care context. We would expect our findings to be most strongly replicated in jobs historically held by white males like the upper echelons of organizations (e.g. Westphal and
Stern, 2007) or in other prestigious professions (e.g. lawyers, professors). In these fields, what constitutes exceptional performance is open to multiple interpretations and so the evaluation of performance would seem highly vulnerable to stereotyping and unconscious biases. Of course, whether our findings are upheld in such samples is a question for future research.

Conclusions

Over the last 40 years, women and ethnic minorities have entered occupations and attained positions of leadership in business and government that at one time were almost entirely the province of white men (Stroh, Brett, & Reilly, 1992). To some observers these trends show that U.S. society has become more egalitarian, fairer, and less discriminatory since the Civil Rights Movement of the 1960s (Gross & Hook, 1978; Lynch, 1991). Yet women and ethnic minorities are still not well represented among the ranks of highly paid managers and professionals in U.S. corporations and in prestigious occupations like law and medicine (e.g., Baldi & McBrier, 1997; Eagly & Karau, 2002; Wilson, Sakura-Lemessy, & West, 1999) and their earnings lag behind those of white men within the same occupational category (Dreher & Cox, 1996; Petersen & Morgan, 1995). Some writers have suggested that women and ethnic minorities should simply try harder (Wenneras & Wold, 1997; Yarkin et al., 1982). Yet the findings of our study suggest that it is an open question whether working longer hours, choosing demanding work assignments, or going above and beyond the minimal expectations of their jobs will necessarily help women and minorities achieve parity with their white male counterparts in the absence of other interventions that directly attempt to counteract the operation of stereotypic biases and unconscious beliefs that lead others’ to discount or ignore these efforts.
References


Heathfield, S. M.; 360 Degree Feedback: The Good, the Bad, and the Ugly
http://humanresources.about.com/od/360feedback/a/360feedback_2.htm.


Westphal, J. D. & Stern, I. 2007. Flattery will get you everywhere (especially if you are a male caucasian): Ingratiation, boardroom behavior, and demographic minority status affect the likelihood of gaining additional board appointments at U.S. companies. Academy Management Journal, in press.


Table 1
Means, Standard Deviations, and Correlations between Predictor Variables and Dependent Variable

<table>
<thead>
<tr>
<th></th>
<th>M</th>
<th>s.d.</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Patient Satisfaction</td>
<td>0.51</td>
<td>0.11</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Practice Busyness</td>
<td>0.66</td>
<td>0.47</td>
<td>-0.30</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Full time status</td>
<td>0.80</td>
<td>0.20</td>
<td>-0.07</td>
<td>0.11</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Number of patients in panel</td>
<td>1749.77</td>
<td>550.63</td>
<td>-0.10</td>
<td>0.26</td>
<td>0.59</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. Panel age</td>
<td>45.84</td>
<td>4.89</td>
<td>0.07</td>
<td>-0.07</td>
<td>0.05</td>
<td>-0.03</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. Chronic sickness of panel</td>
<td>1.04</td>
<td>0.12</td>
<td>0.13</td>
<td>-0.12</td>
<td>-0.15</td>
<td>-0.13</td>
<td>0.55</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7. Tenure with Prohealth (years)</td>
<td>14.81</td>
<td>8.51</td>
<td>0.20</td>
<td>-0.14</td>
<td>0.14</td>
<td>0.08</td>
<td>0.33</td>
<td>-0.20</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8. Age (years)</td>
<td>50.34</td>
<td>6.58</td>
<td>0.09</td>
<td>-0.09</td>
<td>0.16</td>
<td>0.21</td>
<td>0.29</td>
<td>-0.05</td>
<td>0.69</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9. Race (0=white; 1=non-white)</td>
<td>0.12</td>
<td>0.32</td>
<td>-0.15</td>
<td>0.02</td>
<td>0.01</td>
<td>-0.04</td>
<td>-0.14</td>
<td>-0.05</td>
<td>-0.12</td>
<td>-0.03</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10. Sex (0=male, 1=female)</td>
<td>0.38</td>
<td>0.49</td>
<td>-0.06</td>
<td>-0.06</td>
<td>-0.63</td>
<td>-0.44</td>
<td>-0.21</td>
<td>0.04</td>
<td>-0.25</td>
<td>-0.31</td>
<td>0.12</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11. Productivity</td>
<td>23.00</td>
<td>1.97</td>
<td>0.05</td>
<td>0.12</td>
<td>0.22</td>
<td>0.30</td>
<td>-0.06</td>
<td>0.22</td>
<td>-0.25</td>
<td>-0.01</td>
<td>-0.04</td>
<td>-0.15</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>12. Quality</td>
<td>-0.01</td>
<td>1.55</td>
<td>0.11</td>
<td>0.03</td>
<td>0.07</td>
<td>0.08</td>
<td>0.21</td>
<td>0.05</td>
<td>0.14</td>
<td>0.11</td>
<td>0.02</td>
<td>0.04</td>
<td>-0.03</td>
<td>-</td>
</tr>
<tr>
<td>13. Accessibility to Patients</td>
<td>0.16</td>
<td>0.15</td>
<td>0.13</td>
<td>-0.11</td>
<td>-0.18</td>
<td>-0.23</td>
<td>0.05</td>
<td>-0.06</td>
<td>-0.04</td>
<td>-0.16</td>
<td>-0.11</td>
<td>0.21</td>
<td>0.05</td>
<td>0.23</td>
</tr>
</tbody>
</table>

N = 113; All correlations larger than .15 are significant at p<.05
**TABLE 2**
Analysis Examining Moderating Effects of Physician Race, Sex and Objective Performance on Patient Perceptions of Physician Performance Quality

<table>
<thead>
<tr>
<th></th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Constant</strong></td>
<td>.60</td>
<td>.69</td>
<td>.67</td>
</tr>
<tr>
<td><strong>Controls</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Practice Busyness</td>
<td>-.05**</td>
<td>-.04†</td>
<td>-.05*</td>
</tr>
<tr>
<td>Full time status</td>
<td>-.07</td>
<td>-.11</td>
<td>-.13†</td>
</tr>
<tr>
<td>Number of patients in panel</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
</tr>
<tr>
<td>Panel age</td>
<td>-.01†</td>
<td>-.01*</td>
<td>-.01*</td>
</tr>
<tr>
<td>Chronic sickness of panel</td>
<td>.24†</td>
<td>.27*</td>
<td>.22†</td>
</tr>
<tr>
<td>Tenure with Prohealth (years)</td>
<td>.01**</td>
<td>.01**</td>
<td>.01**</td>
</tr>
<tr>
<td>Age (years)</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
</tr>
<tr>
<td><strong>Direct effects</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Race (0=white; 1=non-white)</td>
<td>-.03</td>
<td>-.03</td>
<td>-.05†</td>
</tr>
<tr>
<td>Sex (0=male, 1=female)</td>
<td>-.04†</td>
<td>-.05*</td>
<td>-.04†</td>
</tr>
<tr>
<td>Productivity</td>
<td>.01</td>
<td>.01</td>
<td>.01†</td>
</tr>
<tr>
<td>Quality</td>
<td>.01</td>
<td>.01*</td>
<td>.01†</td>
</tr>
<tr>
<td>Accessibility to Patients</td>
<td>.06</td>
<td>.09</td>
<td>.03</td>
</tr>
<tr>
<td><strong>Interactions</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Productivity X Sex</td>
<td>.00</td>
<td>.00</td>
<td></td>
</tr>
<tr>
<td>Quality X Sex</td>
<td>-.03**</td>
<td>-.03**</td>
<td></td>
</tr>
<tr>
<td>Accessibility X Sex</td>
<td>-.26*</td>
<td>-.27*</td>
<td></td>
</tr>
<tr>
<td>Productivity X Race</td>
<td></td>
<td>-.03**</td>
<td></td>
</tr>
<tr>
<td>Quality X Race</td>
<td>-.04*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Accessibility X Race</td>
<td></td>
<td>-.36</td>
<td></td>
</tr>
<tr>
<td>R²</td>
<td>.222**</td>
<td>.306**</td>
<td>.367*</td>
</tr>
<tr>
<td>ΔR² from previous model</td>
<td>.084**</td>
<td>.061*</td>
<td></td>
</tr>
</tbody>
</table>

---

4 One tailed t-test used for hypothesized coefficients, two-tailed for non-hypothesized coefficients; †p<.10 * p<.05 ** p<.01  N = 113. All participants are medical doctors. The sample consisted of 100 whites, 10 Asian or Pacific Islanders, 2 blacks, and one Native American.
Figure 1. The influence of physician sex and performance of patient-centered behaviors on patient perceptions of physician performance quality

Figure 2. The influence of physician sex and performance of patient-centered behaviors on patient perceptions of physician performance quality

<table>
<thead>
<tr>
<th>**</th>
<th>p &lt; .01</th>
</tr>
</thead>
<tbody>
<tr>
<td>*</td>
<td>p &lt; .05</td>
</tr>
<tr>
<td>+</td>
<td>p &lt; .10</td>
</tr>
<tr>
<td>n.s.</td>
<td>p &gt; .10</td>
</tr>
</tbody>
</table>
Figure 3. The influence of physician race and performance of patient-centered behaviors on patient perceptions of physician performance quality

- Physician Quality
- Percent of patients rating physician as excellent
- White (**) vs. non-white (+)

Figure 4. The influence of physician race and performance of patient-centered behaviors on patient perceptions of physician performance quality

- Physician Productivity
- Percent of patients rating physician as excellent
- White (**) vs. non-white (**)

** p < .01
* p < .05
+ p < .10