Shifting Standards and Stereotype-Based Judgments

Monica Biernat and Melvin Manis

Four studies tested a model of stereotype-based shifts in judgment standards developed by M. Biernat, M. Manis, and T. E. Nelson (1991). The model suggests that subjective judgments of target persons from different social groups may fail to reveal the stereotyped expectations of judges, because they invite the use of different evaluative standards; more "objective" or common rule indicators reduce such standard shifts. The stereotypes that men are more competent than women, women are more verbally able than men, Whites are more verbally able than Blacks, and Blacks are more athletic than Whites were successfully used to demonstrate the shifting standards phenomenon. Several individual-difference measures were also effective in predicting differential susceptibility to standard shifts, and direct evidence was provided that differing comparison standards account for substantial differences in target ratings.

When judging individuals from different social groups, one may implicitly refer to his or her conception of the group mean or standard on the dimension of interest as an important reference criterion. For example, when one is asked, "how tall is Julia?" an answer of "very tall" can generally be taken to mean that Julia is very tall relative to the average woman. She might, however, measure only 5'9—a height that would not be referred to as "very tall" if characteristic of a man. Similarly, what is deemed to be "very assertive" behavior in a woman may be quite different from what is deemed to be "very assertive" in a man. In both of these cases, different standards of judgment are being used to evaluate members of each sex. This occurs, we believe, because people implicitly accept the stereotypes (accurate or not) that men are, on average, taller and more assertive than women. Holding these stereotypes means that the standard one calls to mind when judging a woman's height or assertiveness will be quite different from that called to mind when judging a man. In essence, then, we hypothesize that people routinely shift or adjust their standards of judgment as they think about members of different social groups (see Foddy & Smithson, 1989; Kahneman & Miller, 1986).

In a recent article, Biernat, Manis, and Nelson (1991) presented a schematic model and supporting evidence for a "shifting standards" effect in judgments about male and female targets' heights, weights, and incomes. In that research, subjects rated a series of photographs on these attributes by using either "subjective" (Likert-type) or "objective" response scales. Objective responses included judgments in feet and inches for height, pounds for weight, and dollars earned per year for income. These are stable, externally anchored units of measurement that retain a constant meaning regardless of the type of exemplar being judged. Biernat et al. (1991) argued that such objective ratings should reflect the mental representations of their subjects with reasonable fidelity. They reasoned, however, that subjective ratings might mask these representations, because they allow for the standard shift phenomenon: Subjects may differentially adjust the meaning of labels such as *very short* and *very tall* when judging male versus female targets. Other researchers who have compared objective and subjective response scales have noted the former's lesser sensitivity to context (e.g., contrast) effects (Campbell, Lewis, & Hunt, 1958; Helson & Kozaki, 1968; Krantz & Campbell, 1961). They suggest that subjective scales allow for semantic changes of meaning of the sort we are proposing here (see Manis, 1967, 1971).

Biernat et al. (1991) found that when subjects judged personal income by indicating "dollars earned per year," men were rated as earning more than women. In other words, the stereotype (in this case, an accurate one) that men make more money than women was clearly reflected in judgments of individual targets. However, when subjects judged income using a subjective scale with endpoints labeled *financially very unsuccessful* and *financially very successful*, women were rated higher than men. These results suggest that when using subjective scales, the judges differentially adjusted the meanings of the end anchors for female and male targets. For a man to be labeled *financially very successful*, he had to earn much more money than a woman who was similarly labeled.

In general, we suggest that whenever one is provided with a subjective response scale on which to evaluate a group of targets (such as women), the end-anchors of the rating scale are shifted so as to maximize differentiation among the class members. This idea is not new to the judgment literature. Volkmann's (1951) "rubber band" model assumes that subjects set the end-

Monica Biernat, Department of Psychology, University of Kansas; Melvin Manis, Department of Psychology, University of Michigan.

This research was supported, in part, by a University of Florida Research Development Award, and preparation of the manuscript was facilitated by National Institute of Mental Health Grant 1R29MH48844-01A2 to Monica Biernat. We gratefully acknowledge the helpful comments of Chris Crandall, Mark Schaller, and several anonymous reviewers on an earlier draft of this article.

Correspondence concerning this article should be addressed to Monica Biernat, Department of Psychology, 426 Fraser Hall, University of Kansas, Lawrence, Kansas 66045-2160. Electronic mail may be sent to biernat@ukanvm (Bitnet).

Journal of Personality and Social Psychology, 1994, Vol. 66, No. 1, 5-20
 0022-3514/94/$3.00

points of their rating scales to match the stimulus range that they anticipate; the subjective meaning of various response categories changes as this stimulus range extends or retracts (see also Postman & Miller, 1945). Parducci's "range" and "frequency" principles make similar predictions about the judge's assignment of stimuli to appropriate rating categories (Parducci, 1963, 1965; Parducci & Perrett, 1971). Upshaw's (1962, 1969) variable perspective model also suggests that judgments are based on where stimuli (in his case, attitudinal stimuli) fall within an individual's subjective frame of reference or "perspective." The novelty of the present approach is in tying the phenomenon of differential scaling adjustments to the stereotyping literature.

A stereotype that differentiates two groups on some relevant dimension implies that these groups will differ with respect to (a) the mean or "typical" value and (b) the range of values that might be anticipated from a sample of the individual group members. For example, men are expected, on average, to be more aggressive than women, and the expected range of aggressiveness in men begins (and ends) at a higher level than the expected range of aggressiveness in women. When a subjective rating scale is introduced, the response values are adjusted to fit these expectations. The result is that two targets—one male and one female—who are characterized in identical terms (e.g., *very aggressive*) may nonetheless be perceived to differ systematically. The *very aggressive* man may have engaged in some behavior that is substantially different (e.g., more objectively aggressive) than that of the *very aggressive* woman. If we could measure aggressiveness using an externally anchored or common-rule scale, different descriptive terms would be used to describe them: Consistent with the stereotype, the man would be judged as more aggressive than the woman. Although we know of no way to objectively measure aggressiveness, in this article we present four studies that test the logic of this reasoning. One of our goals is to demonstrate the influence of stereotypes in social judgments even in situations where they appear not to be operating; that is, when subjective responding allows for standard shifts (see Locksley, Borgida, Brekke, & Hepburn, 1980; Locksley, Hepburn, & Ortiz, 1981).

Our first study examines the gender stereotype or belief that men are more competent than women (see Goldberg, 1968), and the second study focuses on both gender and racial stereotypes regarding verbal ability—that women have more verbal ability than men, and that Whites have more verbal ability than Blacks. Study 3 involves the stereotype that Blacks are more athletic than Whites, and Study 4, which uses a slightly different paradigm, investigates gender-based beliefs about aggression and passivity. The four studies are extensions of our earlier work on the topic of gender stereotype-based standard shifts in judgments of height, weight, and income (Biernat et al., 1991). In the present experiments, however, we examine more meaningful social stereotypes regarding both gender and race, and we were therefore forced to be more creative in developing common-rule or objective measurement metrics because, for example, verbal ability cannot be measured in as neat a unit as an inch or a pound. For that reason, we relied instead on common-rule, or "universalist" assessment procedures such as letter grades and rank orderings as a substitute for objective judgments. In each study, we expect to find that our common-rule response scales reveal clear stereotyping effects, but that subjective response scales—because they can be adjusted to fit different classes of exemplars—dilute and sometimes reverse these effects.

The present research has two additional objectives. One is to examine individual differences in the extent to which standard shifts occur. As we indicated in our earlier work (Biernat et al., 1991), the standard shift phenomenon should occur only when people hold differing beliefs (stereotypes) about contrasting social groups. For example, we found no evidence that subjects held different gender stereotypes regarding age or frequency of movie attendance, and, as anticipated, subjective and objective response scales yielded similar patterns of judgment. A corollary to this finding is that individuals who do not personally believe in systematic group differences, even when these beliefs are commonly held by others, should also be less susceptible to standard shift effects. In other words, individuals who do endorse stereotypes should be the most likely to produce the patterns of judgment we have previously described. On objective measures, these respondents should show the full effect of their stereotypes; on subjective rating scales, by contrast, the respondents' stereotypes should lead to the use of different standards, which should, in turn, reduce or reverse the stereotype effect. Respondents who do not accept differential group stereotypes should use similar subjective endpoints when evaluating individuals from disparate groups (e.g., men vs. women). For these subjects, the judgment patterns observed should not be affected by the type of response scale on which ratings are made. In the studies presented here, we examine the effects of both attitudes (e.g., racism, and attitudes toward women) and stereotypes (base rate beliefs about groups) on the judgment patterns generated when respondents express their mental representations using subjective versus objective (common-rule) scales.

Our final objective in this research is to provide more direct evidence that shifting standards are, in fact, responsible for the differing judgment patterns we have observed in our earlier work. Subjective and objective scales may differ in many other ways beyond their (alleged) susceptibility to standard shifts. For example, objective scales may prompt raters to attempt accuracy in their judgments (and therefore to rely on base rate beliefs), whereas subjective scales may prompt less careful attention to stimulus details (and more moderate responding). In our earlier work, we ruled out the possibility that the relative difficulty of making judgments in objective versus subjective units was responsible for the judgment patterns we observed, and we found that objective ratings were more likely than subjective ratings to be unbiased and to provide more accurate readings of our subjects' mental representations (Biernat et al., 1991). For example, objective height ratings more closely matched what we assume is a direct index of mental representation—paired comparison judgments.

This research provides further, direct support for the shifting standards account. In Study 3, we demonstrate that the explicit manipulation of comparison standards for making subjective judgments produces differential rating patterns. That is, the pattern of shifting standards can be obtained by directly manipulating standards, without relying on the subjective–objective response scale distinction. Furthermore, in Study 4, we illustrate that individuals use different decision thresholds in determining whether a behavior is diagnostic of stereotypical traits for female versus male targets. If different standards are re-

cruited when judging individual members of different social categories with respect to stereotyped attributes, those standards should lead people to use different "decision rules" for determining the presence or absence of the attributes (see Dunning & Cohen, 1992; Foddy & Smithson, 1989; Foschi, 1992). For example, because most people believe that men are more aggressive than women, they should have a lower threshold for labeling a behavior aggressive when it is committed by a woman rather than a man. A behavior that might be regarded as normal or average for a man might thus be considered an indication of aggressiveness when enacted by a woman. Evidence that aggressive behaviors are differentially diagnostic of aggressiveness in male and female targets points directly to the importance of shifting standards in social judgment tasks.

Study 1

A classic article by Goldberg (1968) provided evidence that women (as well as men) are prejudiced against women. In Goldberg's work, female subjects were asked to read and evaluate a series of articles that were attributed to either male (e.g., "John T. McKay") or female (e.g., "Joan T. McKay") authors. Subjects tended to judge an article more positively on attributes such as competence and quality if it appeared that it was written by a man rather than a woman. This research prompted a great deal of speculation concerning women's tendency to "self-stereotype" (e.g., Cash & Trimer, 1984; Ruble & Ruble, 1982) and inspired a flurry of research seeking to replicate and better understand the effect.

The abundance of this research warranted the 1989 publication of a meta-analytic review of experimental work using the Goldberg evaluation paradigm (Swim, Borgida, Maruyama, & Myers, 1989). In this review of gender effects on evaluations, the authors reported an overall effect size (d) of only $-.07$ (i.e., men were evaluated slightly more positively than women). In other words, "the size of the difference in ratings between female and male target persons was extremely small" (Swim et al., 1989, p. 419). These authors also pointed out that even in Goldberg's original study, the pro-male bias was found on only some of the dependent measures and that significant findings were more likely to be obtained when the topics of the evaluated articles were "masculine" (e.g., law and city planning) as opposed to feminine or neutral. In their meta-analysis, the effect size was slightly larger ($d = -.12$) for masculine than for feminine ($d = -.01$) stimulus materials, although neutral materials produced the largest effect ($d = -.13$).

Swim et al. (1989) identified a number of possible moderators of the small but heterogeneous effect size (e.g., amount of information provided about the target and type of stimulus material), but they commented that the factors they considered "do not fully account for this variability" (p. 420). They called for further research to identify other potential moderators; the present study represents one attempt to do so. We suggest that the type of response scale (i.e., objective or subjective) on which evaluations are gathered may be an important determinant of the size of the gender bias effect. If subjects rely on a global stereotype that men are more competent than women, we should find that objective judgments reveal this bias, whereas subjective judgments do not. The Swim et al. findings, however, lead us to anticipate that something other than a straightforward pro-male bias operates when subjects judge the quality of magazine articles. What is more likely is that gender stereotypes regarding authorship operate such that subjects believe men are better writers of masculine articles (e.g., men know more about fishing than do women), whereas women are better writers of feminine articles (e.g., women know more about nutrition than do men). If this is so, we should find that objective judgments reveal a pro-male bias on masculine articles, but a pro-female bias on feminine articles; subjective judgments should reveal diminished effects, or reversals, because judges may implicitly use a higher (more demanding) standard when assessing an article that they expect to be very good (e.g., a woman, rather than a man, writing about cosmetics).

Method

Subjects were 169 University of Florida undergraduates (107 women and 62 men) enrolled in introductory psychology courses who participated in return for course credit. Subjects simply read a one-page excerpt of a magazine article attributed to either "Joan T. McKay" or "John T. McKay" and were asked to evaluate the article on three dimensions: quality ("How good an article would you say this is?"), monetary worth ("As a magazine editor, how much money would you be willing to pay the author for his/her article?"), and interest ("Do you think the magazine's readers will find this article interesting?"). Ratings were made on either subjective or objective response scales. The subjective measures were 9-point scales with endpoints labeled *excellent* and *terrible* for the quality question, *very little money* and *lots of money* for the monetary worth question, and *no, not at all* and *yes, very much so* for the interest question. Subjects in the objective condition rated quality by assigning a letter grade to the article (A+ through E),[1] monetary worth by providing a dollar figure (constrained to lie between \$50 and \$1,000), and interest by indicating the percentage of the magazine's readers who would find the excerpt interesting. After evaluating the article, subjects completed the 24-item Attitudes Toward Women Scale (AWS; Spence & Helmreich, 1972).

The excerpted articles that subjects read had actually been published in mass circulation magazines and were selected on the basis of the apparent sex typing of their content. Two articles each were chosen to reflect masculine, feminine, and gender-neutral topics. The two masculine articles concerned bass fishing and salaries of professional baseball players; the feminine articles featured hints on cooking nutritious meals and trends in eye makeup; and the "gender-neutral" articles concerned the mind–body problem as applied to health issues and a debate about whether people could be classified into dichotomous types (e.g., optimist–pessimist, etc.). Pretesting indicated that subjects did indeed perceive the masculine articles to be more masculine than the feminine

[1] We are conceptualizing the assignment of letter grades as an objective response scale because grades fit our implicit criteria of (a) being externally anchored and (b) suffering no change in meaning dependent on whom the grade is describing (i.e., an A is an A regardless of various attributes of the student who obtains the A). Nonetheless, we acknowledge potential criticism that grades are actually very subjective and unreliable in nature. They do, however, invite an objective (external) perspective in which all targets are evaluated with respect to a common standard. By contrast, our natural language habits may lead us to use subjective scales in such a fashion as to accommodate our expectations (stereotypes). To determine what our subject population believed about grades, we simply asked a separate sample of 23 undergraduates whether they thought letter grades received in school were subjective or objective in nature. The vast majority of these ($n = 21$) perceived them as objective.

articles, and vice versa, and the neutral articles to fall between the other two types on a masculine–feminine rating dimension. This pretesting also indicated that the masculine and feminine articles were perceived to be equal in quality ("overall, how good an article would you say this is?"), although both of these types were rated less positively than the neutral articles. Because preliminary analyses indicated no differential effects on evaluation of the specific articles of each type, the six articles were collapsed into the three general categories of feminine, masculine, and neutral.

In sum, the study was based on a 2 (sex of author) × 3 (type of article—feminine, masculine, and neutral) × 2 (response scale—objective and subjective) between-subjects design. Some analyses also included sex of subject as an additional factor, but this was not significant as a main or interactive effect and is thus not discussed further. To render the data comparable across the two response scale conditions (subjective and objective), ratings were appropriately reverse-coded when necessary, then standardized separately within each condition (subjective and objective) before creating a scale based on the average of the three items. Coefficient alpha on the three-item scale was .62 for subjects in the objective judgment condition and .55 for subjects in the subjective condition.

Results and Discussion

Evidence of standard shifts. The data were analyzed using a 2 (sex of author: Joan or John) × 3 (topic of article: masculine, feminine, or neutral) × 2 (response scale: objective or subjective) between-subjects analysis of variance (ANOVA). The only significant findings were a main effect of topic, $F(2, 157) = 4.31$, $p < .02$, such that neutral topics were evaluated more positively overall ($M = .17$) than either feminine ($M = -.22$) or masculine ($M = -.10$) topics, and the three-way interaction (Author × Topic × Response Scale), $F(2, 157) = 3.21$, $p < .05$. This interaction, broken down by article type, is shown in Figure 1. For feminine articles (Panel A), the Author × Response Scale interaction was significant, $F(1, 58) = 6.81$, $p < .02$. Simple effects tests indicated that Joan was rated significantly more positively than John in objective units, but John and Joan were rated similarly in subjective units.[2] For the masculine articles (Panel B), the two-way interaction did not meet the conventional level of significance, $F(1, 56) = 3.89$, $p < .15$, but the observed pattern is obviously very similar to the pattern for feminine articles; that is, John was rated more positively than Joan in objective units, but the two did not differ on the subjective ratings. For neutral articles (Panel C), the Author × Response Scale interaction was not significant ($F < 1$); none of the means significantly differed from the others (all $ps > .30$). The general pattern is that for sex-typed articles, subjects' objective ratings indicated a favoritism toward the author of the "corresponding" sex—John was better at writing masculine articles, Joan at writing feminine articles. The subjective evaluations did not reveal these biases.[3] This pattern of effects also appeared when we separately analyzed each of the three ratings that made up the evaluative index, although in the case of the interest rating, the three-way interaction was only marginally significant ($p < .11$).

Individual differences in standard shifts. To discover whether subjects' scores on the AWS affected their evaluations of the articles, we first performed a median split on these scores. The range of possible scores on the AWS is 25–100; the range in the present sample was 25–73 ($M = 41.63$, $SD = 8.99$). Subjects with scores of 39 or below were classified as relatively "nontraditional" ($n = 84$), and those with scores greater than 39 were classified as "very traditional" ($n = 83$). We then added this factor in an analysis that also included sex of author, topic, and response scale as described above. In this analysis, the AWS categorization significantly interacted with author, $F(1, 143) = 7.33$, $p < .01$, such that highly traditional subjects rated John ($M = .13$) significantly higher than Joan ($M = -.24$), whereas less traditional subjects did not differentiate between John ($M = .13$) and Joan ($M = .13$). AWS classification was also involved in a significant three-way interaction that included topic and response scale, $F(2, 143) = 4.91$, $p < .01$. The relevant means appear in Table 1. Among low-traditional subjects making objective ratings, feminine articles were viewed significantly more positively than masculine articles; among highly traditional subjects, the opposite was true—objective ratings indicated that masculine articles were viewed more positively than feminine articles. In other words, the objective ratings revealed judgment patterns consistent with the attitude profile (more value placed on the masculine for high traditionals and more value on the feminine for low traditionals). The subjective ratings, however, revealed the opposite patterns for both high- and low-traditional subjects. Subjectively, low traditionals preferred masculine to feminine articles, whereas high traditionals preferred feminine to masculine articles. Once again, the neutral articles were generally rated more positively overall. For reasons that remain unclear, low traditionals rated neutral articles more positively on subjective than on objective response scales, whereas high traditionals rated neutral topics more positively on objective than on subjective scales.

Although the four-way interaction between author, topic, response scale, and AWS classification was not significant, $F(2, 143) = 1.43$, $p > .20$, we took the liberty of recalculating the Author × Topic × Response Scale ANOVA separately for low- and high-gender-traditional subjects. These data must be interpreted with the caution that this exploratory analysis calls for. Among highly traditional subjects, the only significant effect was the Topic × Response Scale interaction, $F(2, 71) = 3.46$, $p < .05$, which we have previously described (see Table 1). For low gender-traditional subjects, the interaction between author and response scale was significant, $F(1, 72) = 4.92$, $p < .03$. This interaction indicated that Joan ($M = .38$) was rated higher than John ($M = -.28$) in objective units; however, in subjective ratings, the difference between Joan ($M = -.003$) and John ($M = .07$) was nonsignificant. This effect was subsumed, however, by a significant Author × Topic × Response Scale interaction, $F(2, 72) = 4.65$, $p < .02$. The pattern just described was most marked when the topic was feminine in nature, was reduced but in the same direction when the topic was neutral, and was reversed (although nonsignificantly) when the topic was masculine. In other words, subjects who scored low on the AWS (a) rated Joan

[2] All post hoc tests are simple effects tests (t tests) comparing the indicated means, but using the overall mean squared error term (within) from the multiway ANOVA (see Klockars & Sax, 1986). Throughout this article, when means are reported as being significantly different from each other, this indicates that the t test was significant at $p < .05$.

[3] We also decomposed this interaction by examining the Author × Topic interaction separately for objective and subjective ratings. For objective ratings, this interaction was significant, $F(2, 80) = 4.32$, $p < .02$; for subjective ratings, it was not ($F < 1$).

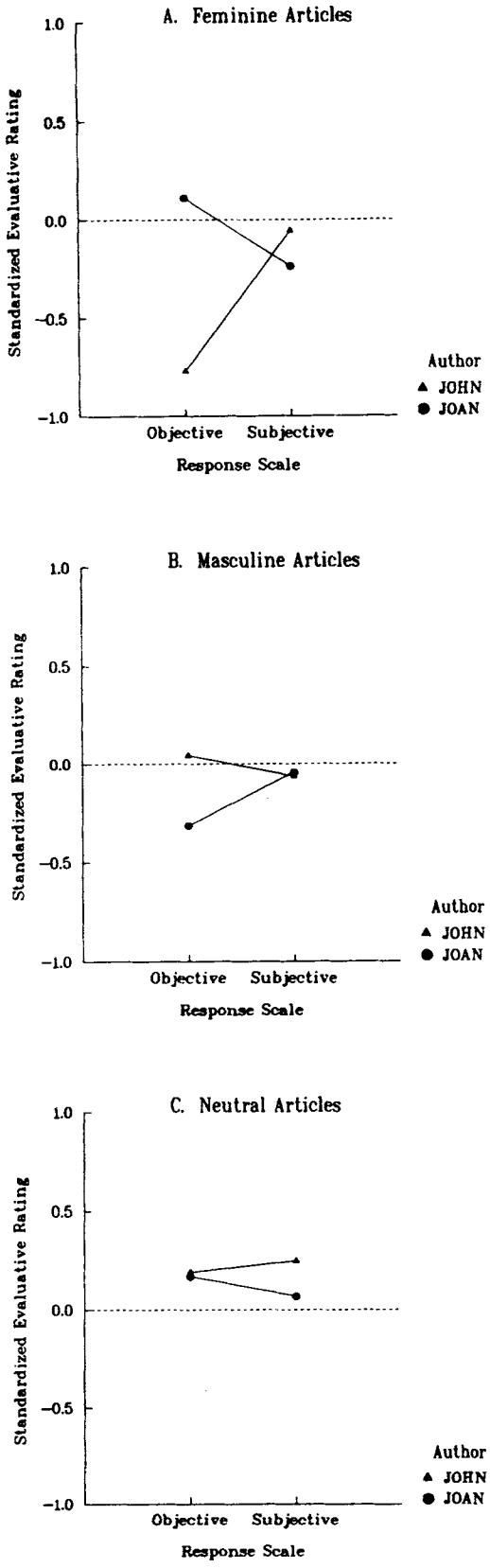

Figure 1. Interaction among author, topic, and response scale in judgments of quality of written articles, Study 1.

more highly than John in objective units when the topic was feminine or neutral, (b) rated John more highly than Joan in objective units when the topic was masculine, and (c) rated Joan and John the same in subjective units on each of the three article types.[4]

It appears, then, that attitudes toward women did influence subjects' judgments to some extent. The most clear-cut finding was that objective ratings revealed biases consistent with subjects' attitude profiles: High traditionals preferred masculine to feminine articles; low traditionals preferred feminine to masculine articles. Consistent with the standard shift account, the subjective ratings showed reversals of these patterns. Our more exploratory analyses also suggested that whereas both high- and low-gender-traditional subjects were affected by a standard shift, the shift took a slightly different form for the two groups. For less traditional subjects, a pro-female author bias was apparent in the objective ratings; for highly traditional subjects, a pro-masculine topic bias appeared in the objective ratings. For both groups, subjective ratings failed to reveal these biases. Thus, we have uncovered some evidence for individual differences in the standard shift effect. Our failure to document a simple male favoritism bias in this study and in others inspired by the original Goldberg (1968) research may perhaps reflect the operation of competing influences; that is, the pro-male "competence" stereotype that was the focus of this research may be offset by a pro-female "verbal ability" stereotype—a bias we examine in Study 2.

Study 2

In Study 2, we demonstrate the shifting standards phenomenon in a substantively different judgment domain. The focus in this case is on judgments of verbal ability. We chose this dimension because it allows us to investigate two social stereotypes simultaneously: the stereotype that women are more verbally able than men, and the stereotype that Whites are more verbally able than Blacks. We predict that objective judgments will, once again, be more likely to reveal evidence of these stereotypes than will subjective judgments. This is the first case in which we have investigated a positive stereotype of the generally disadvantaged group (i.e., women are perceived as more verbally able than men). Demonstration of the standard shift phenomenon in this case will be particularly useful in noting that the effect may generalize to other stereotyped beliefs, regardless of their valence. This study also further examines the influence of individual-difference factors on judgment patterns.

Method

Subjects were 143 White undergraduates at the University of Florida (67 men and 76 women) who received credit in their introductory psychology courses for participating. Sex of subject did not affect judgments in any way, and therefore this variable is not further discussed. On entering the lab, subjects were told the study concerned "social per-

[4] We should also note that when we included AWS scores as covariates in the basic Author × Topic × Response Scale ANOVA, the effect of the covariate was significant, $F(1, 154) = 4.40$, $p < .05$, but the substantive results reported in Figure 1 did not change.

Table 1
Interaction Among Article Topic, Response Scale, and Attitudes Toward Women Scale (AWS) Classification: Study 1

	Low gender-traditionals		High gender-traditionals	
Article topic	Objective scale	Subjective scale	Objective scale	Subjective scale
Masculine	−.278	.019	−.015	−.157
Feminine	.012	−.268	−.545	−.047
Neutral	.193	.499	.173	−.172

ception," and they were handed booklets that began with the following instructions:

> On the following pages, you will find a series of graduation photographs taken at several southern high schools. Prior to graduation, to help evaluate the overall school system of the state, each of these students took part in a systematic educational appraisal that included an oral vocabulary test. Along with each photo you will find two vocabulary definitions that the student in question produced as part of the educational appraisal process. Using this modest pool of information, we would like you to indicate your best judgment as to each student's popularity, maturity, and verbal ability. We realize that this is a difficult task and that you don't really have much to go on; just do the best you can.

The booklet contained 40 photographs, each paired with two word definitions that had supposedly been provided by the pictured individuals. The photo set consisted of 10 Black men, 10 Black women, 10 White men, and 10 White women, whose pictures had been chosen from several nonlocal high school yearbooks. Photocopied reproductions of these photos were used; they were always readily identifiable in regard to race and sex.

The word definitions that appeared with the photos were selected from a set identified by Arnhoff (1953) and supplemented by Fein (1989). These definitions varied in the degree to which they showed evidence of "thought disturbance." In other words, the definitions ranged from the very straightforward (evidence of high verbal ability) to the rather bizarre and confused (evidence of low verbal ability). This manipulation was included in an attempt to pinpoint the influence of the standard shift phenomenon (i.e., does standard shifting affect judgments at all levels of the attribute of interest?). Fein also collected normative data concerning the degree of "thought disturbance" conveyed by each definition as rated on a 9-point scale. We rank ordered the 233 definitions according to these normative data, then divided the set into 10 discrete "levels" of thought disturbance. Level 1 definitions (least disturbed) were rated from 1.08 to 1.77 on the thought disturbance scale, Level 2 definitions were rated from 2.15 to 2.46, Level 3 from 2.69 to 2.92, Level 4 from 3.23 to 3.46, Level 5 from 3.69 to 3.85, Level 6 from 4.08 to 4.31, Level 7 from 4.62 to 4.85, Level 8 from 5.15 to 5.39, Level 9 from 6.23 to 6.69, and Level 10 (most disturbed) from 7.31 to 8.46. For each race and sex combination, one target was attributed definitions from each level of disturbance. That is, each booklet contained a Black woman, Black man, White woman, and White man who gave definitions from each of the 10 levels of thought disturbance (total = 40 targets). Each target was pictured along with two definitions from a given disturbance level; the same definition was never used twice. Examples of definitions from each of the 10 levels are provided in Table 2. In the reported analyses, we converted the 10 disturbance levels into three categories: low, medium, and high disturbance. The low disturbance category consisted of definitions from levels 1–3, medium disturbance included definitions from levels 4–7, and high disturbance included definitions from levels 8–10.

To control for idiosyncratic effects of particular photo-definition combinations, a second version of the booklet was created such that a pair of definitions attributed to a Black in Booklet 1 was attributed to a White in Booklet 2, always keeping the sex constant (i.e., definitions were switched for White men and Black men across booklets, and for White women and Black women). Two different semirandom orders of the booklet were also created, with the stipulation that no more than three targets of the same race were ever depicted in sequence. Booklet type and order did not affect the results in any meaningful way and therefore are not discussed further.

Subjects were asked to rate each target on three attributes: popularity, maturity, and verbal ability. The first two questions were essentially used as fillers to draw attention away from our interest in verbal judgments; all subjects made these ratings on 5-point scales with endpoints labeled *very unpopular* and *very popular*, and *very immature* and *very mature*. On the verbal ratings, however, response scale was manipulated as a between-subjects variable. Half the subjects rated the target on a subjective 5-point scale with endpoints labeled *very low verbal ability* and *very high verbal ability*. The other half rated the target objectively by assigning a letter grade (A through E) that reflected his or her verbal ability.

To summarize, the study used a 2 (response scale: objective and subjective) × 2 (sex of target) × 2 (race of target) × 3 (level of thought disturbance of definitions) design in which judgments of verbal ability served as the dependent variable. The first factor was manipulated between subjects; the latter three were manipulated within subjects. At the end of the study, subjects also completed the 7-item Modern Racism Scale (McConahay, Hardee, & Batts, 1981) and the AWS (Spence & Helmreich, 1972) and were asked to indicate the percentage of Whites, Blacks, women, and men whom they thought had "high verbal ability," as additional measures of gender and racial beliefs.

Results and Discussion

Evidence of standard shifts. Because both types of response scales had five response options, we first analyzed the data with-

Table 2
Examples of Definitions From Each "Thought Disturbance" Level, Study 2

Level	Word	Definition
1	Chaos	Confusion, the opposite of order.
2	Join	One group or part attaches to another part.
3	Gamble	Waste money for good excitement.
4	Fur	A decoration covering for the body.
5	Apple	Nourishment for the stomach.
6	Catacomb	Like a cellar or something to keep stiffs.
7	Pewter	Something that don't smell good.
8	Ballast	A definite kind of some dance.
9	Nuisance	Person who never uses his noodle.
10	Camera	Watcher of the skies, lenses and wings and armies.

out standardizing within question type. However, our analyses indicated a large main effect of response scale, $F(1, 141) = 96.63$, $p < .0001$, such that the objective ratings were generally higher than the subjective ratings (a "grade inflation" effect?). We therefore standardized the data within question type (subjective vs. objective) and analyzed the resulting ratings using a Response Scale × Sex of Target × Race of Target × Level of Psychological Disturbance mixed-design ANOVA, with the last three factors producing 12 repeated measures (mean judged verbal ability of low, medium, and highly "disturbed" Black men, Black women, White men, and White women). Results did not substantially differ in analyses using the standardized versus nonstandardized judgments. Each of the three within-subjects factors emerged as a significant main effect: female targets were rated higher in verbal ability than male targets, $F(1, 141) = 4.90$, $p < .03$, White targets were rated higher than Black targets, $F(1, 141) = 43.16$, $p < .0001$, and low disturbed definitional depictions were attributed higher verbal ability ratings than medium and highly disturbed depictions, which were also ordered appropriately, $F(2, 282) = 372.67$, $p < .0001$. This large effect of definition level provides strong evidence for the validity of these definitions as indicators of verbal ability.

The shifting standards argument suggests that stereotyped categories (e.g., race and sex) should interact with type of response scale to affect judgments. Specifically, objective scales should reveal that Whites and women are perceived as higher in verbal ability than Blacks and men, respectively. Subjective scales, however, should show less evidence of bias; indeed, shifting standards may eradicate (or reverse) the expected stereotype effect. These predictions were supported. The interaction between race and response scale was significant, $F(1, 141) = 6.52$, $p < .02$, as was the interaction between gender and response scale, $F(1, 141) = 7.09$, $p < .01$. These interactions are depicted, in turn, in Figures 2 and 3. In Figure 2, the difference between Black and White targets when rated on subjective scales was still significant, but this difference was significantly smaller than the difference between Blacks and Whites in objective rating units.

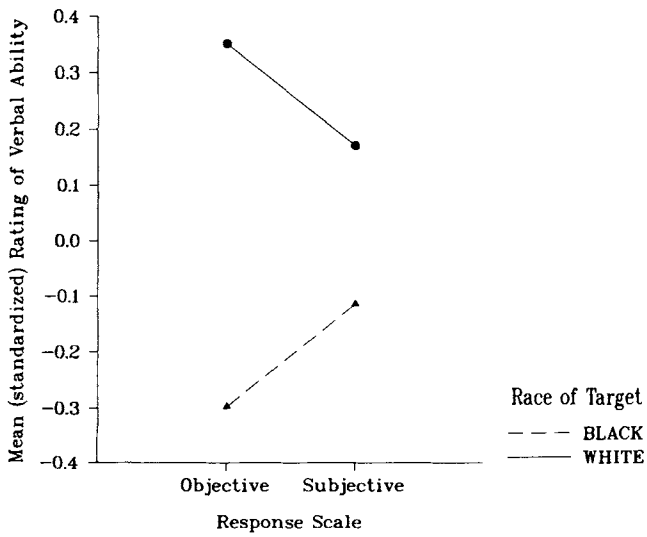

Figure 2. Interaction between race of target and response scale in judgments of verbal ability, Study 2.

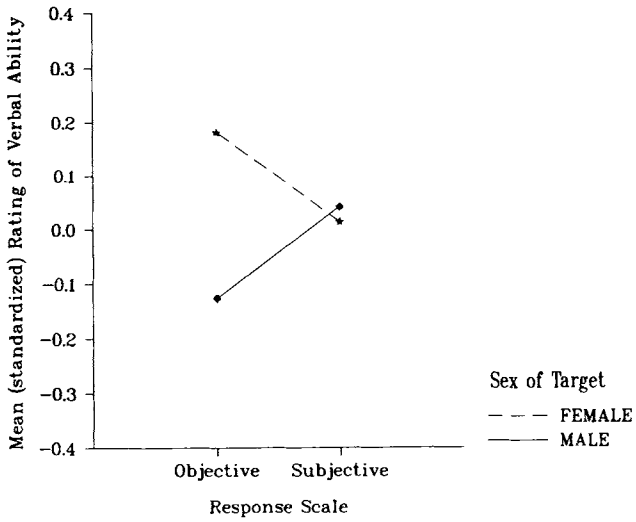

Figure 3. Interaction between sex of target and response scale in judgments of verbal ability, Study 2.

In Figure 3, whereas the objective ratings of men and women were significantly different from each other, the subjective ratings were not.

Gender and race also interacted significantly, $F(1, 141) = 11.00$, $p < .002$. For the White targets, our subjects perceived no difference in verbal ability between women ($M = .23$) and men ($M = .29$), but when the targets were Black, women ($M = -.03$) were rated significantly higher than men ($M = -.38$). The mean ratings of Black men and Black women were each significantly different from every other mean. This was true across response scales (subjective and objective); the three-way interaction between gender, race, and response scale was far from significant, $F < .50$.

The final effect of interest in this analysis was the interaction among race, response scale, and level of definitional disturbance, which was marginally significant, $F(2, 282) = 2.72$, $p < .07$. Simple effects tests indicated that the subjective ratings revealed a significant race difference only at the low level of disturbance, but the objective ratings did so at medium and high levels of disturbance. That is, when targets provided medium or highly disturbed definitions, Whites were rated significantly higher than Blacks on objective scales; these differences were not significant on subjective scales. These data prompt speculation that the standard shift effect may be most marked in situations where the judged phenomenon is particularly striking (in this case, e.g., when the word definitions were clearly disturbed). Of course, these data should be interpreted conservatively, given the marginal significance of the interaction.

Individual differences in standard shifts. One of our goals in this study was to identify individual differences in the standards of evaluation that were used to judge the verbal ability of Blacks and Whites and of women and men. The Modern Racism Scale provided one operationalization of differences in racial standards. We expected that subjects scoring high on this measure were more likely than low scorers to apply different standards to the evaluation of Black versus White targets. Statistically, this would mean that the interaction between race and response

scale would be attenuated for low racists (who presumably do not shift standards when judging these groups) and more striking for high racists (who presumably do).

To examine this possibility, we first divided the sample into high and low racists by performing a median split on Modern Racism Scale scores. Scores on this scale can range from 7 to 35; in this sample, the range was 7–33 ($M = 15.10$, $SD = 5.41$). Subjects who scored 14 or below were classified as *low racists* ($n = 69$), those above a score of 14 were labeled *high racists* ($n = 78$). We then reanalyzed the judgment data as described above but included respondents' racism scores (high vs. low) as another between-subjects variable. The only finding involving this factor was a significant interaction between racism level and race of target, $F(1, 137) = 4.13$, $p < .05$, such that high racists evaluated Whites more positively than Blacks (across both types of response scales), whereas low racists gave comparably low evaluations to both types of targets. The Racism × Race × Response Scale interaction was not significant ($F < 1$). We also used racism as a covariate in the analysis reported above; the covariate effect was not significant ($F < 1$), and its inclusion did not change the results in any substantive way. In general, then, we found no indication that racism level affected judgments of Black and White targets differentially across the two types of response scales. The same was true when we conducted comparable analyses using subjects' base rate estimates of the percentage of Blacks and Whites with "high verbal ability."[5]

To extend this analysis to the issue of different standards based on sex, we examined subjects' AWS scores. None of the various analyses we attempted (including separate ANOVAs for high and low AWS subjects and analysis of covariance) revealed any significant influence of this individual-difference variable. We finally looked to subjects' base-rate estimates of the percentage of women and men with high verbal ability. We first subtracted subjects' female estimates from their male estimates. The resulting mean was −3.25 ($SD = 11.76$); 66 subjects said they believed that women were higher in verbal ability than men, 41 reported no difference between the sexes, and 36 indicated that men were higher in verbal ability than women. We wondered whether individuals with different base rates of this sort would show different patterns of judgment of the targets. Therefore, we recomputed the analysis reported above but added the three-level base rate classification as an additional between-subjects factor. In this analysis, the three-way interaction among sex, response scale, and base rate was of particular theoretical interest; it was marginally significant, $F(2, 137) = 2.46$, $p < .09$. On the basis of this finding, we felt justified in dividing the sample into the three base-rate groups (subjects who indicated women were higher, equal to, or lower than men in verbal ability), and recalculating the Response Scale × Race × Sex × Level of Disturbance analysis within each group.

We were interested in noting two types of effects in each analysis: main effects of target sex and interactions between target sex and response scale. Among subjects who believed, overall, that men were higher in verbal ability than women, neither the main effect of sex nor its interaction with the objective versus subjective response scale were significant ($ps > .25$). Among subjects who believed, overall, that women and men were equal in verbal ability, only a significant main effect of sex was obtained, $F(1, 39) = 4.23$, $p < .05$, with female targets rated higher in verbal ability than male targets. It was only among subjects who believed, overall, that women were higher in verbal ability than men that we found a significant interaction between gender and response scale in the pattern depicted in Figure 3, $F(1, 64) = 12.77$, $p < .001$. This suggests, then, that base-rate beliefs about gender do affect patterns of judgment of individual targets. Only subjects who reported a belief in the cultural stereotype that women are more verbally able than men showed evidence of the shifting standards phenomenon.

Study 3

Study 2 demonstrated the standard shift phenomenon when positive stereotypes of women were operating. This suggests that the effect generalizes to a variety of gender stereotypes, whether they reflect positive or negative views of women. We have not, however, demonstrated that this extends beyond beliefs about gender to include positive beliefs about other generally disadvantaged groups, such as racial groups. To further illustrate the generalizability of the standard shift phenomenon, Study 3 demonstrates it in a context where White subjects view the minority group (Blacks) more positively than the majority group (Whites). We examined the stereotype that Blacks are more athletic than Whites. Our prediction is that when evaluating the athleticism of individual Black and White targets, objective judgments should reveal the full extent of this pro-Black stereotype, whereas subjective judgments, which allow for standard shifts (e.g., "he looks pretty athletic for a White person"), should not. The appearance of this pattern would rule out alternative explanations of the standard shift effect; for example, that it is based on a positive bias toward majority groups.

This study also differs from the previous two in its use of (a) a ranking procedure, which is naturally objective in that it invites use of a common standard by requiring subjects to explicitly order individual targets on the dimension of interest, and (b) a within-subjects design. In this study, subjects are asked to make both subjective ratings and rankings of the same targets. This design provides a more stringent test of the standard shift phenomenon, because subjects are able to directly note (and must directly confront) any inconsistencies in their patterns of ratings across the two types of judgments. If evidence of a standard shift is still obtained, we will have added confidence in the effect.

An additional goal of this study is to demonstrate that directly altering the standards subjects use as they make their subjective ratings causes rating shifts. In our earlier work (Biernat et al., 1991), we found differences in subjective judgments of height when subjects were asked to use the comparison standard "average person" as compared with "average man" (when rating men) or "average woman" (when rating women). Whereas the "average person" ratings resulted in male targets being judged taller than female targets, "average man/woman" ratings

[5] These base-rate data did indicate that our White subjects, on average, believed that Whites are better than Blacks in verbal ability. The mean percentage difference between Whites and Blacks in perceived verbal ability was 16.81; only 12 subjects indicated that Whites and Blacks were equal in verbal ability, and 4 indicated that Blacks were better than Whites. Deleting these latter 16 subjects from the Overall Response Scale × Race × Gender × Level of Psychopathology analysis did not change the pattern of results.

resulted in male and female targets being judged equal in height. A similar logic is used in this study as well. If subjects judge Black and White targets' athleticism with different standards in mind, their subjective ratings should differ such that comparison with "harsh" athletic standards (e.g., "Black men") results in lower ratings than does comparison with relatively "weak" athletic standards (e.g., "women"). At the same time, the use of different standards should not affect our subjects' rankings of individual targets: Explicit orderings of stimuli along a dimension should not be affected by a manipulated standard of comparison.

This point is important because the crux of our argument is that shifting standards account for the differences we have obtained between subjective and objective judgments. Yet, we have little direct evidence that a standard shift is responsible, as the two types of response scales differ on other factors that we may not have considered. If we find that a direct manipulation of standards causes a rating shift within the class of subjective ratings, we advance our argument because the "standard shift" effect can be obtained without relying on the objective–subjective distinction.

Method

Subjects were 44 White undergraduates at the University of Kansas (26 women and 17 men) who participated in exchange for course credit. The title of the project was "Study of Social Perception," and subjects were told that we were interested in "your ability to judge others when you have very little information about them—in this case, the only information you will have about a given individual is his photograph." Subjects worked through a small 10-page booklet. On each page was a photocopied reproduction of a 3.5-in. × 5-in. (8.97-cm × 12.82-cm) photograph of a college undergraduate in a sitting pose (see Biernat et al., 1991; and Nelson, Biernat, & Manis, 1990, for details on these photographs); photographs were also labeled *Person A* through *Person J*. Eight of the photographs depicted White men and two depicted Black men. This unequal race representation was a deliberate attempt to disguise as much as possible the study's concern with race. The photographs were chosen by Monica Biernat on the basis of her subjective impression that the Black target appeared roughly equal in athleticism to the Whites (e.g., similar heights and builds). We did not pretest on this point, however, as we thought it would be odd (and meaningless), given our perspective on shifting standards.[6] Two different orders of the booklets were created; in each case, the Black men filled the fourth and eighth position. No order effects were found in these data.

Manipulation of comparison standards. Subjects were asked to look at each photographed person and judge his athletic ability on subjective (1–7) response scales. As subjects made their subjective ratings of each target, they were invited to use one of five standards of comparison: Black men, White men, men, women, or Americans (manipulated between-subjects). Specifically, the subjective ratings were preceded by instructions to the subject to "Think about the group X. How athletic would you say Person A is compared to all X [e.g., Black men]? Compared to the athleticism of the group X, Person A is closest to:" The 1–7 response scales were labeled at the endpoints by the phrases *the least athletic X* (e.g., Black man) and *the most athletic X* (e.g., Black man). These different standards of comparison were chosen because they vary in the extent to which they are perceived as athletic. We suggest that "Black men" is the harshest athletic standard and should result in relatively lower ratings of both Black and White targets' athleticism. The weakest standards are women and Americans, although we are less sure of which might be perceived as the weakest. On the one hand, women might be perceived as less athletic on average than Americans, but the group of Americans certainly includes women along with other relatively less athletic groups (e.g., the elderly, the physically disabled and children). Of course, the group "Americans" also includes the relatively more athletic groups "males," and "Black males" in particular. We had little way of knowing precisely how people would think about these groups, so we simply suggested them both as relatively weak athletic standards. The groups "White men" and "men" were moderate standards whose ordering, again, was difficult to predict a priori. In general, then, we predicted that judgments of targets relative to "Black men" would result in lower athletic ratings, judgments relative to "men" and "White men" would result in moderate athletic ratings, and judgments relative to "women" and "Americans" would result in the highest athleticism ratings (because of the relative ease of surpassing these standards). This ordering should be true of both Black and White targets.

The final page of the booklet contained the ranking task. Subjects were told

> Now that you've finished rating these individuals by comparing them to group x, the final thing we'd like you to do is rank order the ten people in order of their athletic ability. Next to the letter identifying the person you think is the most athletic, write a "1", . . . ending with a "10" next to the least athletic of the ten individuals.

The purpose of the ranking procedure was to objectify athleticism judgments as much as possible; that is, to avoid a standard shift by inviting direct comparisons between targets. The rank order procedure provided an external (objective) assessment of different targets; an assessment that is assumed to largely bypass the complications that result from shifting standards of evaluation. The reader will note that this study employs a within-subjects manipulation of type of response scale (subjective and ranks) rather than the between-subjects design used in the previous studies.

Individual-difference measures. In the third week of the semester during which this study was run, the majority of the subjects (n = 31) had also participated in a mass-testing procedure during which several measures relevant to this project were obtained. We administered the Modern Racism Scale and asked subjects about their base-rate beliefs regarding the athleticism of Black and White men. Specifically, subjects completed a distributional task of the sort used by Linville and her colleagues (Linville et al., 1986, 1989). Subjects were asked to distribute 100 White men and 100 Black men across five levels of the trait "athleticism." These levels were *very unathletic, somewhat unathletic, neither athletic nor unathletic, somewhat athletic,* and *very athletic.* From these distributions, we calculated both the probability of differentiation (P_d) and the mean perceived athleticism of Black and White men (see Linville et al., 1986, for details on these computations). On average, subjects were more differentiated in their perceptions of the athleticism of White men ($M\ P_d$ = .72) than of Black men ($M\ P_d$ = .70), $t(31) = 2.71$, $p < .02$, and they perceived Black men (M = 3.66) as significantly more athletic than White men (M = 3.36), $t(31) = 4.23$, $p < .001$. In fact, only 2 subjects perceived White men as being more athletic than Black men, 6 perceived no difference, and 23 perceived Black men as more athletic than White men. The main study was conducted from the 11th to the 14th week of the semester; therefore, the individual-difference data were collected between 8 and 11 weeks before subjects' participation in the judgment study.

Results and Discussion

For each type of judgment, we created a variable that indicated the number of times the Black target was rated or ranked

[6] An additional study using this paradigm and a different set of photographs produced similar results, increasing our confidence that the effect can be generalized beyond this set of targets.

as more athletic than the Whites. As there were eight White and two Black targets, this number could range from 0 (neither Black ever rated or ranked more athletic than the Whites) to 16 (each of the two Blacks rated or ranked more athletic than every White). When ties existed (this was only true in the subjective condition), 0.5 was added toward the sum. We then submitted these scores to a Standard (Americans, women, men, White men, and Black men) × Scale Type (subjective rating and ranking) repeated measures ANOVA, in which the latter factor was within-subjects. The effect of standard was nonsignificant ($F < 1$), but the effect of scale type was highly significant, $F(1, 39) = 10.23$, $p < .003$. Black targets were more likely to be viewed as more athletic than the White targets when rankings ($M = 14.14$) rather than subjective ratings ($M = 13.56$) were used. The Scale Type × Standard interaction was marginally significant, $F(4, 39) = 2.21$, $p < .09$. In general, the rankings were more likely than the ratings to show the pattern of Black targets judged more athletic than White targets in every condition except the "White male" standard (in that condition, Ms = 14.06 vs. 14.00 for the ratings and rankings, respectively).

We also analyzed these data by looking more closely at the "ties" between the Black and White targets made in the subjective rating condition. Overall, there were 83 such ties. Our question concerned how these ties were resolved in the ranking procedure: Was the Black target ranked more or less athletic than the White(s) with whom he subjectively "tied?" If our shifting standards premise is correct, these ties should more frequently be resolved by ranking the Black more athletic than the Whites. Of the 83 ties, 56 resulted in the Black target being ranked lower (more athletic) than the White, and in 27 cases the opposite was true. A sign test for matched pairs indicated this difference was significant ($z = 3.07$, $p < .01$). However, some of these 83 ties had been made by the same subjects; specifically, 58 of the ties had been made by subjects with multiple ties, and thus the assumption of independence of observations was violated. To correct for this problem, we looked more closely at the 58 multiple ties. First, we counted a subject only once if he or she resolved his or her multiple ties consistently, and then we recalculated the sign test. This resulted in a total of 68 ties, of which 44 were paired with rankings in the predicted direction—Black more athletic than White—and 24 in the other direction. This difference was also significant ($z = 2.30$, $p < .05$).

As a more conservative test, we then looked at those subjects with multiple ties who resolved their ties in a predominantly consistent manner. This included, for example, subjects with three ties, two of which resulted in rankings in one direction and the third in rankings in the opposite direction. Twenty-four of the remaining 68 ties fit this description; when these ties were cut from the sample under consideration such that a given subject was "counted" only once, 44 ties remained. Of these, 30 were paired with rankings of Black targets as more athletic than White targets, and 14 with the opposite pattern; this too was a significant difference ($z = 2.26$, $p < .05$).

Finally, we dropped all those subjects with multiple ties who were matched with rankings in one direction half the time and in the other direction the other half of the time. Eighteen ties fit this description. We were now left with 26 ties in which a subjects' ties were counted only once. Of these, 21 were resolved such that Black targets were ranked more athletic than Whites and 5 such that Whites were ranked more athletic than Blacks

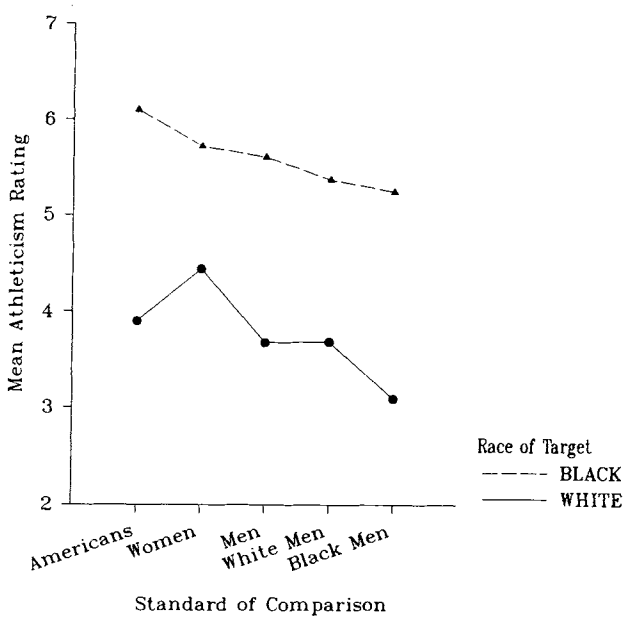

Figure 4. Subjective athleticism ratings of Black and White targets, by standard of comparison, Study 3.

($z = 2.75$, $p < .05$). In sum, in every manner of examining the ties data, the same pattern resulted: When Black and White targets were rated equivalently, Black targets were nonetheless likely to be ranked as more athletic than White targets.

Standard manipulation. Our concern with the effects of explicitly induced standards on athleticism judgments led us to examine the direct ratings and rankings of Black and White targets more closely. First, we entered the mean subjective ratings of the Black (averaged across two targets) and White (averaged across eight targets) stimulus persons as repeated measures in an ANOVA that also included standard of comparison as a between-subjects factor. The two main effects were significant: for race of target, $F(1, 39) = 198.01$, $p < .0001$; for standard, $F(4, 39) = 5.27$, $p < .002$. Overall, Black targets ($M = 5.64$) were rated more athletic than White targets ($M = 3.78$). Although the two-way interaction between race and standard was not significant, $F(4, 39) = 1.72$, $p > .16$, the data are depicted in this form in Figure 4 so as to clearly illustrate the two main effects. In general, the pattern of judgments based on differential standards fit our expectations: The "Black male" standard produced the lowest athleticism ratings for both Black and White targets, the "women" and "Americans" standards produced the highest athleticism judgments, and the "men" and "White male" standards produced moderate judgments. The effect of standard was more striking, however, in subjective judgments of White than Black targets: A one-way (standard) ANOVA on the White judgments was significant, $F(4, 39) = 7.52$, $p < .0001$, whereas the comparable effect of standard on ratings of Black targets was not significant, $F(4, 39) = 1.78$, $p = .15$.

Differing standards should not affect rankings, as these are presumably based on direct relative comparisons across targets. An ordering of targets from most to least athletic should not be affected by differential standard use. To test this point, we also submitted the mean rankings of Black and White targets as re-

peated measures in an ANOVA that included standard of comparison as a between-subjects factor.[7] The main effect of race was very strong, $F(1, 39) = 487.31$, $p < .0001$, with Black targets being ranked lower (more athletic; $M = 2.43$) than White targets ($M = 6.32$). The effect of standard, as predicted, was not significant ($F < 1$), nor was the interaction between race and standard, $F(4, 39) = 1.70$, $p > .16$. Although differing standards clearly affected subjective ratings, they had no influence on rankings. To explicitly test this observation of differential sensitivity to comparison standards between ranking and rating procedures, we conducted an additional analysis in which we converted both the rankings and ratings into z scores (after reverse-coding the rankings) and submitted these to a Race of Target × Standard × Response Scale ANOVA. The only significant effects were the main effect of standard, $F(4, 39) = 4.11$, $p < .01$, and the Race × Standard interaction, $F(4, 39) = 4.28$, $p < .01$. As depicted in Figure 5, standard did not affect rankings but did affect ratings in the predicted directions. This finding increases our confidence in the use of a ranking procedure to tap "common standard" judgments.

Individual differences in standard shifts. Finally, we examined whether individual differences in Modern Racism and base-rate beliefs about the athleticism of Black and White men affected judgment patterns. The probability of differentiation measures, analyzed in a variety of ways, did not affect athleticism judgments and will therefore not be discussed further. We did, however, find some suggestive effects using Modern Racism scores and the mean base-rate perceptions of Black and White men's athleticism.

Scores on the Modern Racism Scale ranged from 7–32 ($M = 15.87$, $Mdn = 16.0$). We performed a median split on these data and added this factor to the Scale (ratings or rankings) × Standard of Comparison (five levels) repeated measures ANOVA described earlier, in which the dependent variables were the number of times (out of 16) Black targets were rated and ranked more athletic than White targets. In this analysis, we omitted the 13 subjects who did not participate in the pretest session, when Modern Racism was measured. In this smaller sample of subjects, we continue to find the results described earlier, along with a significant interaction between scale type and Modern Racism categorization, $F(1, 21) = 9.65$, $p < .006$. This interaction is depicted in Figure 6. Subjects scoring high in racism were the most likely to show evidence of the shifting standards phenomenon described above: They judged Black targets as more athletic than White targets significantly more often in rankings than in ratings. Subjects scoring low in racism did not differentially use the response scales.

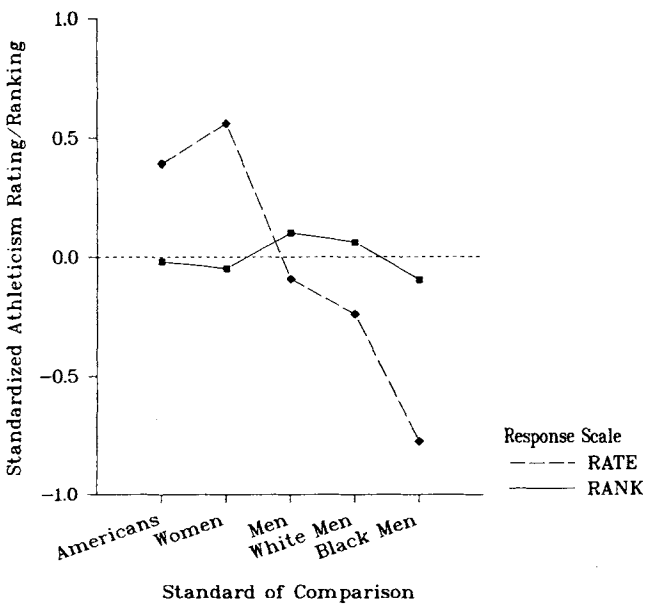

Figure 5. Interaction between standard of comparison and response scale on standardized athleticism judgments, Study 3.

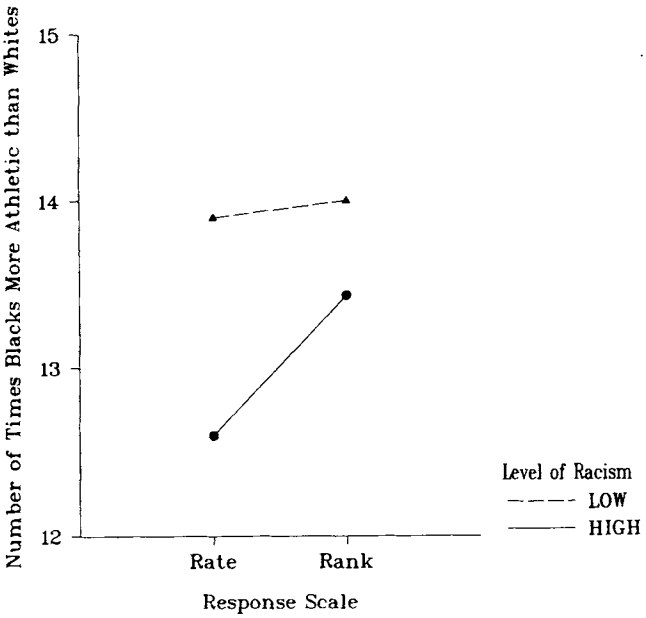

Figure 6. Interaction between racism level and response scale (rankings and ratings) on number of times Blacks are viewed as more athletic than Whites, Study 3.

A comparable analysis using base-rate perceptions of the mean difference in athleticism between Black and White men (collected during pretesting) was also conducted. The mean perceived difference in athleticism between White and Black men (from the pretest questions) was −.297 (on separate 1–5 scales); $Mdn = -.25$; Blacks were viewed on average as more athletic than Whites. We performed a median split on these perceptions, thus creating a group with a strong tendency to perceive Black men as more athletic than White men, and a group with a weaker tendency to do so (because most subjects believed Blacks were more athletic, we could not create groups who clearly did and did not perceive a Black–White differential in athleticism). This factor was included in the Scale (ratings or rankings) × Standard analysis described above. The previously

[7] We recognize that because ranks are ipsative, the use of the Black and White means as repeated measures is not quite appropriate. When the analysis was repeated as one-way (standard) ANOVAs on the Black and White means separately, we continued to find no effects of standard of comparison.

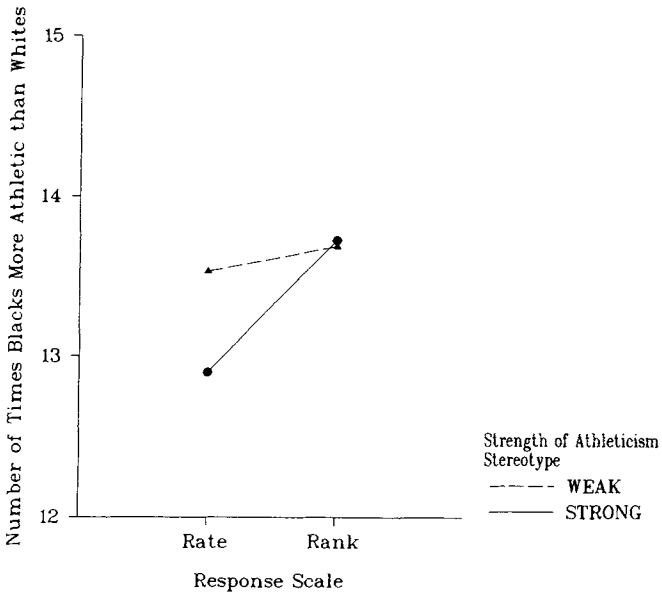

Figure 7. Interaction between base rate beliefs about athleticism and response scale (rankings and ratings) on number of times Blacks are viewed as more athletic than Whites, Study 3.

described main effects of scale and standard remained significant, and the interaction between base-rate perception (Black men much more athletic or Black men modestly more athletic) and scale was marginally significant, $F(1, 21) = 2.50$, $p < .13$. This interaction is depicted in Figure 7. It was among subjects who endorsed the "Black more athletic" belief most strongly that the response scales produced different patterns. For these subjects, Black targets were ranked as more athletic than Whites 13.73 times and rated as more athletic 12.90 times, $t(14) = 2.62$, $p = .02$. For subjects with weak or no endorsement of the stereotype, the corresponding means were 13.69 and 13.53 for ranks and ratings, respectively, $t(15) < 1$. In other words, the strong stereotype endorsers produced a pattern of response similar to that of subjects scoring high in racism. This interaction, however, was only marginally significant and should be interpreted conservatively.[8]

In sum, this study provides additional evidence in favor of the shifting standards hypothesis: Ranked (common-rule) judgments were more likely than subjective ratings to reveal the operation of the Black athletic stereotype. Furthermore, Study 3 supports the corollary regarding individual differences in standard shifts and provides direct evidence that shifts in comparison standards account for rating changes on subjective scales.

Study 4

In this final study, we leave behind the issue of individual differences in standard shifts but extend the previous study's concern for direct evidence of the importance of standards to the judgment patterns we have observed. This study uses a rather different paradigm. Rather than make direct judgments of targets, subjects were asked to keep in mind either a male or female target and to judge the diagnosticity of that target's behaviors for the personality traits aggressive, assertive, and unassertive. We have argued that different standards are recruited for judging members of stereotyped groups; these standards are based on expectations regarding the expected levels of those group members on the dimension of interest. If one group's standard is at a lower level than another's, this should mean that members of the former group can more easily surpass that standard than can members of the latter group; that is, the threshold for "qualifying" for a trait is lower in the former case. Thus, if subjects hold a particular stereotype—for instance, that men are more aggressive than women—this should lead them to have lower thresholds for diagnosing that the attribute (aggressiveness) exists in members of the group presumed to have lower levels of the attribute overall (i.e., women). To take another example, if subjects believe that women are more passive than men, their threshold for diagnosing passivity should be lower for male than for female targets. Evidence to this effect would advance our case that comparison standards are important to a wide variety of judgment settings.

We examine three gender-linked traits in this study: aggressiveness, assertiveness, and unassertiveness (passivity). For aggressiveness and assertiveness, the male standard is expected to be higher than the female standard, but for unassertiveness, the female standard is expected to be higher than the male standard. In each case, the group with the lower standard should produce the higher diagnosticity judgment, as this low standard can be more readily surpassed. We surmised, however, that the assertiveness stereotype is the weakest of the three and that the hypothesized pattern of effects would be least striking in this case. In fact, when Rasinski, Crocker, and Hastie (1985) examined a similar question, using the Locksley et al. (1980, 1982) behaviors, they found no significant difference between male and female targets in the perceived diagnosticity of behaviors for determining assertiveness.

Method

Subjects were 44 female and 31 male undergraduates at the University of Florida who obtained credit toward the experimental participation requirement of their introductory psychology class. Subjects performed three sets of ratings of various behavioral statements adapted from Locksley et al.'s (1980, 1982) work on assertiveness judgments (described below). Specifically, subjects were asked to read about a behavior performed by either "Linda" or "Larry" and to indicate whether that behavior was diagnostic of (a) assertiveness, (b) aggressiveness, and (c) unassertiveness. Twenty behavioral statements were used for each type of rating, and the order in which these ratings was performed was varied across subjects. Because order of rating did not affect judgments, this factor is not discussed further.

[8] The correlation between Modern Racism and White–Black athleticism base rates was $r(n = 31) = .07$, *ns*. However, a chi-square test of independence between the median split categorizations on the two variables was significant, $\chi^2(1, N = 31) = 3.89$, $p < .05$. Ten of the 15 subjects who endorsed the athleticism stereotype most strongly scored low in racism; 11 of the 16 in the other group scored high in racism. Interestingly, however, evidence for differential response scale use was found among subjects who scored high in racism and were strong stereotypers, yet only 5 subjects fell into both categories. Thus, the two individual difference effects we report are based on rather different subsets of subjects.

Before beginning the procedure, subjects read the following (the word *assertive* was substituted with *aggressive* and *unassertive* when appropriate):

> Linda [Larry] is a 25 year old woman [man]. Please think about Linda [Larry]; imagine meeting her [him]. Now imagine that someone has asked you the question, "Is Linda [Larry] assertive?" From what you know so far, it would probably be difficult for you to answer that question. What kind of information would you need to know before you were able to answer "Yes, Linda [Larry] is assertive?" Below is a list of behaviors that Linda [Larry] may have engaged in during the past month. Please read each behavior, and put an "X" next to that behavior if you think it gives you information that Linda [Larry] is an assertive person. In other words, put an "X" next to the behavior if you think that by engaging in the behavior, Linda [Larry] has provided you with evidence that she [he] is an assertive person.

These instructions were repeated before each set of ratings (assertiveness, aggressiveness, and unassertiveness).

The behaviors we used were chosen from those developed by Locksley and her colleagues in their work on base rates (gender categories) versus individuating information as influences on assertiveness judgments (Locksley et al., 1980, 1982). Locksley et al. pretested a set of 85 statements by having 40 subjects rate how passive or assertive each behavior was on a 0 (*passive*) to 10 (*assertive*) scale. The sex of the actor was always unspecified in each behavioral statement. We obtained these pretest ratings from the Locksley team and selected the 20 behaviors that had been rated most passive (range = 1.65–2.74) and most assertive (range = 7.80–8.88) by the set of 40 judges. The 20 passive statements were those our subjects considered when making their unassertiveness judgments, and the 20 assertive statements were used for both the assertiveness and aggressiveness judgments. Examples of the assertive behaviors included "grabbed his/her wallet back from a pickpocket on the bus" and "drew up a petition and persuaded people to sign it." Passive behavioral examples included "bought a worthless product in order to get rid of the salesman" and "was talked into going to see a fairly bad movie for the second time." For each set of ratings, we presented the 20 behaviors in one of two random orders; this ordering factor also did not affect the diagnosticity judgments and therefore will no longer be considered.

In sum, subjects considered either Linda (n = 41) or Larry (n = 34), and rated (a) the diagnosticity of 20 behaviors (Locksley et al.'s most assertive behaviors) for assertiveness, (b) the diagnosticity of those same 20 behaviors for aggressiveness, and (c) the diagnosticity of a different set of 20 behaviors (Locksley et al.'s most passive behaviors) for unassertiveness.

Results and Discussion

For each behavior, we calculated the proportion of subjects who indicated that it was diagnostic of the relevant personality trait for Linda and for Larry. Of 60 comparisons, only 2 indicated that sex of subject had an impact on subjects' judgments. Because these two differences were likely to have occurred by chance only, we do not discuss sex of subject further.

First, we consider aggressiveness judgments. We suggest that because most people believe that men are more aggressive than women, they have a lower threshold for labeling a behavior aggressive when it is committed by a woman rather than a man. That is, the same behavior is more likely to be considered aggressive when enacted by a woman than a man. For each behavior, we determined whether a higher percentage of subjects who read about Linda found the behavior diagnostic of aggressiveness, or whether a higher percentage of subjects who read about Larry did so. In 14 of the 20 cases, Linda's behaviors were more diagnostic of aggressiveness than Larry's behaviors. A sign test indicated that this difference was significant (z = 1.57, $p < .03$, one-tailed). A comparable analysis of the assertiveness judgments was also performed. In this case, only 8 of the 20 behaviors were more likely to be judged as diagnostic of assertiveness for Linda than for Larry. Not surprisingly, the sign test indicated that this difference was not significant (z = .67, *ns*). Finally, we examined judgments of unassertiveness. In this case, because people generally believe that women are more likely than men to be unassertive, we expected subjects to have a lower threshold for diagnosing unassertiveness in men than in women. That is, the same behavior should be more diagnostic of unassertiveness in Larry than in Linda. The data supported this argument: In 16 of 20 cases, the behavior was more likely to be judged diagnostic of Larry's unassertiveness than of Linda's unassertiveness (z = 2.46, $p < .01$).

General Discussion

These studies make four important contributions to our understanding of the shifting standards model and of the stereotyping process more generally. First, the studies replicate our past work on shifting standards (Biernat et al., 1991) and extend those findings to include more meaningful, traditional social stereotypes based on both negative and positive beliefs about relatively disadvantaged groups (e.g., Blacks and women). Second, the studies show how the process of shifting standards may be relevant to a prominent subliterature in the stereotyping field (e.g., the Joan vs. John McKay effect). Third, we provide evidence supporting the individual difference corollary of the shifting standards model; and fourth, we offer more direct evidence concerning the processes that underlie the shifting standards effect.

Our research has been focused primarily on the distinction between objective ("common rule") and subjective assessment procedures, and their stability (vs. instability) when a judge evaluates diverse targets (men vs. women, Blacks vs. Whites). The three experiments relevant to this point yielded clear, consistent results at this level of analysis (see Figures 1–3). Although the various experiments investigated diverse stereotypes and used a variety of methodologies, the results were remarkably uniform: Judgment procedures that invited an objective, or common rule, point of view showed clearer evidence of stereotyping in the assessment of individual targets than did subjective rating procedures. We interpret these results as further support for the view that objective or common rule assessments encourage the judge to rely on a relatively unchanging evaluation standard. As a result, these judgments may reflect the judge's mental representations with reasonable fidelity; they typically indicate that the evaluations of individual targets may be biased (through assimilation) to broadly shared stereotypes regarding the target's membership group.

Subjective ratings, on the other hand, appear to invoke systematic shifts in the judge's frame of reference, in which targets from disparate social groups are evaluated with respect to different standards. The resulting judgments may consequently show only modest evidence (if any) that the judge's evaluations have been systematically affected by the target's group membership. For example, when judges assess individual men and

women on some attribute where substantial group differences might plausibly be expected (e.g., verbal ability), the meaning they attach to the various rating categories appear to shift, depending on the target's gender (see Parducci, 1963, 1965; Postman & Miller, 1945; Volkmann, 1951). That is, in arriving at subjective evaluations (e.g., high vs. low verbal ability), there is a general tendency to compare the target with others from the same group rather than to evaluate successive targets against a common, unchanging set of standards.

Kahneman and Miller (1986) offered a related conception. They contended that judgment is typically based on an active recruitment process that involves imagined alternatives to the target case at hand; the target is evaluated by comparing it with these imagined alternatives. In line with this approach, we believe that the subjective standards against which an individual is evaluated are importantly affected by expectations (imagined alternatives) based on the target's group membership. The recruitment of alternatives explanation is not fully consistent with our model, however, as it does not account for the strong stereotype effects we found on common-rule judgments. We should further note that in most of the cases we have studied, subjective judgments reveal a diminution rather than an elimination of assimilative stereotyping effects. For example, in judging verbal ability in Study 2, subjects using subjective rating scales continued to view White targets as more verbally able than Black targets, although this difference was significantly smaller than that observed when subjects used objective rating scales. That some assimilation to stereotypes continues to emerge on subjective scales suggests that there may be some pooling or merging of standards, or in Kahneman & Miller's terms, recruitment of at least some alternatives from more than one social category.[9] That is, when judging Black and White targets for verbal ability—particularly when these judgments are made successively, as they were in Study 2—a subject may not completely disregard her standards for one race as she judges a member of another race. This idea is also consistent with Higgins and Stangor's (1988) premise that our judgments incorporate the standards we have used at different points in time.

Studies 3 and 4 are important in demonstrating that shifting standards can directly affect judgmental shifts. In Study 3, we explicitly manipulated the standard of judgments subjects were to use as they subjectively rated targets on athleticism. When the standard was harshest (Black male), athleticism ratings of both Black and White targets decreased; when the standard was weakest (women), athleticism ratings increased. Similarly, in Study 4, subjects' diagnosticity judgments indicated that the threshold for a behavior to qualify as aggressive was lower for women than for men, whereas the threshold for a passive action was lower for men than for women. This pattern of results runs counter to expectancy-confirmation models, which predict that behaviors will be interpreted consistently with stereotypes (e.g., "if it's done by a man, it must be aggressive"), but is quite consistent with the shifting standards model. If the reference standard for a group is relatively low, it can more readily be surpassed by members of that group. As a consequence, a behavior that is seen as being only moderately aggressive for a man (a member of the high-standard group) may be seen as very aggressive if enacted by a woman. These data indicate that differential standard use can directly account for differential judgments—the basic premise of our work.

Individual Differences Among Judges

A corollary of the shifting standards model is that the subjective reference norms associated with a given target may vary from one judge to the next; these norms may depend on the judge's acceptance of familiar group stereotypes. We reasoned that subjects who subscribe to divergent stereotypes of the relevant target groups should show clear evidence of standard shifts when they evaluate individual targets from these contrasting social categories. Hence, their subjective assessments might fail to show the sort of stereotype (assimilation) effects that would be revealed in a more stable, objective judgment procedure. Those who reject group stereotypes, on the other hand, may invoke a common standard for their subjective evaluations of men and women; their common rule and subjective assessments would show similar patterns as a consequence.

The present experiments supported this corollary, although there were also some inconsistencies. In Study 2, we measured stereotypes by asking subjects to indicate the percentage of men and women they thought to possess "high verbal ability." Those who believed that women, as a group, exceed men in this regard showed evidence of the standard shift phenomenon: They rated the female targets as superior to the male targets when the judgment procedure invited a common-rule, objective frame of reference, but not when they made judgments in subjective units. This distinctive pattern of results was not observed among the respondents whose base-rate estimates indicated that they saw no difference between the verbal abilities of men and women, nor among those who indicated that men had higher verbal ability. For these subjects, the common rule and subjective rating procedures yielded similar patterns of results. These results suggest that subjects who rejected the stereotype that women have higher verbal ability than men had not shifted their standards when evaluating men versus women.

Despite these hypothesis-supporting results, however, we also found that subjects who denied that women were generally superior in verbal ability (in their base-rate estimates) nonetheless judged the individual female targets, on average, to be more verbally able than the male targets, whether their assessments were made in subjective or objective units. In essence, then, the stated base rates of these subjects proved to be inconsistent with their assessments of individual targets. These unexpected results suggest that our stereotype measure, based on simple base rates, may have provided an inadequate or incomplete measure of respondents' beliefs.

Respondents' attitudes toward women were also implicated in their evaluations of the "authors" in Study 1. Here, we assumed that subjects who endorsed traditional sex role attitudes would be most susceptible to the standard shift phenomenon; they should show clear evidence of a male-superior bias when assessing the individual targets from a common-rule, objective frame of reference, but should appear to evaluate the men and women more comparably when rating them subjectively. This simple pattern was not confirmed, although some related phenomena were observed. Among the respondents with traditional sex role attitudes, authors who wrote about masculine

[9] Other interpretations are also possible: Assimilation to expectation (stereotypes) at the representational level may be a stronger effect than the contrastive standard shifts that we posit.

topics received more favorable objective evaluations than those who wrote about feminine topics (presumably because of the authors' association with "important" masculine topics). The subjective ratings again reduced this difference. Respondents who favored a more contemporary, nontraditional attitude toward women showed a rather different pattern. In the objective judgment task, in keeping with their more "feminist" sentiments, these respondents favored the female author over the male. Their subjective ratings of "John" and "Joan" did not differ, however, presumably because the targets were now evaluated against different, gender-specific standards.

In contrast to these data in support of individual differences in standard use, we were not successful in accounting for individual differences in our respondents' assessments of Blacks and Whites in Study 2. The results here indicated that neither the Modern Racism Scale nor the subjects' base-rate estimates of Black versus White verbal ability were related to their assessments of individual Black and White targets. It is conceivable, however, that our attitude and opinion measures did not provide relevant, valid information, in part because they were collected directly after the judgment task took place. These measures may have failed because White subjects were sensitive to racial issues (made salient by the judgment task) and may have hidden their true beliefs in this important area (see Biernat & Vescio, 1993, Study 3).

In Study 4, where racial attitudes and stereotypes were measured weeks before the judgment task, the results were more consistent with our individual-difference hypothesis. Subjects who scored high on the racism scale, and subjects who endorsed the "Blacks more athletic" stereotype most strongly, were those who showed the most striking evidence of the standard shift effect. Their judgments varied significantly when we compared the rating and ranking tasks; the difference between the perceived athleticism of Black and White targets (Blacks more athletic) was particularly marked in our subjects' rankings (the common-rule scale) as compared with their ratings. We should note two additional features of this study that distinguish it from the others. First, this was the only experiment in which each subject made both subjective and common-rule judgments. We believed that this within-subjects design would provide a more stringent test of the individual-difference hypothesis, because subjects were in a position to directly note (and presumably avoid) any inconsistencies in their judgments across ratings and rankings. It is therefore particularly striking that the effect was obtained here, using two different individual-difference measures (see Figures 6 and 7). Second, of all the stereotypes considered in these studies, the race and athleticism stereotype appeared to be the strongest. All subjects were more likely to rate and rank Blacks as more athletic than Whites. That we nonetheless find effects of individual differences in this judgment domain suggests that both cultural and personal stereotypes guide judgment processes and that the individual-difference approach to understanding the standard shift phenomenon warrants continued attention.

Concluding Comments

The basic pattern of findings in these experiments is clear: When judgments are made with respect to attributes that are associated with widespread stereotypes, objective assessments consistently reveal bias effects based on group membership, whereas subjective ratings will reveal these effects less clearly (or not at all). At a broader level, this work suggests the need for caution in interpreting "bias-free" evaluations at face value, for subjective assessments may not yield a faithful picture of the judge's mental representations. That is, when targets from different social groups are evaluated in the same subjective terms ("he/she is quite assertive"), these targets may nonetheless reflect very different mental representations. These divergent representations may be masked, however, because targets are judged against different subjective norms—norms that are importantly affected by the judge's stereotypes. As others have noted, prejudice in evaluative judgment may take different forms. The most obvious and typical form is that evaluations are assimilated to stereotypes—for example, a woman is judged less competent than a man. However, these data also illuminate a more subtle form of prejudice: Members of different groups may be evaluated against different standards. What is disturbing is that people who show either form of prejudice may feel that they are behaving in a nonprejudiced, egalitarian manner. For example, individuals who use shifting standards might believe they are color-blind because they evaluate Blacks and Whites comparably. Others, who succeed in avoiding the standard shift, may proclaim that they too are color-blind—"Even though I believe that this White target is more competent than the African-American target, at least I am using a common standard." The shifting standards data therefore raise the complicated issue of what constitutes prejudicial evaluation in our culture. Paradoxically, strong stereotypes may often underlie apparently "fair-minded," bias-free judgments, and the evocation of common standards may promote stereotype-consistent judgments.

Everyday speech is chock-full of subjective assertions. For example, we are likely to characterize Carol as being "very tall" rather than refer to her 5'10 height; or we might comment on her "wonderful" writing style, as opposed to the likelihood that she might earn an "A" in a writing class. This raises the important question of how such subjective comments are understood by listeners and whether they are properly "corrected" to take account of the speaker's standards for describing men versus women. After all, a man who is described as being very tall is likely to be taller than a woman who is similarly characterized (see Roberts & Herman, 1986).

In many cases, it is clear that listeners automatically take account of differences in the subjective standards that underlie everyday speech. No one is surprised to hear of a "large frog" that nonetheless fits very comfortably into a "small car." Without thinking, we recognize that the adjectives *large* versus *small* are applied in accordance with different subjective standards, depending on what is being described (frogs vs. cars). It is, however, unclear whether we apply similar cognitive "corrections" when decoding statements that apply to social groups such as men versus women, where different subjective standards might plausibly affect the speaker's descriptive comments. Our ability to take account of the changing subjective standards that are required for everyday speech is apparently far from perfect. Higgins and Lurie (1983) demonstrated a "change of standard" effect, in which subjects apparently remembered the verbal label they had attached to the criminal sentences of a fictitious "Judge Jones" (how harsh or lenient his sentences seemed to be, compared with other judges), but not the comparative context

that led to these characterizations (see also Higgins & Stangor, 1988). By focusing on the evaluative, subjective language, we may lose sight of the original mental representation on which it was based. What this implies for the present research is that while we make subjective judgments using shifting standards, it is the subjective language itself that may be best remembered by ourselves and by others. Thus, the label *good verbal ability* applied to a man and woman may ultimately lead others to accept these two targets as comparable. However, if Julia's verbal ability is described as "good," we should probably infer that her skills in this area are "very, very good," because the speaker's stereotypes may well have led to the use of a very high set of standards when evaluating the verbal abilities of women. This is an intriguing area, which we are now beginning to investigate.

References

Arnhoff, F. H. (1953). *Some factors influencing the unreliability of clinical judgments.* Unpublished doctoral dissertation, Northwestern University, Evanston, IL.

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.

Campbell, D. T., Lewis, N. A., & Hunt, W. A. (1958). Context effects with judgmental language that is absolute, extensive, and extra-experimentally anchored. *Journal of Experimental Psychology, 55,* 220–228.

Cash, T. F., & Trimer, C. A. (1984). Sexism and beautyism in women's evaluations of peer performance. *Sex Roles, 10,* 87–98.

Dunning, D., & Cohen, G. L. (1992). Egocentric definitions of traits and abilities in social judgment. *Journal of Personality and Social Psychology, 63,* 341–355.

Fein, S. (1989). [Normative ratings of psychopathology of word definitions]. Unpublished raw data, University of Michigan, Ann Arbor.

Foddy, M., & Smithson, M. (1989). Fuzzy sets and double standards: Modeling the process of ability inference. In J. Berger, M. Zelditch, Jr., & B. Anderson (Eds.), *Sociological theories in progress: New formulations* (pp. 73–99). Newbury Park, CA: Sage.

Foschi, M. (1992). Gender and double standards for competence. In C. L. Ridgeway (Ed.), *Gender, interaction, and inequality* (pp. 181–207). New York: Springer-Verlag.

Goldberg, P. (1968). Are women prejudiced against women? *Transaction, 5,* 28–30.

Helson, H., & Kozaki, A. (1968). Anchor effects using numerical estimates of simple dot patterns. *Perception and Psychophysics, 4,* 163–164.

Higgins, E. T., & Lurie, L. (1983). Context, categorization and recall: The "change of standard" effect. *Cognitive Psychology, 15,* 525–547.

Higgins, E. T., & Stangor, C. (1988). A "change of standard" perspective on the relations among context, judgment, and memory. *Journal of Personality and Social Psychology, 54,* 181–192.

Kahneman, D., & Miller, D. T. (1986). Norm theory: Comparing reality to its alternatives. *Psychological Review, 93,* 136–153.

Klockars, A. J., & Sax, G. (1986). *Multiple comparisons.* Newbury Park, CA: Sage.

Krantz, D. L., & Campbell, D. T. (1961). Separating perceptual and linguistic effects of context shifts upon absolute judgments. *Journal of Experimental Psychology, 62,* 35–42.

Linville, P. W., Fischer, G. W., & Salovey, P. (1989). Perceived distributions of the characteristics of in-group and out-group members: Empirical evidence and a computer simulation. *Journal of Personality and Social Psychology, 57,* 165–188.

Linville, P. W., Salovey, P., & Fischer, G. W. (1986). Stereotyping and perceived distributions of social characteristics: An application to ingroup–outgroup perception. In J. F. Dovidio & S. L. Gaertner (Eds.), *Prejudice, discrimination, and racism* (pp. 165–208). San Diego, CA: Academic Press.

Locksley, A., Borgida, E., Brekke, N., & Hepburn, C. (1980). Sex stereotypes and social judgment. *Journal of Personality and Social Psychology, 39,* 821–831.

Locksley, A., Hepburn, C., & Ortiz, V. (1982). Social stereotypes and judgments of individuals. *Journal of Experimental Social Psychology, 18,* 23–42.

Manis, M. (1967). Context effects in communication. *Journal of Personality and Social Psychology, 5,* 325–334.

Manis, M. (1971). Context effects in communication. In M. H. Appley (Ed.), *Adaptation-level theory* (pp. 237–255). San Diego, CA: Academic Press.

McConahay, J. B., Hardee, B. B., & Batts, V. (1981). Has racism declined in America? It depends on who is asking and what is asked. *Journal of Conflict Resolution, 25,* 563–579.

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.

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

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

Parducci, A., & Perrett, L. F. (1971). Category rating scales: Effects of relative spacing and frequency stimulus values. *Journal of Experimental Psychology Monograph, 89,* 427–452.

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

Rasinski, K. A., Crocker, J., & Hastie, R. (1985). Another look at sex stereotypes and social judgments: An analysis of the social perceiver's use of subjective probabilities. *Journal of Personality and Social Psychology, 49,* 317–326.

Roberts, J. V., & Herman, C. P. (1986). The psychology of height: An empirical review. In C. P. Herman, M. P. Zanna, & E. T. Higgins (Eds.), *Physical appearance, stigma, and social behavior: The Ontario Symposium* (Vol. 3, pp. 113–140). Hillsdale, NJ: Erlbaum.

Ruble, D. N., & Ruble, T. L. (1982). Sex stereotypes. In A. G. Miller (Ed.), *In the eye of the beholder: Contemporary issues in stereotyping* (pp. 188–251). New York: Praeger.

Spence, J. T., & Helmreich, R. (1972). The Attitudes Toward Women Scale: An objective instrument to measure attitudes toward the rights and roles of women in contemporary society. *JSAS: Catalog of Selected Documents in Psychology, 2,* 66.

Swim, J., Borgida, E., Maruyama, G., & Myers, D. G. (1989). Joan McKay versus John McKay: Do gender stereotypes bias evaluations? *Psychological Bulletin, 105,* 409–429.

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). San Diego, CA: Academic Press.

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.

Received May 19, 1992
Revision received May 13, 1993
Accepted May 13, 1993 ■