All That You Can Be: Stereotyping of Self and Others in a Military Context

Monica Biernat and Christian S. Crandall
University of Kansas

Lissa V. Young
United States Military Academy

Diane Kobrynowicz
The College of New Jersey

Stanley M. Halpin
U.S. Army Research Institute for the Behavioral and Social Sciences

The authors tested the shifting standards model (M. Biernat, M. Manis, & T. E. Nelson, 1991) as it applies to sex- and race-based stereotyping of self and others in the military. U.S. Army officers attending a leadership training course made judgments of their own and their groupmates' leadership competence at 3 time points over a 9-week period. We examined the effects of officer sex and race on both subjective (rating) and objective/common-rule (ranking/Q-sort) evaluations. Stereotyping generally increased with time, and in accordance with the shifting standards model, pro-male judgment bias was more evident in rankings than in ratings, particularly for White targets. Self-judgments were also affected by sex-based shifting standards, particularly in workgroups containing a single ("solo") woman. Differential standard use on the basis of race was less apparent, a finding attributed to the Army's explicit invocation against the use of differential race-based standards.

> A soldier needs physical and moral courage, ingenuity and integrity, determination and loyalty, a sense of humor, and of course luck, to be successful. . . . I do not believe and did not see any evidence that these qualities are distributed on the basis of gender or race. (Major Rhonda Cornum, Army flight surgeon taken prisoner by Iraqi soldiers during Operation Desert Storm. Presidential Commission on the Assignment of Women in the Armed Forces, 1992, p. 111)

> I can tell you, without statistics and without detailed reporting from the field, that American ability to wage war has already been seriously weakened by the deployment of relatively large numbers of women troops to an overseas battlefield. (David Horowitz, President of the Center for the Study of Popular Culture, Presidential Commission on the Assignment of Women in the Armed Forces, 1992, p. 59).

These two quotes illustrate conflicting views about women's ability, relative to men's, to succeed in military contexts. The former view suggests that neither sex nor race has anything to do with "military ability"; the latter reflects the dominant cultural stereotype that women are less competent than men in this setting. In this article, we examined the influence of both sex and racial stereotypes on U.S. Army captains' judgments of their own and each other's leadership competence. The theoretical perspective guiding this work is the *shifting standards* model (Biernat & Manis, 1994; Biernat et al., 1991), which incorporates the idea that stereotypes exert an influence on judgment through their activation of category-specific judgment standards.

The Shifting Standards Model

Judgments and evaluations of others are always made with reference to some standard. This standard may be externally imposed (e.g., does the individual measure up to some explicit

Monica Biernat and Christian S. Crandall, Department of Psychology, University of Kansas; Lissa V. Young, Department of Behavioral Sciences and Leadership, United States Military Academy; Diane Kobrynowicz, Department of Psychology, The College of New Jersey; Stanley M. Halpin, U.S. Army Research Institute for the Behavioral and Social Sciences, Fort Leavenworth, Kansas.

This research was supported in part by National Institute of Mental Health Grant R29MH48844 and by funds from the U.S. Army Research Institute for the Behavioral and Social Sciences. The views, opinions, and findings in this report are those of the authors and should not be construed as official Department of the Army position, policy, or decision, unless so designated by other authorized documents.

We are grateful to Colonel A. E. Bryant, director of the Combined Arms and Services Staff School (CAS3) at Fort Leavenworth, for providing access to the study samples and to Doug Spiegel, Angela Karrasch, and Kathy Fuegen for their assistance with data collection. The results of Study 1 were presented to Colonel A. E. Bryant on May 7, 1996. He subsequently authorized additional studies to verify these findings and to identify changes that could be made within CAS3 to ameliorate stereotyping effects.

Correspondence concerning this article should be sent to Monica Biernat, Department of Psychology, 426 Fraser Hall, University of Kansas, Lawrence, Kansas 66045. Electronic mail may be sent to biernat@ukans.edu.

Journal of Personality and Social Psychology, 1998, Vol. 75, No. 2, 301–317

0022-3514/98/$3.00

performance criterion?), but in many cases, it is likely to be at least partially determined by the group membership of the person being evaluated. This is the argument we have advanced in an approach to understanding stereotype-based judgment called the shifting standards model (Biernat, 1995; Biernat & Kobrynowicz, 1997; Biernat & Manis, 1994; Biernat et al., 1991; Biernat, Vescio, & Manis, 1998; Kobrynowicz & Biernat, 1997).

Specifically, this model suggests that when perceivers judge individual members of stereotyped groups on stereotype-relevant dimensions, they use within-category judgment standards. For example, given stereotypes that men are better leaders than women (Brown & Geis, 1984; Butler & Geis, 1990; Eagly, Karau, & Makhijani, 1995; Eagly, Makhijani, & Klonsky, 1992; Heilman & Kram, 1983; Malloy & Janowski, 1992), they are likely to judge the leadership competence of a particular woman relative to (lower) standards of competence *for women* and the leadership competence of a particular man relative to (higher) standards of competence *for men.* The result is that evaluations of men and women on leadership competence may not be directly comparable, as their meaning is tied to different contexts: "Good" for a woman does not mean the same thing as "good" for a man (see also Kobrynowicz & Biernat, 1997).

A standard incorporates the average and range that is expected from members of a group on a particular dimension and aids the judge in anchoring the endpoints of a subjective rating scale (e.g., high to low competence). Rating points are defined to reflect the expected distribution of category members on the dimension, with high numbers reserved for targets with the highest expected level of the attribute among members of the category. When groups are expected to differ (i.e., when a stereotype is held), these endpoints are differentially anchored for the contrasting groups (see variants of this theme in classic judgment models by Parducci, 1963, 1965; Postman & Miller, 1945; Upshaw, 1962, 1969; Volkmann, 1951).

Evidence supporting the operation of stereotype-based standard shifts can be gleaned from comparisons between judgments that are made on such subjective rating scales ("slippery" scales whose units can be differentially defined and adjusted) to those made on objective rating scales (externally anchored, "common-rule" scales whose judgment units maintain a constant meaning across contexts and types of targets; see Biernat, 1995). The key prediction of the shifting standards model is that objective judgments are more likely than subjective judgments to reveal the influence of stereotypes; because subjective scales can be differentially adjusted for different target categories, they may mask this influence. Thus, when perceivers make height judgments of male and female targets, men are decisively judged taller than women in inches (an objective, common-rule scale), but this sex-differential is significantly reduced when the subjective labels "short" and "tall" are applied (Biernat et al., 1991). That is, a man and a woman may be perceived quite differently in objective height (e.g., 6 feet 2 in. [1.88 m] if a man, 5 feet 10 in. [1.78 m] if a woman), but both be labeled "tall."

To date, the signature shifting standards pattern (stronger stereotyping effects on common-rule than subjective response scales) has been documented in a variety of judgment domains and for both sex and racial groups. Specifically, we have found that judges shift their standards in ratings of women versus men on the physical dimensions of height and weight and on the social dimensions of income, verbal ability, writing competence, aggression, parenting involvement, and job-related competence; standards for Blacks versus Whites similarly shift in the social domains of verbal ability, athleticism, and job-related competence (Biernat & Kobrynowicz, 1997; Biernat & Manis, 1994; Biernat et al., 1991; Kobrynowicz & Biernat, 1997).

Several important questions about the model remain. First, does extended acquaintance with individual group members increase or decrease the use of within-category judgment standards? Second, are within-category standards also applied when perceivers make self-judgments? Third, does judgment bias always take the form of standard shifts; more specifically, do social categories such as sex and race function similarly (i.e., do judgment standards shift on the basis of both sex and race?), or do they exert different influences on judgment patterns? And finally, to what extent do contextual factors—for example, the number of women relative to men present in a judgment setting—influence the extent to which shifting standards are applied? The present research represents an attempt to address these questions, as we moved from the controlled laboratory and the use of undergraduate psychology student participants to a military setting—a U.S. Army training facility—where male and female officers representing a variety of ethnic groups judged themselves and each other on the dimension of leadership competence during a 9-week training course. We viewed the Army setting as a means of theory testing, but this context had the added advantage of allowing examination of one additional research question: Can the laboratory-based findings regarding shifting standards be documented in a naturalistic setting, where meaningful judgments are made of live, interacting targets (see Sears, 1986)?

Stereotyping Over Time

The present studies incorporated a longitudinal design and thus allowed us to examine shifting standards patterns as they develop, change, or persist over time. How might increased acquaintance affect patterns of stereotyping and standard use? The broader literature on intergroup relations offers some answers. The contact hypothesis suggests that with increasing (positive) intergroup contact, prejudice and stereotyping should decrease (Allport, 1954; Amir, 1969, 1976; Cook, 1978). Positive contact includes conditions characterized by "equal status, stereotype disconfirmation, cooperation, high acquaintance potential, and equalitarian norms" (Hewstone, 1996, p. 327).

To the extent that interaction among the Army captains in our studies meets many or all of the above criteria, stereotyping on the basis of sex or race should decrease over time. A similar prediction can be derived from models of impression formation such as Fiske and Neuberg's (1990) continuum model and Brewer's (1988) dual process model. These models suggest that stereotypes will be relied on less and individuating information relied on more when motivation to individuate is high or when targets of perception do not fit relevant stereotypes. As participants get to know each other over time and presumably discover that their peers do not neatly comply with stereotyped expectations, they should reject the use of these stereotypes as they make judgments of leadership competence. From the shifting standards perspective, decreased use of stereotypes precludes

the operation of within-category standards to define the meaning of subjective judgment scales (see Biernat et al., 1991). Thus, along with an overall reduction of stereotyping over time, differences in judgment based on response scale (objective vs. subjective) should be reduced as well.

However, the competitive nature of military training may not provide optimal contact conditions, and there are also theoretical reasons to expect that reliance on stereotypes may increase, not decrease, with time and exposure. For example, Darley and Gross (1983) have argued that stereotype-based expectancies function as hypotheses about a target person and that perceivers therefore require some data (behavioral evidence) before they are willing to use their stereotypes to render judgment. If this is the case, it may be that on first meeting individual members of stereotyped groups (i.e., women and ethnic minorities), raters are unwilling to use these social categories as a basis of judgment. This inhibition may be driven by social desirability or "political correctness" norms, by the desire to protect an egalitarian self-image (see Gaertner & Dovidio, 1986), or by epistemic concerns (Leyens, Yzerbyt, & Schadron, 1994; Yzerbyt, Schadron, Leyens, & Rocher, 1994). In any case, if stereotypes are avoided in judgment, differential response scale effects should not be observed. However, with time and exposure to individuating information, the inhibition on use of stereotypes may be withdrawn and perceivers may read the behavioral evidence they collect in a stereotype-confirming manner (Darley & Gross, 1983; Hilton & Von Hippel, 1996; Snyder & Cantor, 1979; Snyder & Swann, 1978; Stangor & Lange, 1994; Von Hippel, Sekaquaptewa, & Vargas, 1995). Over time, then, perceivers may be increasingly likely to use their stereotypes as a basis of judgment, and this should be most strongly evidenced on objective judgment scales.

In summary, the shifting standards model posits that if use of sex or race as a judgment cue decreases with time (as might be predicted from the contact hypothesis), the signature shifting standards pattern should dissipate as well; if use of sex stereotypes increases with time (as might be predicted by extensions of the Darley & Gross, 1983, model), this should be particularly marked on objective, relative to subjective, judgment scales.

Self-Judgments

The shifting standards model has thus far focused on how within-category standards are used to judge others, but whether individuals similarly judge themselves relative to their in-group standards (and therefore show different self-judgments on subjective vs. objective response scales) has not been tested. The extensive literature on social comparison theory suggests that we do compare ourselves (i.e., assess our opinions and abilities) to similar others (Festinger, 1954; see also Goethals & Darley, 1977; Halpin, 1970; Wood, 1989), and in a considerable amount of research from this perspective, *similarity* is based on social category memberships such as sex (e.g., Buunk & VanYperen, 1991; Major, 1989, 1993; Major & Forcey, 1985; Major & Testa, 1989; Zanna, Goethals, & Hill, 1975).

There is also substantial evidence that as a consequence of self-categorization processes, individuals engage in *self-stereotyping*—ascribing to themselves the attributes of their groups (e.g., Biernat, Vescio, & Green, 1996; Hardie & McMurray, 1992; Haslam, Oakes, Turner, & McGarty, 1996; Hogg & Turner, 1987; Lau, 1989; Lorenzi-Cioldi, 1991; Simon, Glässner-Bayerl, & Stratenwerth, 1991; Simon & Hamilton, 1994; Turner, 1982; Turner, Hogg, Oakes, Reicher, & Wetherell, 1987). To the extent, then, that the self is categorized as a member of a stereotyped group and that comparison of self to in-group others occurs, the processes previously outlined for stereotype-based judgment of others may also apply to the self. That is, the use of within-category judgment standards may lead to reductions of self-stereotyping effects on subjective compared with objective judgment scales: If women judge themselves relative to women and men relative to men, subjective judgment scales will mask female–male differentials in self-evaluation. Such a pattern would link the shifting-standards model with the literatures on social comparison, social identity, and self-stereotyping by demonstrating that within-group comparison processes allow individuals to shift or adjust the meaning of subjective evaluative dimensions for judgments of both others and themselves.

Comparing Sex and Race

The military provides an ideal context in which to examine stereotyping effects. Widely publicized concerns about sexual harassment in military contexts, women's admission to formerly male-only military academies, sex-segregated basic training, double standards in the treatment of sexually active military men and women, and the role of female soldiers in combat attest to the relevance of sex and sex stereotyping in this setting (Francke, 1997; Martindale, 1990; Presidential Commission on the Assignment of Women in the Armed Forces, 1992; Pryor, 1988; U.S. Department of Defense, 1988). Furthermore, in their recent meta-analysis on sex and perceived leader effectiveness, Eagly et al. (1995) reported a significantly larger effect size (d = .42) for studies done in military settings compared with those done in other organizational contexts (ds ranged from −.15 to .07): In the military, but generally not in other settings, men fared better than women in perceived leadership effectiveness. Problems with racial stereotyping and bias have also been at issue at various points in military history (see Smither & Houston, 1991; St. Pierre, 1991; C. Young, in Terkel, 1980), though some have argued that this institution represents one of the great success stories on issues of racial discrimination since the passage of the 1948 Executive Order that integrated the armed services (Moskos & Butler, 1996; Pulakos, White, Oppler, & Borman, 1989).

Although stereotypes based on both sex and race may affect judgments of U.S. Army officers, there is also reason to suspect that sex may be the more salient categorical distinction in military contexts and that the Army setting may be more likely to instantiate differential standards based on sex than on race. Women are less well represented in the Army (and in the present studies) than are racial minorities (the total active duty Army is roughly 45% non-White and 14% female; see Defense Equal Opportunity Management Institute, 1995), and women's potential roles in the military are actively limited in a way that is not true for racial minorities (e.g., combat exclusions apply to women). Furthermore, leadership competence—the judgment dimension on which we focused in the present research—is a marked component of sex stereotypes (Bem, 1974; Eagly et al.,

1992, 1995; Spence, Helmreich, & Stapp, 1974) but appears in no list of common racial stereotypes (e.g., see Devine & Elliot, 1995; Niemann, Jennings, Rozelle, Baxter, & Sullivan, 1994). Thus, we may be more likely to find evidence of sex- than race-based judgment bias in this context.

Differences based on sex versus race may be even more apparent when one considers the issue of standards in the military. Explicit in Army code, for example, is the fact that the adjustment of standards on the basis of race does not occur. In their book detailing this facet of the Army, Moskos and Butler (1996) wrote: "The Army does not lower its standards; it elevates its recruits and soldiers" (p. 74), and "the Army does not patronize or infantilize Blacks by implying that they need special standards in order to succeed" (p. 72). Even more relevant to the shifting standards model, "the military has no hint of two promotion lists in which Whites are compared only with Whites, Blacks only with Blacks" (p. 70). This stands in contrast with recommendations for military policy on the issue of sex set by the Presidential Commission on the Assignment of Women in the Armed Forces (1992): "The Services should retain gender-specific physical fitness tests and standards" (p. 5), "Entry level training may be gender-specific as necessary" (p. 9), "Military pre-commissioning training may be gender-normed" (p. 11), and "women should be excluded from direct land combat units and positions" (p. 24; see also Francke, 1997).

If military personnel attend to these policies and procedures, and if, therefore, judgment standards are adjusted on the basis of sex but not race, the implications for the shifting standards model in this context are clear: We should find evidence of standard shifts (i.e., stronger evidence of stereotyping on common-rule than subjective scales) only when sex, but not race, is the relevant category cue. A lack of judgment scale differences does not necessarily imply, however, that all signs of bias based on race will be absent. For example, if negative stereotypes are applied to minorities, but perceivers avoid shifting their standards, a race bias, but no differences across judgment scales, should emerge. Such a finding would delimit the shifting standards model by indicating that stereotyping need not always prompt standard shifts; contextual factors that discourage differential standard use may override this tendency. More generally, differential findings for sex and race could indicate that stereotypes are a necessary, but not sufficient, precursor to shifting standards effects.

Context and Category Salience

In addition to norms and policies, one aspect of a group setting that may increase the salience of a social category is the number of category members present. Specifically, research suggests that an individual who is the sole representative of his or her group (a "solo") draws increased attention; this attention, in turn, leads perceivers to judge the solo more extremely—often, more stereotypically—than they otherwise would (Biernat & Vescio, 1993; Taylor & Fiske, 1978; Taylor, Fiske, Etcoff, & Ruderman, 1978). Furthermore, solos themselves tend to experience increased self-focus and encounter problems such as unrealistic expectations, uninformative feedback, and social isolation (Kanter, 1977; Pettigrew & Martin, 1987). Solo status therefore has implications for both the impressions perceivers form and for self-perception (i.e., for stereotyping of others as well as the self).

As already indicated, women are a distinct minority in the broad context of the military as well as in the studies described here. Furthermore, some of the groups we examined in our research included a solo woman, whereas others included 2 women, providing an opportunity to examine the impact of gender status (1 vs. 2 women present) on judgments of own and others' leadership competence. A unique prediction from the shifting standards model is that stereotyping effects will take a form that follows from the increased use of within-sex judgment standards in groups containing only one woman. Specifically, if solo status draws perceivers' attention to the solo and draws the solo's attention to herself, the use of sex as a cue to judgment will likely increase. On objective or common-rule rating scales, these sex effects will be clearly revealed: Men will be judged and judge themselves as more competent than solo women. However, if sex stereotypes have their impact through the activation of within-category judgment standards, subjective judgments will mask these stereotyping effects: Solo women will be particularly likely to judge themselves and be judged relative to women, and men in these solo-woman groups will have a heightened tendency to judge themselves and be judged relative to men. The result of these processes is little or no effect of sex stereotypes on subjective ratings. In short, the shifting standards pattern—stronger stereotyping on common-rule than on subjective scales—will be intensified in groups containing a solo woman.

To summarize, the present studies were designed to extend previous theory and research on the shifting standards model in several ways: (a) by examining longitudinal patterns of stereotyping on objective and subjective judgment scales; (b) by assessing whether self-judgments, like other-judgments, are affected by category-specific standard use; (c) through testing the scope and limits of the shifting standards model by examining whether sex and race categories have differential effects on judgment in a context that actively sanctions standard shifts in one case (sex) but not in the other (race); and (d) by considering the effects of a group contextual factor—the number of women present—on patterns of standard shifts in both self- and other-judgments.

Study 1

Method

Sample

Participants were 100 students at the Combined Arms and Services Staff School (CAS3) at Fort Leavenworth, Kansas. All were U.S. Army Commissioned Officers at the rank of captain who represented each of the three general branches of service (Combat Arms, Combat Support, and Combat Service Support). The Combat Arms specialty includes aviators, infantry personnel, and special forces personnel (this is often perceived as the most prestigious of the branches); the Combat Support branch includes military police, military intelligence, engineers, and chemical corps; and the Combat Service Support staff include finance, ordnance, transportation, and quartermaster corps. Table 1 presents the complete breakdown of the sample by the categories of sex, race, and branch of service.

Table 1
Sample Frequencies by Branch of Service, Sex, and Race, Study 1

	Combat arm and sex						
	Combat Arms		Combat Support		Combat Service Support		
Race	Men	Women	Men	Women	Men	Women	Total
White	31	2	31	4	10	1	79
Black	3	0	6	1	2	0	12
Asian	3	0	0	2	0	0	5
Hispanic	1	0	0	0	1	1	3
Native American	0	0	0	0	1	0	1
Total	38	2	37	7	14	2	100

Course and Population Description

The officers participated in the present research during the course of their 9-week training at Fort Leavenworth. The CAS3 program provides training in advanced tactical decision making and division level staff skills and is a requirement for promotion to major. The goals of the course are to "improve ability to analyze and solve military problems, improve communications skills, and improve ability to interact and coordinate as a member of a staff" (CAS3 Office, 1997). Each session of CAS3 involves roughly 500 students who are divided into groups of 12 or 13; we received permission to study 8 of these groups (100 students). The policy is to establish groups such that proportionate distribution based on sex, race, and branch of service is obtained: Each group of 12–13 officers typically includes at least 1 woman, 2–3 officers of minority ethnic origin, and representatives from a variety of service branches. The students live, eat, work, and conduct physical training together for 12–14 hr a day. Of the eight groups we studied, five included solo women, and three included 2 women.

The 9-week course centers around a series of individual and group tasks (e.g., war games, decision making and planning, and oral communication exercises). As part of the regular course curriculum, each exercise performance by an individual or the group is evaluated on nine leadership competencies: communications, teaching and counseling, soldier team development, technical and tactical proficiency, supervision, decision making, planning, use of available systems, and professional ethics. These evaluations are made by section leaders and are treated as confidential communications that do not appear on officers' permanent records.

Data Collection Procedures

At three time points during the 9-week course, participants were asked to both rate and rank their groupmates and themselves with regard to their overall effectiveness as "leaders/commanders." The order in which ratings and rankings were made was counterbalanced, and this variable had no effect on the findings reported below. Within each group, officers were given an alphabetical list of group members' names, alongside which they made their judgments. The rating questionnaire required officers to judge the leadership competence of their groupmates and themselves on 5-point scales. Scale points were labeled *outstanding, excellent, satisfactory, needs improvement,* and *needs much improvement;* this is the same rating system that Army personnel use to evaluate students' progress through the course. The ranking questionnaire included the same alphabetical list and required officers to rank order each member of their group (including themselves) with regard to leadership competence. We have argued elsewhere that rank orders meet our criteria of objectivity in the sense that they invite the use of a single dimension on which to evaluate all individuals in a given context (in this case, one's small group; see Biernat & Manis, 1994). This imposition of a single judgment array stands in contrast to the multiple and shifting meanings that are possible when subjective ratings are made.

The first data collection took place at zero acquaintance (Albright, Kenny, & Malloy, 1988) on Day 1, in the first minutes of the course. At this point, students had not yet introduced themselves to each other, but each was dressed in full uniform and seated behind a name plate. Time 2 data collection took place at the end of Week 3 of the course. This point marked a transition in curriculum from an emphasis on intensive individual work to group work. Therefore, Time 2 judgments were made after students had lived, studied, and recreated together for a considerable time, but before they had explicitly worked together as a unit on group decision making and tactical training projects. The final data collection took place at the end of Week 8, after the completion of a 5-week period of highly intensive group work.

Additional demographic and background information was also collected at Time 1. On average, participants were 31 years old (range = 27–47) and had been in commissioned service for 8 years (range = 4–14). Officers were also provided with a checklist of possible military honors (badges and medals) and asked to indicate those they had personally been awarded. Badges and medals were weighted by prestige to create the variables medals and badges, described below.

Results

Overview

We treated the target of judgment (rather than the judger) as the unit of analysis. Thus, the dependent variables of interest were (a) the mean leadership ranking and rating the target received from his or her groupmates (excluding the self-judgment) and (b) the target's self-rating and ranking, at each time point (corrected for number of group members, 12 or 13).[1] Judgments were reverse scored such that high numbers indicated more favorable evaluations. We first report preliminary analyses that consider each captain's individual achievements (medals and badges) and the relationship between these awards and the judgments received. Next, we turn our attention to the effects of the two different social categories on evaluations: target sex and race (White vs. non-White; more specific racial breakdowns were prohibited by the low sample sizes). Because of the distribution of our sample across these categories as well as across branch of service (see Table 1), we could not examine the category joint effects; instead we report separate Category × Rating Scale (rating vs. ranking) × Time analyses of variance (ANOVAs), focusing on sex and race, in turn. The same analyses are repeated for self-ratings and self-rankings.

Medals and Badges

Participants provided three types of information that we considered indicative of their general past achievement: The number and types of medals and badges they had earned in their Army

[1] The estimates of judgments received by one's groupmates were highly reliable at Times 2 and 3 (for Time 2 and 3 rankings, average Cronbach's αs = .87 and .93, respectively; for ratings, .82 and .91), though Time 1 interjudge agreement was lower (mean alphas for rankings and ratings, respectively, were .66 and .44).

careers and whether they had a combat patch, signifying deployment to a combat zone. These variables were significantly correlated with the mean judgments captains received from their groupmates at each of the three time points. The average correlations were .20 between number of medals and evaluation received, .38 between number of badges and evaluation received, and .25 between combat deployment (no or yes) and evaluation received, $ns = 100$, $ps < .05$. That is, targets were judged more favorably the more honors they had received. The relationship between achievement and evaluation also remained stable across time and across judgment type (rating vs. ranking), though the correlations with rankings tended to be stronger.

Analyses also indicated that male captains tended to have slightly higher achievement ($Ms = 9.94$ badges, 13.39 medals, and 42% having undergone combat deployment) than female captains (comparable $Ms = 4.45$, 9.64, and 18%), though these differences were not reliable ($ps < .14$). White captains ($M = 10.67$) had significantly more badges than non-White captains ($M = 4.33$), $t(98) = 2.72$, $p < .01$, and nonsignificantly more medals ($M = 13.58$) and combat deployment (43%) than non-White captains ($Ms = 10.71$ and 24%, $ps < .14$).[2]

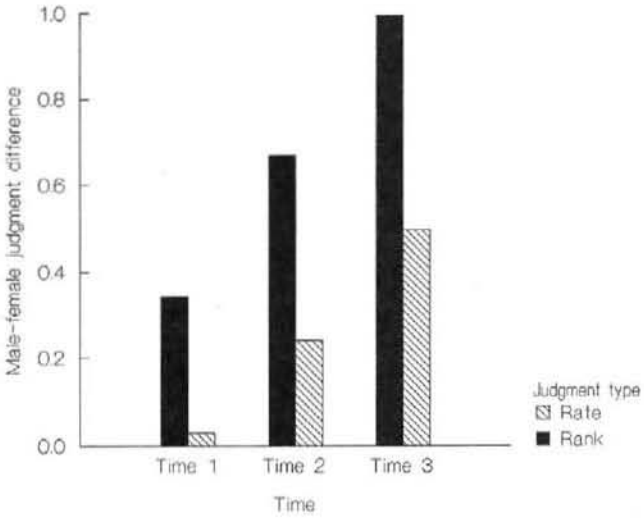

Figure 1. Sex differential in leadership judgments of groupmates by response scale and time point, Study 1.

Social Category-Based Ratings and Rankings Over Time

Given the findings described above, we thought it was important to control for past awards in our analyses examining social category effects on evaluation. The use of controls for these and other factors described below provides some assurance that any observed effects are stereotyping effects, rather than perceptions based on the arguably diagnostic cues of past achievement (however much these may have been influenced by the judgmental biases of other Army personnel). We therefore took the following steps before computing the critical Category × Time × Judgment Scale ANOVAs. First, to make it possible to directly compare rankings and ratings, we standardized judgments within scale type (rating and ranking) and across time points. These standardized scores were then regressed on the following control factors: Medals, badges, combat deployment, number of years of commissioned service, CAS3 group size (12 or 13), and branch of service (a 3-level variable). This latter factor was included because we found that individuals from the high-status service branch (Combat Arms) received more favorable evaluations than those from the lower status branches at each point in time (mean r between branch and evaluations = .31). Finally, in addition to the six control factors, we also regressed judgments on target race when we wished to focus on sex effects and on target sex when we wished to focus on race effects. Thus, in the Sex × Time × Judgment Scale mixed-design ANOVA reported below, the dependent measures were the residuals that remained after controlling for race, branch, medals, badges, combat deployment, number of years of commissioned service, and group size; similarly, the Race × Time × Scale ANOVA was based on the residuals that remained after controlling for sex, branch, medals, badges, combat deployment, years of commissioned service, and group size.[3]

Target sex. Figure 1 depicts the male–female sex differential for rankings and ratings at each of the three time points. All the differences were positive, indicating that male captains were judged more favorably (as better leaders) than female captains. In addition, the sex differential was greater on leadership rankings than ratings, and the judgment differential appeared to increase with time. The ANOVA revealed two significant effects: a main effect of target sex, $F(1, 98) = 3.92$, $p = .05$, and the predicted Sex × Scale interaction, $F(1, 98) = 4.26$, $p < .05$. However, neither the Sex × Time nor the Sex × Time × Scale interaction was significant ($ps > .22$).

Overall, in standardized units, women ($M = -.41$) were judged less favorably than men ($M = .05$). Furthermore, this difference was reliably larger for rankings ($Ms = -.60$ and .07 for women and men, respectively) than for ratings (comparable $Ms = -.23$ and .03). This latter pattern is the trademark shifting standards effect: Stereotyping effects were stronger on objective (ranking) than subjective (rating) judgment scales. Analyses within time point and scale type also revealed that the male–female ranking differential was marginally significant at Time 1 ($p < .12$) and reliable at both Times 2 and 3 ($ps < .05$); however, the male–female rating differential was not reliable at any of the time points ($Fs < 1$). Finally, separate Sex × Time analyses on rankings and ratings revealed that for rankings, the main effect of sex was reliable, $F(1, 98) = 7.81$, $p < .01$, and the Sex × Time interaction approached significance, $F(2, 196) = 2.35$, $p < .10$. For ratings, neither effect was significant ($Fs <$

[2] See Francke (1997) for a discussion of how gender bias may enter into the awarding of medals and badges. To the extent that bias based on either sex or race occurs, these awards cannot be considered pure indicators of competence or merit. Nonetheless, they are visible indicators of past (acknowledged) achievement, which may provide some legitimate basis for competence judgments.

[3] These control factors accounted for an average of 21% of the variance in target judgments. In analyses without these controls, the effects of social category membership (sex and race) either remained the same or increased in strength. In general, the analyses with controls offer a more conservative test of stereotyping effects; where effects emerge, we can be more confident of them.

1). In summary, although it appears that sex-based stereotyping effects increased with time, statistical support for this was limited to the ranking conditions and, even in this case, was not strong. Overall, however, the data support the shifting standards pattern of stronger evidence of sex bias on leadership rankings than on ratings.

In a follow-up analysis, we examined whether the number of women present in a group (1 or 2) moderated the above effects. It did not—in a Sex × Time × Scale × Number of Women ANOVA, all Fs involving this factor were less than 1.

Target race. A comparable Race × Time × Scale ANOVA on judgment residuals revealed a marginal main effect of race, $F(1, 98) = 3.49, p < .07$, and a Race × Time interaction, $F(2, 196) = 5.11, p < .01$. Non-White captains ($M = -.27$) were evaluated more negatively than White captains ($M = .07$), and the pro-White bias increased over time (White/non-White differential was −.12 at Time 1, .52 at Time 2, and .60 at Time 3). Neither the Race × Scale nor the Race × Scale × Time interaction was significant (ps > .30).[4]

Because each CAS3 group included at least 2 non-White officers, each individual officer was evaluated by both White and non-White groupmates. It was therefore possible to examine whether race of judge, in addition to race of target, affected patterns of evaluation.[5] For each target person, we calculated the mean judgment received from White groupmates and the mean judgment received from non-White groupmates. These means were standardized within scale type and residualized as described earlier, then submitted to a Target Race × Judge Race × Time × Scale ANOVA. Significant effects of judge race, $F(1, 98) = 7.08, p < .01$; the Target Race × Judge Race interaction, $F(1, 98) = 21.06, p < .0001$; and the Judge Race × Time interaction, $F(2, 196) = 4.33, p < .05$, were subsumed by the significant Target Race × Judge Race × Time interaction, $F(2, 196) = 12.88, p < .0001$ (no other effects were significant). This three-way interaction is depicted in Figure 2; judgments by White raters and non-White raters are shown in separate panels.

As can be seen in this figure, White and non-White judges showed markedly different judgment patterns. Although both groups showed no race bias in judgments at Time 1 (White/non-White Time 1 differences were nonsignificant; ps > .10), each group generally demonstrated in-group bias (more favorable evaluations of own race than other race)[6] at Times 2 and 3. However, simple effects tests indicated that only the race differences at Times 2 and 3 for White judges were reliable. To better interpret the interaction, we computed separate Target Race × Time × Scale ANOVAs for White and non-White judges. Among White judges, significant effects were obtained for target race, $F(1, 98) = 7.22, p < .01$; time, $F(2, 196) = 3.99, p < .05$; and the Target Race × Time interaction, $F(2, 196) = 11.85, p < .0001$. Among non-White judges, however, no effects were reliable (all ps > .20). Separate Target Race × Judge Race × Scale ANOVAs within each time point also indicated that the Target Race × Judge Race interaction was not significant at Time 1 ($F < 1$) but was reliable at both Times 2 and 3, Fs(1, 98) = 21.34 and 26.76, respectively, ps < .0001. Thus, this overall pattern of results supports three conclusions: (a) White evaluators demonstrated stereotypical (pro-White) judgment bias at Times 2 and 3; (b) non-White evaluators demonstrated some (nonreliable) tendency toward bias favoring their minority groupmates, also at Times 2 and 3; and (c) neither group showed evidence of using race-based shifting standards—in no case did type of response scale moderate judgments.

Self-Rankings and Ratings

Medals and badges. Although awards were significantly related to judgments received, we found little evidence that past achievements (medals, badges, combat deployment) affected self-evaluations. When these variables were correlated with each of the six self-judgments (self rating and ranking at each of three time points), rs ranged from −.14 to .21, with a mean of .07.

Sex effects. To be consistent with our analyses of other-judgments, we standardized self-judgments within scale type and across time and regressed these judgments on the set of control variables described earlier (including race). The Sex × Time × Judgment Scale ANOVA on the residuals indicated a main effect of sex, $F(1, 87) = 5.26, p < .05$, and a significant three-way interaction, $F(2, 174) = 7.26, p < .0001$. However, this interaction was further clarified by including the group context factor in our analysis—the number of women who were present in a given group (1 or 2). Of course, conducting this analysis meant that we had to divide the already small number of women (11) into even smaller groups of 5 solos and 6 non-solos. With the appropriate caveats prompted by these small samples, the reanalysis nonetheless documented a reliable Sex × Time × Response Scale × Number of Women (1 vs. 2) interaction, $F(2, 170) = 3.68, p < .05$.

Figure 3 depicts this interaction as separate three-way interactions for solo and non-solo groups. Looking only at groups containing solo women (top panel of graph), we found that the three-way ANOVA revealed a main effect of sex, $F(1, 51) = 7.55, p < .01$, and a Sex × Scale interaction, $F(1, 51) = 5.57$,

---

[4] Although we grouped all non-White targets into a single racial category, there was some variability in the judgments specific minority group members received. The mean standardized but otherwise unadjusted judgments (across time and judgment scale) received by Whites, Asians, Hispanics, Blacks, and the single Native American target, respectively, were .15, −.05, −.66, −.74, and −.72. In general, the Asian targets were judged more similar to White targets than to the other minority groups. When we deleted judgments of the 5 Asian officers from the analysis reported in the text, the only change was that the main effect of race was reliable, rather than marginal, in the reduced data set, $F(1, 93) = 4.74, p < .05$. In short, in this and other analyses, we found no reason to believe that our gross White–non-White distinction disguised any meaningful effects.

[5] The analogous analysis was not possible with regard to sex, as 5 of the 11 women in the sample were only evaluated by men (i.e., they were solo women in their groups), and the other 6 women were evaluated by only 1 woman and by 10 or 11 men.

[6] Because the group of non-White officers includes Blacks, Hispanics, Asians, and Native Americans, the "in-group bias" label is not technically correct—for example, a single Black target may have been judged by another Black officer (in-group member) as well as an Asian or Hispanic officer (other non-Whites, but not in-group members). Nonetheless, we use the label to refer to the more global categories of White versus non-White.

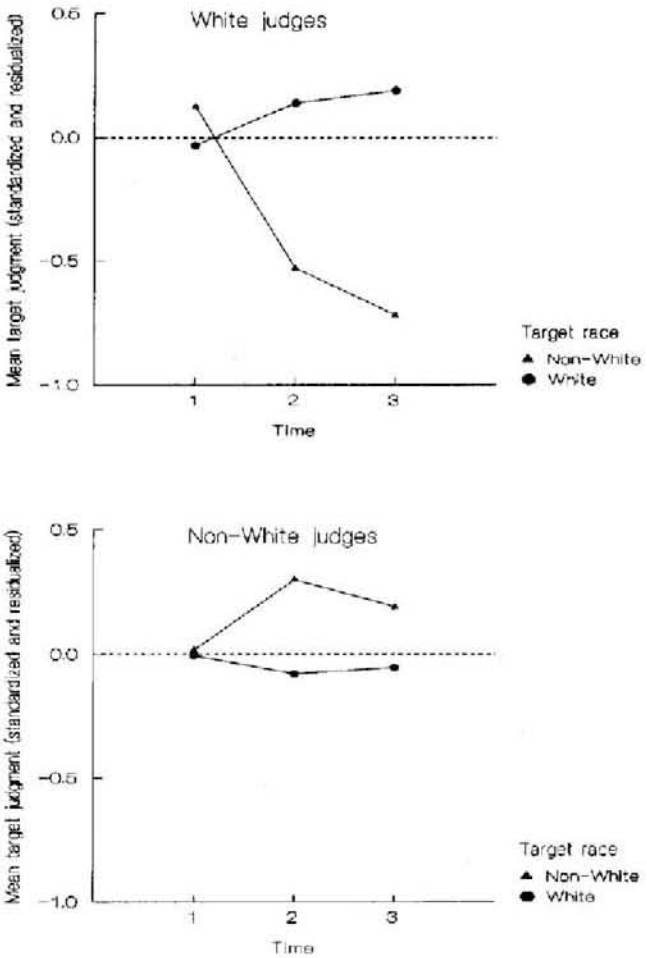

Figure 2. Judge Race × Target Race × Time interaction on judgments of groupmates, Study 1.

$p < .01$. For these groups, the shifting standards pattern—greater sex differentiation on ranks than on ratings—was apparent at all three time points. However, for groups containing 2 women (bottom panel of Figure 3), the main effect of sex was nonsignificant ($F < 1$), but the three-way interaction was reliable, $F(2, 68) = 9.59$, $p < .001$. Here, the shifting standards pattern appeared only at Time 1 ($p < .05$); at Time 2, the rank–rate difference was reliable but in the opposite direction predicted by the model ($p < .01$), and no sex effects were apparent at Time 3. Although we have no explanation for this Time 2 effect, one general conclusion to be drawn from the self-judgments presented in Figure 3 is that a pattern of shifting standards was clearly evident, at all time points, in groups with solo women, but only appeared at Time 1 for groups that included 2 women.

Race effects. The comparable analysis of self-judgments by race revealed no significant effects (all $ps > .25$).

Relationship Between Self- and Other Judgments

Table 2 presents the correlations between (unadjusted) self- and other-judgments by time point, scale type, and group membership. The sets of correlations are not independent, as the men and women groupings include both Whites and non-Whites, and the White and non-White groupings include both men and women (see Table 1). At Time 1, there was no relationship between self-ranks and ranks assigned by others, and for all groups except women, self- and other-ratings were negatively related. By Time 2 and continuing to Time 3, however, self- and other-judgments were positively correlated for both rankings and ratings, but only for the high-status groups (men and Whites). For women and for non-Whites, there was virtually no relationship (in some cases, a slight negative relationship) between self- and other-judgments.

Discussion

The judgment data from this study provided clear evidence of sex-based shifting standards in evaluations of others' leadership competence. Men were consistently judged to be better leaders than women, but this effect was reliable only for rankings and not for ratings. This is the central shifting standards pattern:

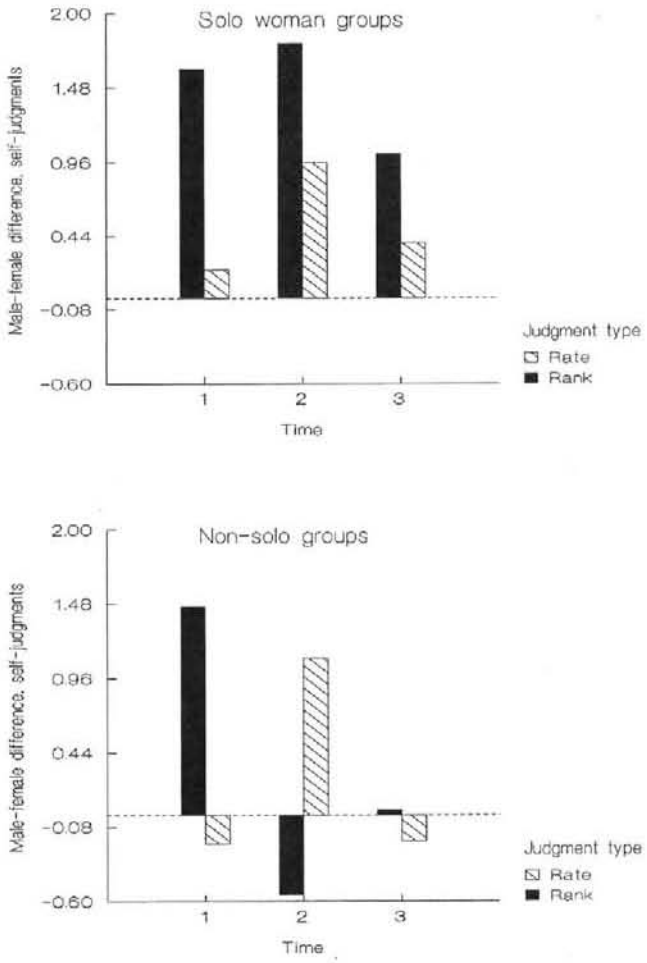

Figure 3. Sex differential in leadership self-judgments by response scale and time point, separately for groups with 1 versus 2 women, Study 1.

Table 2
Zero-Order Correlations Between Self-Judgments and Judgments by Others

Judgment type	Time 1	Time 2	Time 3
	By sex		
Ranks			
Men	.03	.25*	.31**
Women	.20	−.16	.09
Ratings			
Men	−.35**	.15	.40***
Women	.24	−.19	−.09
	By race		
Ranks			
White	.13	.34**	.38***
Non-White	−.03	.10	.15
Ratings			
White	−.19†	.23*	.45***
Non-White	−.53*	.03	−.30

† $p < .10$ (marginally significant). * $p < .05$. ** $p < .01$. *** $p < .001$.

Objective rankings revealed evidence of stereotypes, and subjective ratings masked these effects. We believe this occurred because Army captains judged women relative to women and men relative to men, but rankings forced them to array their groupmates on a single judgment continuum. Although the interaction between target sex, judgment scale, and time was not significant, it was also the case that rankings, but not ratings, produced some evidence of increased stereotyping with time, as participants evolved from strangers to familiars.

This general pattern of increased stereotyping with time is consistent with Darley and Gross's (1983) stereotypes as hypotheses model. In general, sex and race were not used as a basis for judgment at Time 1; neither the Time 1 ratings nor the Time 1 rankings revealed reliable sex or race effects. Participants may have felt that they needed to see some performance evidence before assuming that women and non-Whites would be less competent than men and Whites. By Time 2, after 3 weeks of interactive contact, they apparently had seen enough evidence to confirm their stereotype-based hypotheses.

What evidence do we have that these effects represent stereotypic biases rather than an accurate assessment of relative performance? First, stereotyping effects held even after controlling for a variety of factors that might conceivably be associated with actual performance—medals, badges, combat deployment, and branch/specialty. Controlling for these factors should "level the playing field" such that what remains are "pure" category effects (i.e., bias).[7] Second, there was little agreement between women's self-judgments and the judgments they received from others, or between self- and other-judgments for non-Whites (see Table 2). Self–other agreement is one accuracy criterion (see Funder, 1995; Judd & Park, 1993), and it was generally not met here. Overall, then, we suggest that our data reflected bias based on category membership (sex and race) and that this bias, when assessed by rankings, generally increased with time.

We must note, however, that only in the case of target sex (but not race) did judgment bias take the form predicted by the shifting standards model. That is, the pro-male sex bias was more pronounced on rankings than on ratings, but in the case of race, White judges evaluated Whites more favorably than non-Whites, regardless of judgment scale. Why might this be? On the basis of both anecdotal and more formal accounts, we believe this is true at least partly because the military is explicit in its use of differential standards for women but not for racial minorities (see Moskos & Butler, 1996). If Army captains incorporate these "rules" regarding standards, their subjective leadership judgments should be adjusted for sex but not race. This is precisely the pattern that emerged.

Furthermore, captains' self-judgments were affected by sex but not race. Women judged themselves more negatively than men judged themselves, but non-Whites and Whites showed comparable patterns of self-judgments. Consistent with the use of within-sex standards to judge the self, evidence of sex-based shifting standards—stronger sex effects on rankings than on ratings—appeared at Time 1. In other words, within-category standards were used to judge the self, just as they were to judge others. In groups with solo women, the shifting standards effect on self-judgments also continued to be documented at Times 2 and 3 (see Figure 3). Much prior literature has documented that solo contexts produce increased attention to and stereotyping of the solo member (Biernat & Vescio, 1993; Kanter, 1977; Taylor, 1981; Taylor et al., 1978; cf. Oakes, 1987). In the present study, we found no evidence that the judgments women received from their groupmates varied as a result of the number of women in the group, but the solo context clearly increased and sustained the tendency for sex to be used as a standard in making self-judgments.

In summary, the data from Study 1 both support and extend predictions from the shifting standards model. However, given the small sample size, and particularly the small number of women, we felt it would be valuable to replicate the findings in a larger, independent sample.

Study 2

Method

Participants were 373 U.S. Army captains attending the 9-week CAS3 training course at Fort Leavenworth. These individuals, like those in Study 1, were assigned by Army personnel to 12- or 13-person groups, and 30 of those groups were designated to participate in this study.[8] This sample was completely independent of the Study 1 sample; training began about 7 months after the session attended by captains in Study

[7] Ideally, it would have been valuable to have objective evidence regarding captains' performance during the CAS3 training (perhaps as assessed by the group leader, though these too may have been subject to various forms of bias), but by policy these were not available to us.

[8] An additional group of 12 captains participated in this study, but we discarded these data as 4 members of the group failed to provide any information. This rendered the judgment estimates less stable than those in the other groups and also signified that the group leader was not supportive of the study. In 26 of the other groups, all members participated, and in the remaining 4 groups, 1–3 members failed to participate. These latter groups were retained in all analyses.

1. A description of the sample by sex, race, and branch of service appears in Table 3.

The procedures for this study were generally the same as those in Study 1, but three key differences were introduced. First, in place of a ranking procedure, participants were asked to perform a modified Q-sort on members of their group. Specifically, they were asked to think about a six-category evaluative system ranging from best to worst. Their task was to place 1 member of their group in the *best* category and 1 in the *worst*, then to place 2 members in the next best category and 2 in the next worst, and finally to place 3 groupmates in each of the two remaining middle categories (if the group contained 13 members, a 4th member was to be placed in the third best category). This procedure is similar to a ranking task, and therefore we conceptualized it as the objective or common-rule assessment in our tests of the shifting standards model. However, because the Q-sort (unlike the ranking task) allows for equivalence in placement of some group members, we viewed it as a less optimal objective measure. Its inclusion therefore created a more conservative testing ground for the shifting standards model (i.e., the Q-sort was less distinct from the rating task than was the ranking procedure used in Study 1). Contributing to the similarity between the judgment tasks in this study, the rating procedure also used a 1–6 response format. Thus, the same number of judgment categories was available for both the Q-sort and the rating task (in Study 1, rankings used a 12- or 13-point system, whereas ratings used a 5-point system).

The second major change was that participants did not judge themselves and each other on a global leadership dimension but rather made two sets of judgments on more specific leadership components—"interpersonal skills important to leadership," and "technical/professional competence." However, because these judgments were highly correlated (*r*s for Q-sorts and ratings, respectively, were .81 and .86 for judgments of others and .67 and .69 for judgments of self) and because nearly identical patterns of effect appeared on each dimension, we combined them into a single leadership assessment.[9] Thus, the critical dependent variables in this study were the mean Q-sort and rating score each target received from his or her groupmates on these two dimensions (other-judgments) and the mean Q-sort and rating score assigned to self on these dimensions (self-judgments).

The final change was that although judgment data were collected at three different points in time, the timing differed slightly from that of Study 1. Initial data collection took place near the end of Day 2 of training (after introductions and considerable formal and informal interaction took place) rather than at Hour 1 (as in Study 1), and the second data collection took place on Day 10 rather than on Day 15. Time 3 data collection took place as in Study 1, but nonparticipation was a serious problem at this point: Of the 30 groups, complete nonparticipation occurred in one group, more than half of the members did not fill out questionnaires in 3 groups, and up to one third of the members did not complete the Time 3 task in 7 groups. For these reasons, we felt that the Time 3 data were suspect and therefore did not include them in our analyses (the Time 3 results in no way challenge the conclusions of this article). Thus, the data reported here were based on judgments of self and groupmates at two points in time—on Days 2 and 10 of the course.

Participants were provided with an alphabetized list of their groupmates and performed the judgment tasks in this order: Q-sort on interpersonal skills, interpersonal skills rating, Q-sort on technical competence, technical competence rating. Although the race, sex, and service branch of each captain was available, participants did not provide information on medals, badges, or year of commissioning, and this information was not procurable. Thus, the analyses do not include the same controls for past awards as were possible in Study 1.

Table 3
Sample Frequencies by Arm, Sex, and Race, Study 2

	Combat arm and sex						
	Combat Arms		Combat Support		Combat Service Support		
Race	Men	Women	Men	Women	Men	Women	Total
White	88	0	105	19	55	15	282
Black	12	0	26	10	10	0	58
Asian	1	0	4	0	3	0	8
Hispanic	2	0	8	3	2	0	15
Native American	2	0	0	0	0	0	2
Other	3	0	3	1	0	1	8
Total	108	0	146	33	70	16	373

Results and Discussion

Social Category-Based Ratings and Q-Sort Judgments Over Time

Because of the larger sample size in this study compared with Study 1, we were able to conduct analyses that simultaneously included target race (coded as White vs. non-White[10]) and sex. However, because of the lack of women in the combat arms and the single non-White female service support officer (see Table 3), we could not include branch of service as an additional

[9] The patterns for interpersonal skill and technical competence judgments were always in the same direction and nearly always significant in each separate analysis; when differences appeared, they were small in size (e.g., a p value of $<.05$ in one analysis might be $<.07$ in the other). To further examine differences in these two sets of judgments, we entered a leadership component as an additional repeated measure in our Sex × Race × Scale × Time analyses described in the *Results and Discussion* section. Almost every interaction F involving the leadership component was <1; the only effect that approached significance was a Target Sex × Component interaction, $F(1, 369) = 3.55$, $p < .07$. Judgments of interpersonal skills were unaffected by target sex (Ms = .01 and −.002 for female and male targets, respectively), but, consistent with stereotypes, women tended to be judged less favorably than men on technical competence (Ms = −.15 and .02). This finding cut across time and judgment scale and did not change or challenge our discussion of higher order interactions (which appeared on each component as well as on the combined index) in the *Results and Discussion* section.

[10] As in Study 1, we found some variability across specific minority groups in judgments received. The average standardized evaluations received by White, Asian, Hispanic, Native American, Black, and other targets, respectively, were .10, .08, .18, .29, −.49, and −.67. In this study, Blacks and "other" minorities were clearly discrepant from (judged more negatively than) any other group. For this reason, all the analyses reported in the text were recomputed in a variety of ways: (a) by comparing Whites with Blacks/others only (deleting the other minority group members), (b) by comparing Whites, Asians, Hispanics, and Native Americans with Blacks/others, and (c) by comparing Whites and Asians with all other minorities. In general, because Blacks represent the largest proportion of our non-White sample, these various deletions and regroupings of the data all produced very similar results (though analyses using the White vs. Blacks/others distinction produced stronger effects than those reported in the text). For ease of comparison with the Study 1 findings, we chose to focus on the White/non-White racial distinction.

variable. We did, however, control for branch of service as in Study 1 by first regressing judgments on this variable (along with group size) and then analyzing the residuals. Similar to Study 1, the average correlation between branch of service and judgments received was .30. To summarize, the mean Q-sort and rating scores (averaged across judgments of technical competence and interpersonal skills) each target received from his or her groupmates[11] were first standardized within judgment type and across time points, regressed on branch of service and group size, and then submitted (in residualized form) to a Race × Sex × Judgment Type × Time mixed-design ANOVA (repeated measures on the last two factors). High numbers again indicate more favorable evaluations.

This analysis revealed significant main effects of target race, $F(1, 369) = 9.85$, $p < .01$, and target sex, $F(1, 369) = 4.25$, $p < .05$, as well as significant Race × Time, $F(1, 369) = 6.79$, and Sex × Time, $F(1, 369) = 10.71$, interactions ($ps < .01$). These interactions were not moderated by judgment scale (three-way interaction $ps > .17$), but for comparison with Study 1, the sex effect is depicted separately for the Q-sort and rating procedures in Figure 4. At Time 1, captains showed no evidence of sex bias in either the Q-sort or the rating task (sex $Fs < 1$), but by Time 2, the sex effect was reliable for each type of judgment, $Fs(1, 369) = 7.14$ and 7.96 for Q-sort and rating, respectively, $ps < .05$. In the case of race, Time 1 judgments (collapsed across judgment type) indicated a significant main effect, $F(1, 369) = 4.31$, $p < .05$, with Whites being judged more favorably ($M = .07$) than non-Whites ($M = -.21$); by Time 2, this effect was magnified, $F(1, 369) = 12.42$, $p < .001$ (comparable $Ms = .22$ and $-.30$). Thus, both sex and race bias notably increased from Time 1 to Time 2, though there was no evidence that the effects were stronger on the Q-sort versus the rating task.

The critical prediction from the shifting standards model is a Target Category × Judgment Scale interaction. The Sex × Scale effect was not reliable in the present study, nor was the Race × Scale interaction, $Fs < 1$. Although only the latter null effect replicates Study 1 findings, the present analysis did reveal a reliable Sex × Race × Scale interaction, $F(1, 369) = 3.69$, $p < .05$. As one might expect because of the lack of sex bias at Time 1, this three-way interaction was reliable only in the Time 2 data, $F(1, 369) = 5.01$, $p < .05$, but not at Time 1, $F(1, 369) = 1.00$, *ns*. Figure 5 depicts the Time 2 interaction in terms of the difference between men and women on each judgment type, separately for each racial group (Whites and non-Whites). Numbers above zero indicate that male targets were judged more favorably than female targets. As can be seen in Figure 5, the signature shifting standards pattern—greater evidence of sex stereotyping on the common rule measure (Q-sort) than on subjective ratings—was detected only for White targets. Indeed, a Sex × Judgment Type ANOVA including only White targets produced a significant interaction, $F(1, 280) = 3.69$, $p = .05$; the sex effect was reliable on the Q-sort judgments, $F(1, 280) = 6.54$, $p < .05$, but not on ratings, $F(1, 280) = 1.45$, $p > .20$. For non-White targets, the comparable analysis indicated a main effect of target sex, $F(1, 89) = 5.22$, $p < .05$, but no interaction with judgment type, $F < 1$. For non-Whites, the sex effect was reliable for both the rating task and the Q-sort ($ps < .05$).[12]

Thus, sex was used as a basis of judgments of non-White officers—women were judged to be less competent than men—but there was no evidence that judgment standards shifted by sex for these targets. For Whites, however, the sex-based shifting standards pattern was documented. We also analyzed these data to see if there was evidence of race-based shifting standards by conducting separate Race × Judgment Type ANOVAs for male and female targets; the interaction was nonsignificant in both cases. Thus, standards shifted on the basis of sex (for White targets) but not on the basis of race. As in Study 1, we found no evidence that the number of women in a group (1 vs. 2) influenced judgment patterns (all Fs involving this factor < 1).

Race of judge effects. We also examined whether the race of the judge affected judgments received by White and non-White targets. We separately calculated the evaluations targets received from their White and non-White groupmates, then submitted these to a Target Race × Target Sex × Time × Judgment Type × Judge Race mixed-model ANOVA (again in residualized form). In addition to the effects described previously, this analysis indicated a significant Target Race × Judge Race interaction, $F(1, 369) = 11.06$, $p < .01$, which was not moderated by either scale type or time (three-way interaction $Fs < 1.40$, $ps > .20$). White judges evaluated non-Whites ($M = -.31$) significantly more negatively than Whites ($M = .20$), $F(1, 369) = 14.26$, $p < .001$, whereas non-White judges' evaluations were unaf-

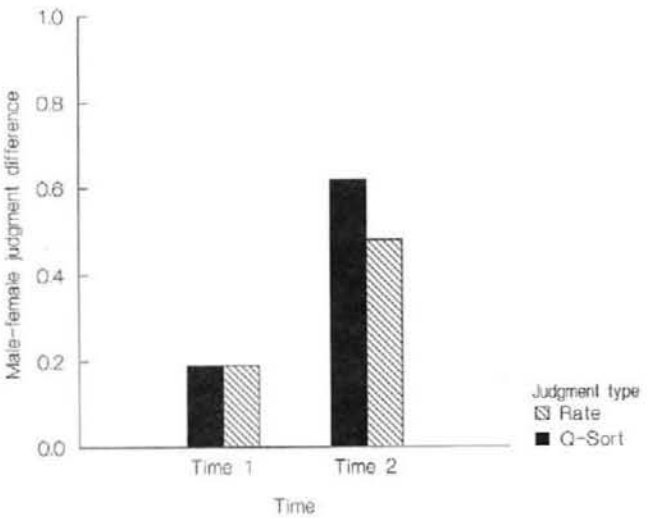

Figure 4. Sex differential in leadership judgments of groupmates by response scale and time point, Study 2.

[11] Interjudge agreement was modest at Time 1 (mean Cronbach's αs across groups = .69 and .59, respectively, for Q-sort and ratings) and notably higher by Time 2 (mean αs = .79 and .70).

[12] In Study 1, we did not explicitly test for the Sex × Race × Judgment Type interaction, given our small N. After noting the Study 2 findings, however, we revisited Study 1 to compute this interaction. Though it was not statistically reliable, $F(1, 96) = 2.05$, $p < .16$, it was the case that the Sex × Judgment Type interaction was significant for White targets, $F(1, 77) = 5.63$, $p < .05$, but not for non-White targets, $F < 1$ ($n = 21$). Thus, both data sets generally supported the finding of sex-based standard shifts for White but not for non-White targets.

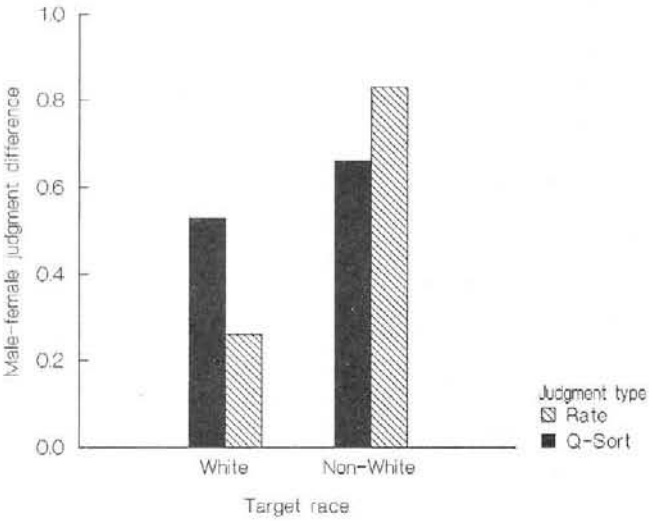

Figure 5. Sex differential in leadership judgments of groupmates by target race and response scale, Time 2 of Study 2.

fected by target race, $F(1, 369) = 1.10$, *ns* (comparable *M*s = −.04 and .01). In Study 1, we found that the pro-White bias by White judges increased over time; the lack of a time effect in this study may be attributable to the fact that Time 1 data collection took place after a day of interaction among group members rather than on first meeting.

Sex of judge effects. Given the larger number of groups in this study, it was also possible to examine whether the sex of the judge affected the judgments captains received. This analysis had to be restricted to groups that contained 2 women so that any given woman in a group received evaluations from men and 1 woman; men received evaluations from 2 women (*n*s = 36 women and 199 men). In addition to findings reported earlier, this analysis revealed a significant Judge Sex × Target Sex × Time interaction, $F(1, 231) = 5.25$, $p < .05$. Among male judges, the Target Sex × Time interaction was reliable, $F(1, 234) = 4.61$, $p < .05$; among female judges, it was not ($F < 1$). Neither male nor female judges showed evidence of sex bias at Time 1 (male–female difference = .01 for male judges and −.03 for female judges; Judge Sex × Target Sex interaction, $F < 1$). At Time 2, the Judge Sex × Target Sex interaction was reliable, $F(1, 232) = 6.02$, $p < .05$. Only male judges showed a significant pattern of in-group bias (male $M = .04$, female $M = -.26$); female judges tended to judge women nonsignificantly more favorably than men (male $M = -.02$, female $M = .08$, $F < 1$). These effects cut across judgment scale type and support a general tendency toward in-group bias by Time 2, particularly among male judges.

Self-Rating and Q-Sort Judgments

Finally, participants' standardized self-judgments on the Q-sort and rating task (averaged across the two dimensions of technical competence and interpersonal skills) were regressed on branch of service and group size, and the resulting residuals were submitted to a Target Sex × Target Race × Judgment Type × Time × Number of Women in Group mixed-model ANOVA (complete self-judgment data were available from 10 women in solo groups and 35 in nonsolo groups). There was no evidence that self-judgments varied by race ($F < 1$), but they were affected by sex, as evidenced in the Sex × Judgment Type × Number of Women interaction, $F(1, 333) = 6.00$, $p < .02$ (no other significant effects emerged). Time did not moderate this effect, $F(1, 333) = 2.06$, $p < .16$; however, Figure 6 depicts the data separately for each time point so that comparisons can be made with Study 1 (see Figure 3). The bars in the figure reflect the male–female differential in self-judgments—numbers above zero indicate that men's self-judgments were more favorable than women's self-judgments.

Consistent with the pattern reported in Study 1, groups including solo women, but not those including 2 women, produced a pattern of sex-based judgment shifts in self-ratings: In solo-woman groups, the male–female difference was marked on the Q-sort task but eliminated on the rating task: A separate Sex × Race × Time × Judgment Type ANOVA on these groups revealed a significant Sex × Judgment Type interaction, $F(1, 121) = 5.69$, $p < .05$. In groups including 2 women, sex was not a

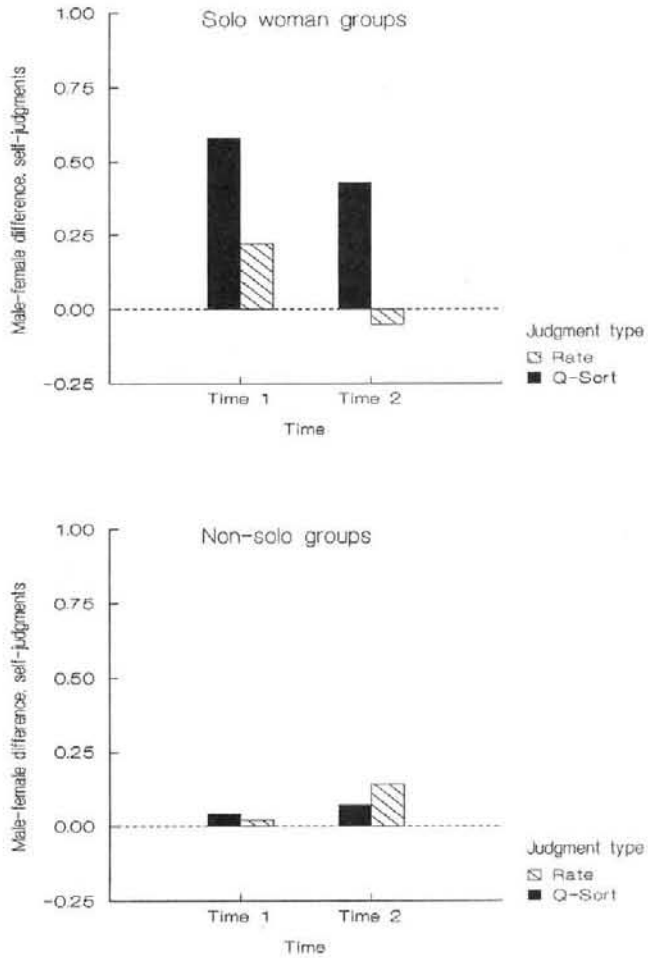

Figure 6. Sex differential in leadership self-judgments by response scale and time point, separately for groups with 1 versus 2 women, Study 2.

basis of self-judgments, either as a main effect or in interaction with judgment type ($Fs < 1$). Thus, with the larger N of Study 2, we replicated the Study 1 data regarding groups with solo women (including the lack of change with time), and found no evidence of within-sex standard use in groups with 2 women.

Summary

The Study 2 data provided a nearly complete replication of the central Study 1 findings. First, judgment standards shifted on the basis of sex but not race (though only for White targets). These findings paint a consistent and remarkable picture: Officers appeared to duplicate Army policy by applying different standards to women and men, yet evaluating Whites and non-Whites relative to a single criterion. More on this point appears in the General Discussion. Second, patterns of stereotyping increased with time, a finding consistent with the stereotypes as hypotheses approach (Darley & Gross, 1983). Third, White officers showed marked evidence of pro-White bias in target evaluations, whereas non-White officers did not reliably distinguish between the groups; men also showed more evidence of in-group bias than did women. And finally, self-judgments were unaffected by race, but there was clear evidence of sex-based shifting standards in self-judgments in groups containing solo women.

General Discussion

The Army setting in which these data were collected provided a meaningful real-world context in which to examine and test a number of theoretical extensions of the shifting standards model. We predicted that standards were more likely to shift on the basis of sex than race; that longitudinal trends in stereotyping effects would be more marked on objective than on subjective scales; that self-judgments, like other-judgments, would show evidence of standard shifts; and that groups with solo women relative to groups with 2 women would show magnified evidence of the operation of sex-based shifting standards. The data were largely supportive of each of these predictions.

Sex and Race Bias in Judgments

Despite changes in procedure and measurement across studies, the data yielded consistent evidence of the use of sex-based shifting standards in Army captains' judgments of each others' leadership competence. In Study 1, the expected shifting standards pattern appeared at all time points (greater evidence of sex bias in rankings than in ratings), and in Study 2, this pattern emerged at Time 2, though it was specific to judgments of White targets.

This latter effect should be explored more closely. It is important to note that in both studies, judgments of non-White officers were influenced by sex—non-White women were judged less competent than non-White men—but there was no evidence of differential sex bias on subjective versus common-rule (Q-sort and ranking) scales. This suggests that for non-White targets, subjective judgments were not made with reference to sex-specific standards; instead, the subjective judgment scale functioned like the objective scale. Why might this be? If the Army's invocation against the use of different race-based standards is taken to heart, officers may have been reluctant to apply different standards to judge non-White officers. Instead, they seem to have applied a single high standard to judge both non-White males and females, as evidenced in the lower evaluations of non-Whites on both objective and subjective rating scales in both studies. Given default values, we assume that the standard was most likely based on expectations for White males, the prototypical officers in the Army and at CAS3. Use of a White male standard for judging non-Whites would allow for the revelation of perceived differences between men and women regardless of the response scale being used.

These data indicate that category-based bias will not always produce standard shifts that are captured in divergent results on objective and subjective scales. In this sense, the findings add an important caveat to the shifting standards model: Although the shifting standards pattern (greater evidence of bias on objective than subjective scales) indicates that stereotypes are being used, a stereotype's influence need not be manifested in this pattern—it can take other forms. In these studies, the race-based judgment patterns may have been due to general in-group favoritism or in-group bias (Figure 2). However, racial minorities did not reliably show a prominority bias in Study 1, and their judgments tended to be pro-White, though not significantly so, in Study 2. It seems more likely that captains (particularly White captains) held a general antiminority stereotype (e.g., that minorities are less competent than Whites) that manifested itself on both objective and subjective scales because of the Army's explicit policy and instruction that race-based standard shifts are inappropriate. A common judgment framework essentially converts the subjective scale into a common-rule scale; antiminority stereotyping is then evident regardless of judgment format.

Although we believe that the Army's differential policy on sex- and race-based standards was responsible for the different forms that sex and race bias took in this study, further work is clearly necessary to better delineate the conditions under which standard shifts do and do not follow from the activation of group stereotypes. The present research suggests that stereotypes are a necessary, though not sufficient, contributor to shifting standards effects. Obviously, if no stereotype exists, no differential standards will be called to mind (see also Biernat et al., 1991). Similarly, a target person must be categorized as a member of a group in order for the group-specific standard to be relevant. Given categorization and a relevant stereotype, however, a variety of situational factors may moderate the application of differential judgment standards.

For example, differential standard use may be either normatively inappropriate (as in the case of race in the present study) or normatively appropriate (as in the case of sex); the context may also dictate the rationality of standard use (e.g., it seems both reasonable and kind to evaluate the verbal competency of foreign graduate school applicants relative to a lower standard than U.S. applicants). Social desirability concerns may be particularly likely to enhance the use of within-group judgment standards, as subjective language that is defined in reference to a low category standard will be more favorable to the target. In some circumstances, the context may impose its own evaluative standards that override those suggested by the stereotype (e.g., a job may require a specific set of qualifications against which

applicants are compared, regardless of their group membership). Some standard shifts may also be more habitual, perhaps automatic, than others. One can contrast the case of sex-specific height standards, which our earlier research indicates are used tenaciously (Biernat et al., 1991; Nelson, Biernat, & Manis, 1990), with the case of sex-specific athletic standards, which are more readily put aside in response to instructional sets (Biernat, 1995; Biernat & Manis, 1994).

Furthermore, a variety of other motivational orientations may affect the tendency to use differential standards when judging members of stereotyped groups. Motives for accuracy or accountability may promote the use of a single judgment standard, and interdependence or relevance of the target for the self may focus perceivers on individuating rather than category attributes of the target (e.g., Brewer, 1988; Fiske & Neuberg, 1990), thereby reducing the likelihood that the individual will be thought of (and judged relative to) his or her group. On the other hand, strong anti–out-group or pro–in-group sentiment may be evidenced regardless of the judgment scale (objective or subjective) in use; in Study 2, for example, men showed an overarching tendency to judge women as less competent than men at Time 2. To date, the shifting standards model has been largely cognitive in its emphasis; integrating the cognitive mechanisms of this model with motivational factors seems a worthy endeavor.

Regardless of the specific contributors to judgment in a given setting, a central message of the shifting standards model is that if researchers seek to accurately assess perceivers' mental representations of targets, they should use common-rule (objective) judgment measures whenever possible. These measures avoid the interpretational problems that are introduced when the meaning of rating units can be adjusted in category-specific or idiosyncratic ways. In the present studies, such measures best indicated sex bias in leadership perceptions and in self-judgments, whereas subjective assessments masked these effects. Common-rule measures also suffice in the absence of standard shifts, as in the race-based effects described here. Because they avoid the potential for within-category meaning shifts (and the subsequent difficulty of making cross-group comparisons), common-rule assessments such as rankings and externally anchored judgment units (inches, dollars, hours, test scores) will better serve the researcher.

We should emphasize, however, that these measurement recommendations apply to situations in which researchers are examining judgments of individual members of stereotyped groups. When measuring stereotypes of groups as a whole (e.g., how good at leadership are men vs. women?), shifting standards are not likely to be introduced; rather, judges understand that they are to use a single interpretation of the trait dimension such that the two groups can be reasonably compared and distinguished. For this reason, subjective (e.g., Likert-type) measures may be quite appropriate for measuring group-level stereotypes (see Biernat & Crandall, 1996). Judges are likely to evaluate two different groups against a common standard, but two different individuals against shifting standards on the basis of their group memberships.

Longitudinal Trends

In both studies, we found a general pattern of increased sex stereotyping with time. However, this time effect was not statistically reliable in Study 1, though it appeared most clearly on rankings rather than ratings (consistent with the shifting standards model). In Study 2, the pattern of sex-based shifting standards for White targets emerged only at Time 2. Both studies also supported a clear pattern of increased racial stereotyping with time, particularly by White officers. Virtually no race bias was evident at Time 1, but by Time 2, non-White officers were derogated relative to White officers. These data are clearly inconsistent with the hypothesis that contact decreases stereotyping, but are quite compatible with the Darley and Gross (1983) suggestion that stereotypes serve as initial hypotheses, which require behavioral information to confirm. At Time 1 of Study 1, when participants were strangers, there was mild evidence of sex stereotyping and no evidence of racial stereotyping. By Time 2 (3 weeks later), evidence of these stereotypes was in full bloom (and Time 2 to Time 3 comparisons indicated slight further increases in these stereotyping trends). At Time 1 of Study 2, when participants had known each other for roughly 2 days, no sex and little race bias was evident; by Day 10, however, women and non-Whites were judged more negatively than men and Whites. To the extent that perceivers require behavioral evidence before they are willing or able to express their stereotypes, the data from Study 2 suggest that 2 days' acquaintance are not sufficient—after 2 days, participants were no more biased by sex stereotypes than they were at zero acquaintance in Study 1.

Self-Judgments and Context Effects

The present data additionally suggest that the shifting standards model can be applied to the domain of self-judgments (see also Biernat, Manis, & Kobrynowicz, 1997). Similar to the pattern documented for judgments of others, pro-male bias in self-judgments was stronger on rankings and Q-sorts than on ratings, particularly in groups that included only 1 woman. We believe that women and men engaged in a process of self-stereotyping along gender lines. On common-rule scales (rankings, Q-sort), this stereotyping was revealed in a straightforward fashion—women judged themselves as less competent than men judged themselves. On subjective scales (ratings), self-stereotyping was manifested in captains' judgments of themselves relative to their sex category—women evaluated themselves relative to women, and men to men, resulting in decreased sex differentiation in judgments.

Thus, just as perceivers may evaluate a female groupmate whom they ranked low in leadership competence as subjectively "good (for a woman)," they may also apply similar reasoning to evaluations of themselves. The intensification of this pattern in groups that included only 1 woman was likely due to the heightened salience of sex as a judgment cue. Solo women may have been particularly likely to view themselves as women (i.e., as stereotypically low in leadership competence), resulting in low self-placement in the Q-sort array and in (higher) subjective evaluation relative to other women. Men in groups with only 1 woman may also have been more likely to self-categorize as men, resulting in relatively high self-placement in the Q-sort and (lower) subjective evaluation relative to other men.

Solo status did not, however, affect how women and men were judged by their groupmates; there was no evidence in either study that solo women were more strongly stereotyped

than nonsolo women. Perhaps to perceivers, a 1:12 women-to-men ratio makes sex no more salient than a 2:11 ratio, particularly in the broader military context where women are always a small minority. But when judging the self, the solo woman is particularly likely to be cognizant of her sex and, therefore, likely to self-stereotype (Mullen, 1983). These findings point to the need for further research on how context may differentially affect the salience of category cues for judgments of others versus the self. More generally, the self-judgment data indicate that the shifting standards model may aid in understanding and elaborating on the processes and outcomes of social comparison and self-stereotyping.

Conclusion

The present data simultaneously support, extend, and delimit the shifting standards model. Both studies replicated a pattern of sex-based shifting standards in judgments of others and demonstrated that self-judgments are similarly affected by differential standard use. At the same time, the fact that racial stereotyping occurred without the operation of differential standards indicates that the use of within-category judgment standards is not an automatic consequence of stereotype activation (see Biernat et al., 1998). More important, these studies demonstrated that contextual factors—for example, the military's normative structure and policy regarding standard use, the number of category members present in a group—may either nullify or intensify the tendency for individuals to use within-category judgment standards.

At an applied level, these data indicate that sex and, in a different manner, race remain important distinguishing characteristics among advanced Army officers. Judgments of groupmates' leadership competence were biased by both of these cues, and sex affected self-judgments as well. In the confines of our data collection procedures, we could not tease apart the precise mechanism by which these categories affected judgments, though potential candidates include biased or confirmatory information processing and limited opportunity for women (and perhaps to a lesser extent, racial minorities) to demonstrate leadership skills in the context of the course. Though we favor the former account, further work in both military and other contexts should seek to establish the precise processes through which stereotypes and standard shifts exert their influence on judgment and behavior.

References

Albright, L., Kenny, D. A., & Malloy, T. E. (1988). Consensus in personality judgments at zero acquaintance. *Journal of Personality and Social Psychology, 55,* 387–395.

Allport, G. W. (1954). *The nature of prejudice.* Reading, MA: Addison-Wesley.

Amir, Y. (1969). Contact hypothesis in ethnic relations. *Psychological Bulletin, 71,* 319–342.

Amir, Y. (1976). The role of intergroup contact in change of prejudice and ethnic relations. In P. Katz (Ed.), *Towards the elimination of racism* (pp. 245–308). New York: Pergamon Press.

Bem, S. L. (1974). The measurement of psychological androgyny. *Journal of Consulting and Clinical Psychology, 42,* 155–162.

Biernat, M. (1995). The shifting standards model: Implications of stereotype accuracy for social judgment. In Y. T. Lee, L. J. Jussim, & C. R. McCauley (Eds.), *Stereotype accuracy: Toward appreciating group differences* (pp. 87–114). Washington, DC: American Psychological Association.

Biernat, M., & Crandall, C. S. (1996). Creating stereotypes and capturing their content. *European Journal of Social Psychology, 26,* 867–898.

Biernat, M., & Kobrynowicz, D. (1997). Gender- and race-based standards of competence: 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., & Kobrynowicz, D. (1997). Simultaneous assimilation and contrast effects in judgments of self and other. *Journal of Personality and Social Psychology, 73,* 254–269.

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

Biernat, M., & Vescio, T. K. (1993). Categorization and stereotyping: Effects of group context on memory and social judgment. *Journal of Experimental Social Psychology, 29,* 166–202.

Biernat, M., Vescio, T. K., & Green, H. L. (1996). Selective self-stereotyping. *Journal of Personality and Social Psychology, 71,* 1194–1209.

Biernat, M., Vescio, T. K., & Manis, M. (1998). Judging and behaving toward members of stereotyped groups: A shifting standards perspective. In C. Sedikides, J. Schopler, & C. Insko (Eds.), *Intergroup cognition and intergroup behavior* (pp. 151–175). Hillsdale, NJ: Erlbaum.

Brewer, M. B. (1988). A dual process model of impression formation. In T. K. Srull & R. S. Wyer (Eds.), *Advances in social cognition* (Vol. 1, pp. 1–36). Hillsdale, NJ: Erlbaum.

Brown, V., & Geis, F. L. (1984). Turning lead into gold: Evaluations of men and women leaders and the alchemy of social consensus. *Journal of Personality and Social Psychology, 46,* 811–824.

Butler, D., & Geis, F. L. (1990). Nonverbal affect responses to male and female leaders: Implications for leadership evaluations. *Journal of Personality and Social Psychology, 58,* 48–59.

Buunk, B. P., & VanYperen, N. W. (1991). Referential comparisons, relational comparisons, and exchange orientation: Their relation to marital satisfaction. *Personality and Social Psychology Bulletin, 17,* 709–717.

Combined Arms and Services Staff School Office. (1997). *Combined arms and services staff school* [description]. Retrieved June 16, 1998, from the World Wide Web: http://www-cgsc.army.mil/cas3/cas3info.htm

Cook, S. W. (1978). Interpersonal and attitudinal outcomes in cooperating interracial groups. *Journal of Research and Development in Education, 12,* 97–113.

Darley, J. M., & Gross, P. H. (1983). A hypothesis-confirming bias in labeling effects. *Journal of Personality and Social Psychology, 44,* 20–33.

Defense Equal Opportunity Management Institute. (1995). *Active duty Army: Distribution of active duty forces* [On-line]. Available: http://www.pafb.af.mil/deomi/dr_stat1.htm

Devine, P. G., & Elliot, A. J. (1995). Are racial stereotypes really fading? The Princeton trilogy revisited. *Personality and Social Psychology Bulletin, 21,* 1139–1150.

Eagly, A. H., Karau, S. J., & Makhijani, M. G. (1995). Gender and the effectiveness of leaders: A meta-analysis. *Psychological Bulletin, 117,* 125–145.

Eagly, A. H., Makhijani, M. G., & Klonsky, B. G. (1992). Gender and

the evaluation of leaders: A meta-analysis. *Psychological Bulletin, 111,* 3–22.

Festinger, L. (1954). A theory of social comparison processes. *Human Relations, 7,* 71–82.

Fiske, S. T., & Neuberg, S. L. (1990). A continuum of impression formation, from category-based to individuating processes: Influences of information and motivation on attention and interpretation. In M. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 23, pp. 1–74). New York: Academic Press.

Francke, L. B. (1997). *Ground zero: The gender wars in the military.* New York: Simon & Schuster.

Funder, D. C. (1995). Stereotypes, base rates, and the fundamental attribution mistake: A content-based approach to judgmental accuracy. In Y. T. Lee, L. J. Jussim, & C. R. McCauley (Eds.), *Stereotype accuracy: Toward appreciating group differences* (pp. 141–156). Washington, DC: American Psychological Association.

Gaertner, S. L., & Dovidio, J. F. (1986). The aversive form of racism. In J. F. Dovidio & S. L. Gaertner (Eds.), *Prejudice, discrimination, and racism* (pp. 61–89). Orlando, FL: Academic Press.

Goethals, G. R., & Darley, J. M. (1977). Social comparison theory: An attributional approach. In J. Suls & R. Miller (Eds.), *Social comparison processes: Theoretical and empirical perspectives* (pp. 259–278). Washington, DC: Hemisphere.

Halpin, S. M. (1970). Complex social comparison (Doctoral dissertation, Purdue University, 1970). *Dissertation Abstracts International, 31,* 5518.

Hardie, E. A., & McMurray, N. E. (1992). Self stereotyping, sex role ideology, and menstrual attitudes: A social identity approach. *Sex Roles, 27,* 17–37.

Haslam, S. A., Oakes, P. J., Turner, J. C., & McGarty, C. (1996). Social identity, self-categorization, and the perceived homogeneity of ingroups and outgroups: The interaction between social motivation and cognition. In R. M. Sorrentino & E. T. Higgins (Eds.), *Handbook of motivation and cognition: Vol 3. The interpersonal context* (pp. 182–222). New York: Guilford Press.

Heilman, M. E., & Kram, K. E. (1983). Male and female assumptions about colleagues' views of their competence. *Psychology of Women Quarterly, 7,* 329–337.

Hewstone, M. (1996). Contact and categorization: Social psychological interventions to change intergroup relations. In C. N. Macrae, C. Stangor, & M. Hewstone (Eds.), *Stereotypes and stereotyping* (pp. 323–368). New York: Guilford Press.

Hilton, J. L., & Von Hippel, W. (1996). Stereotypes. *Annual Review of Psychology, 47,* 237–271.

Hogg, M. A., & Turner, J. C. (1987). Intergroup behaviour, self-stereotyping and the salience of social categories. *British Journal of Social Psychology, 26,* 325–340.

Judd, C. M., & Park, B. (1993). Definition and assessment of accuracy in social stereotypes. *Psychological Review, 100,* 109–128.

Kanter, R. M. (1977). *Men and women of the corporation.* New York: Basic Books.

Kobrynowicz, D., & Biernat, M. (1997). Do the same traits imply the same behavior? Shifting standards in the interpretation of trait concepts. *Journal of Experimental Social Psychology, 33,* 529–601.

Lau, R. R. (1989). Individual and contextual influences on group identification. *Social Psychology Quarterly, 5,* 220–231.

Leyens, J. P., Yzerbyt, V., & Schadron, G. (1994). *Stereotypes and social cognition.* London: Sage.

Lorenzi-Cioldi, F. (1991). Self-stereotyping and self-enhancement in gender groups. *European Journal of Social Psychology, 21,* 403–417.

Major, B. (1989). Gender differences in comparisons and entitlement: Implications for comparable worth. *Journal of Social Issues, 45,* 99–115.

Major, B. (1993). Gender, entitlement, and the distribution of family labor. *Journal of Social Issues, 49,* 141–159.

Major, B., & Forcey, B. (1985). Social comparisons and pay evaluations: Preferences for same-sex and same-job wage comparisons. *Journal of Experimental Social Psychology, 21,* 393–405.

Major, B., & Testa, M. (1989). Social comparison processes and judgments of entitlement and satisfaction. *Journal of Experimental Social Psychology, 25,* 101–120.

Malloy, T. E., & Janowski, C. L. (1992). Perceptions and metaperceptions of leadership: Components, accuracy, and dispositional correlates. *Personality and Social Psychology Bulletin, 18,* 700–708.

Martindale, M. (1990). *Sexual harassment in the military: 1988.* Report, Defense Manpower Data Center, Arlington, VA.

Moskos, C. C., & Butler, J. S. (1996). *All that we can be: Black leadership and racial integration the Army way.* New York: Basic Books.

Mullen, B. (1983). Operationalizing the effect of the group on the individual: A self-attention perspective. *Journal of Experimental Social Psychology, 19,* 295–322.

Nelson, T. E., Biernat, M., & Manis, M. (1990). Everyday base rates (sex stereotypes): Potent and resilient. *Journal of Personality and Social Psychology, 59,* 664–675.

Niemann, Y. F., Jennings, L., Rozelle, R. M., Baxter, J. C., & Sullivan, E. (1994). Use of free responses and cluster analysis to determine stereotypes of eight groups. *Personality and Social Psychology Bulletin, 20,* 379–390.

Oakes, P. (1987). The salience of social categories. In J. C. Turner, *Rediscovering the social group: A self-categorization theory* (pp. 117–141). Oxford, England: Basil Blackwell.

Parducci, A. (1963). Range-frequency compromise in judgment. *Psychological Monographs, 77*(2, Whole No. 565).

Parducci, A. (1965). Category judgment: A range frequency model. *Psychological Review, 72,* 407–418.

Pettigrew, T., & Martin, J. (1987). Shaping the organizational context for Black American inclusion. *Journal of Social Issues, 43,* 41–78.

Postman, L., & Miller, G. A. (1945). Anchoring of temporal judgments. *American Journal of Psychology, 58,* 43–53.

Presidential Commission on the Assignment of Women in the Armed Forces. (1992). *Women in combat: Report to the President.* Washington, DC: Brassey's.

Pryor, J. B. (1988). *Sexual harassment in the United States military: The development of the DOD survey.* Patrick Air Force Base, FL: Defense Equal Opportunity Management Institute.

Pulakos, E. D., White, L. A., Oppler, S. H., & Borman, W. C. (1989). Examination of race and sex effects on performance ratings. *Journal of Applied Psychology, 74,* 770–780.

Sears, D. O. (1986). College sophomores in the laboratory: Influences of a narrow data base on psychology's view of human nature. *Journal of Personality and Social Psychology, 51,* 515–530.

Simon, B., Glässner-Bayerl, B., & Stratenwerth, I. (1991). Stereotyping and self-stereotyping in a natural intergroup context: The case of heterosexual and homosexual men. *Social Psychology Quarterly, 54,* 252–266.

Simon, B., & Hamilton, D. L. (1994). Self-stereotyping and social context: The effects of relative in-group size and in-group status. *Journal of Personality and Social Psychology, 66,* 699–711.

Smither, R. D., & Houston, M. R. (1991). Racial discrimination and forms of redress in the military. *International Journal of Intercultural Relations, 15,* 459–468.

Snyder, M., & Cantor, N. (1979). Testing hypotheses about other people: The use of historical knowledge. *Journal of Experimental Social Psychology, 15,* 330–342.

Snyder, M., & Swann, W. B. (1978). Hypothesis-testing processes in social interaction. *Journal of Personality and Social Psychology, 36,* 1202–1212.

Spence, J. T., Helmreich, R., & Stapp, J. (1974). The Personal Attributes Questionnaire: A measure of sex-role stereotypes and masculinity-femininity. *JSAS Catalog of Selected Documents in Psychology, 4,* 43–44.

St. Pierre, M. (1991). Accession and retention of minorities: Implications for the future. *International Journal of Intercultural Relations, 15,* 469–489.

Stangor, C., & Lange, J. E. (1994). Mental representations of social groups: Advances in understanding stereotypes and stereotyping. *Advances in Experimental Social Psychology, 26,* 357–416.

Taylor, S. E. (1981). A categorization approach to stereotyping. In D. L. Hamilton (Ed.), *Cognitive processes in stereotyping and intergroup behavior* (pp. 83–114). Hillsdale, NJ: Erlbaum.

Taylor, S. E., & Fiske, S. T. (1978). Salience, attention, and attribution: Top-of-the-head phenomena. In L. Berkowitz (Ed.), *Advances in Experimental Social Psychology* (Vol. 11, pp. 249–288). New York: Academic Press.

Taylor, S. E., Fiske, S. T., Etcoff, N., & Ruderman, A. (1978). The categorical and contextual bases of person memory and stereotyping. *Journal of Personality and Social Psychology, 36,* 778–793.

Terkel, S. (1980). *American dreams: Lost and found.* New York: Ballantine Books.

Turner, J. C. (1982). Towards a cognitive redefinition of the social group. In H. Tajfel (Ed.), *Social identity and intergroup relations* (pp. 15–40). Cambridge, England: Cambridge University Press.

Turner, J. C., Hogg, M. A., Oakes, P. J., Reicher, S. D., & Wetherell, M. (1987). *Rediscovering the social group: A self-categorization theory.* Oxford, England: Basil Blackwell.

Upshaw, H. S. (1962). Own attitude as an anchor in equal-appearing intervals. *Journal of Abnormal and Social Psychology, 64,* 85–96.

Upshaw, H. S. (1969). The personal reference scale: An approach to social judgment. In L. Berkowitz (Ed.), *Advances in Experimental Social Psychology* (Vol. 4, pp. 315–371). New York: Academic Press.

U.S. Department of Defense. (1988). *Report: Task force on women in the military.* Washington, DC: Author.

Volkmann, J. (1951). Scales of judgment and their implications for social psychology. In J. H. Rohrer & M. Sherif (Eds.), *Social psychology at the crossroads* (pp. 273–294). New York: Harper.

Von Hippel, W., Sekaquaptewa, D., & Vargas, P. (1995). On the role of encoding processes in stereotype maintenance. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 27, pp. 177–254). New York: Academic Press.

Wood, J. V. (1989). Theory and research concerning social comparisons of personal attributes. *Psychological Bulletin, 106,* 231–248.

Yzerbyt, V. Y., Schadron, G., Leyens, J. P., & Rocher, S. (1994). Social judgeability: The impact of meta-informational cues on the use of stereotypes. *Journal of Personality and Social Psychology, 66,* 48–55.

Zanna, M., Goethals, G. R., & Hill, J. (1975). Evaluating a sex-related ability: Social comparison with similar others and standard setters. *Journal of Experimental Social Psychology, 11,* 86–93.

Received January 31, 1997
Revision received January 2, 1998
Accepted January 21, 1998 ■