Stereotypes and Standards of Judgment

Monica Biernat
University of Florida

Melvin Manis and Thomas E. Nelson
University of Michigan
and Ann Arbor Veterans Administration Medical Center

People routinely adjust their subjective judgment standards as they evaluate members of stereotyped social groups. Such shifts are less likely to occur, however, when judgments are made on stable, "objective" response scales. In 3 studies, subjects judged a series of targets with respect to a number of gender-relevant attributes (e.g., height, weight, and income), using either subjective (Likert-type) or objective response scales (e.g., inches, pounds, and dollars). Objective judgments were consistently influenced by sex stereotypes; subjective judgments were not. Results were also consistent with the expectation that when a judgment attribute is unrelated to gender, male and female targets evoke the same judgment standards. A schematic model of how stereotyped mental representations are expressed on subjectively defined rating scales is presented, and implications for the study of person perception are discussed.

In ordinary language use, words such as *small* and *large* or *short* and *tall* are understood to have different referents, depending on the object being described. For example, one would not be at all surprised to learn that a large dog was smaller than a small car. It is obvious that in using words like small and large, people routinely use different standards as they move from one category of referents (dogs) to another (cars).

Now consider the categories *male* and *female.* Because men differ from women (or at least *seem* to differ) with respect to a number of attributes (e.g., aggressiveness, height, and income), one may also have different referents in mind when one characterizes a man versus a woman as being aggressive, tall, or financially successful. This line of reasoning suggests that verbal assessments (or ratings) of different male and female targets may not provide an accurate reflection of the perceiver's mental representations of those targets. Male and female targets who are characterized in the same terms (e.g., "very aggressive") may nonetheless be perceived to differ systematically, because they are being evaluated with respect to different standards.

Much of the past research on the "loose linkage" between internal representations and overt assessments has focused on the fact that respondents who have had disparate experiences, particularly in the recent past, may label or describe their mental representations in distinctive ways (Eiser, 1971; Eiser & Stroebe, 1972; Gravetter & Lockhead, 1973; Parducci, 1965; Upshaw, 1965, 1969; Volkmann, 1951). The present research points to a related but relatively unexplored problem with the rating scale methodology when it is applied to the area of stereotypes. Here, there is reason to believe that the rater may "shift" the end anchors of his or her rating scale, depending on the type of exemplar that is to be evaluated (see Manis, 1967, 1971). In general, we assume that the end anchors of a scale will be positioned so as to maximize differentiation among the class members. Taking a physical gender stereotype as an example, this would mean that the description "very short" might be reserved for women under 5 feet 2 inches tall, whereas this same statement might be applied to any man who was under, say, 5 feet 6 inches.

In a series of studies by Locksley and her colleagues (Locksley, Borgida, Brekke, & Hepburn, 1980; Locksley, Hepburn, & Ortiz, 1982), subjects judged the assertiveness of male and female targets who were described as engaging in a number of behaviors that were diagnostic of assertiveness (e.g., interrupting a student in class). The key finding was that female and male targets who behaved assertively were judged to be *equally* assertive, suggesting that these subjects had ignored the widespread stereotype that men are more assertive than women. This research has found a prominent place in the literature on the "base rate fallacy"—the tendency of perceivers to underuse base rate or prior probability information when individuating information is provided (Bar-Hillel, 1980; Ginossar & Trope, 1980, 1987; Kahneman & Tversky, 1973; Kreuger & Rothbart, 1988; Manis, Dovalina, Avis, & Cardoze, 1980; Nisbett & Ross, 1980; Zukier & Jennings, 1984).

A rather different account of the Locksley et al. findings also seems plausible, however. Because most subjects believe that as a group, men are more assertive than women (Bem, 1974; Broverman, Vogel, Broverman, Clarkson, & Rosenkrantz, 1972; McKee & Sherriffs, 1957; Ruble & Ruble, 1982; Spence &

This research was supported by a grant from the Veterans Administration.

We gratefully acknowledge the helpful comments of Chris Crandall, Barry Schlenker, Roger Blashfield, and several anonymous reviewers on an earlier draft.

Correspondence concerning this article should be addressed to Monica Biernat, Department of Psychology, 292 Psychology Building, University of Florida, Gainesville, Florida 32611.

Journal of Personality and Social Psychology, 1991, Vol. 60, No. 4, 485-499
 0022-3514/91/$3.00

Helmreich, 1978), they may use different standards when rating the assertiveness of male versus female targets. What is considered "average" assertiveness for a man may be seen as "very" assertive for a woman. When trait assertions are made (e.g., "Beatrice is smart"), the standard of judgment is usually the "average person," and the implication is that Beatrice is smarter than the average person (see Higgins, 1977; Huttenlocher & Higgins, 1971). However, a standard other than the average person is evoked when a target person exhibits an attribute that has its own standard associated with it. For example, because men are expected to be relatively assertive as compared with the average woman, the standard of comparison that is evoked when judging assertiveness may differ depending on the target's sex (see Higgins & Lurie, 1983; Kahneman & Miller, 1986). Two targets, one male and one female, may thus receive the same assertiveness rating not because the respondent is unaffected by sex stereotypes, but because these stereotypes have inadvertently led to the use of very different judgment *standards* for assessing men and women.

This analysis is presented schematically in Figure 1. We start with the assumption that our hypothetical perceiver accepts the stereotypic view that men are more assertive than women. Moreover, we assume that individuals who hold this belief, upon encountering a physical event attributed to a male or female target (in this case, "Ann" and "Andy" each interrupt a student), automatically encode and represent those targets in stereotypical terms. At a representational level, Andy is thought to be more assertive than Ann (because he is male). If subjects use different standards for judging the assertiveness of men and women, however, then the rating category associated with a *very* assertive woman may reflect the same degree of assertiveness that is normally associated with a somewhat *unassertive* man. Nonetheless, given the rather different end anchors for assessing the assertiveness of women and men, *both* targets are assigned a rating of 6. But underlying these overt characterizations is the implicit representation of Andy, a male, as more assertive than Ann.

The discrepancy between "male" and "female" standards is

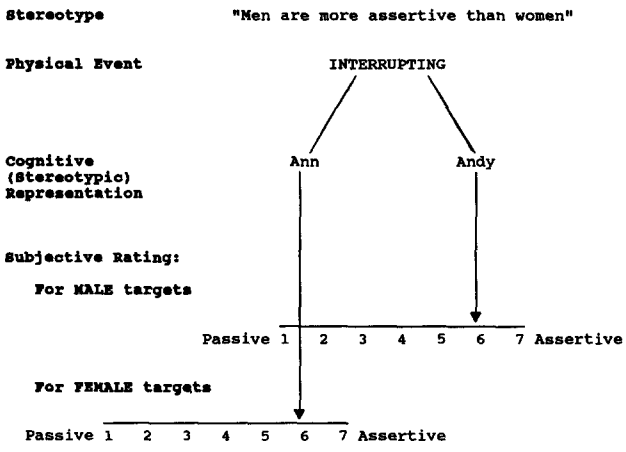

Figure 1. Schematic model of target judgments as mediated by stereotyped representations.

difficult to determine with any precision. It seems clear, nonetheless, that one would obtain *reversed* stereotyping effects, with female targets rated higher (more assertive) than male targets, if the male standard in Figure 1 was displaced to the right or the female standard to the left. Alternatively, higher ratings for men than women would result if the female standard (scale) was shifted to the right or if the male standard was shifted to the left. The point is that when a dependent variable is based on a subjective response scale, one cannot conclude that sex stereotypes are inoperative even when male and female targets receive similar ratings. Unfortunately, unstable subjective scales are frequently used in the person perception literature; hence, one of our main goals in this article is to demonstrate the need for caution in interpreting data of this sort.

This problem is moot, however, if subjects have access to a stably defined response scale, on which measurement units have the same meaning whether applied to men or women. Assertiveness cannot at present be measured on an objective, consensually accepted scale of this sort, but men and women are thought to differ on a number of attributes that *do* have empirical, objective measurement scales. Three such attributes that we consider are height, weight, and income. These are familiar dimensions, and the sex stereotypes associated with them are also well-known: On average, men are taller, weigh more, and earn more money than do women. Moreover, the same units of judgment—feet and inches, pounds, and dollars per year—can be readily applied to both male and female targets without any change of meaning. Referring back to Figure 1, assessments using these "objective" standards should result in judgments that more faithfully reflect the perceived differences between male and female targets.

In previous work, we have used a height judgment paradigm to investigate our respondents' reliance on *stereotypes* as opposed to the *individuating height cues* that photographs provide (Nelson, Biernat, & Manis, 1990). We were attracted to height as a judgment attribute not only because feet and inches are a common metric applicable to both men and women, but also because the stereotype associated with height is accurate and thoroughly grounded in everyday experience. What we have found is that subjects strongly and consistently rely on group stereotypes when judging the heights of individual targets. As a result, other things being equal, male targets are judged to be taller than female targets. In subsequent work, motivational manipulations (including cash rewards for accuracy) failed to eliminate this stereotype effect. Moreover, even when the heights of the various male and female targets were "matched" (so that the men and women were equal in height, on average), and subjects were *informed* of this matching, male targets were nonetheless judged to be significantly taller than female targets. These data were interpreted as evidence of a robust reliance on base rates (or stereotypes) in height estimation. This stereotype effect was particularly strong for judgments of sitting targets, presumably because the height information provided by such a pose is ambiguous, leading to a greater reliance on the relevant base rates.

In designing the present experiments, we wondered whether comparable results would be observed among respondents who used a *subjective* rating scale (e.g., "short" to "tall") in place of

the *objective* height scale (feet and inches) we had used in our earlier work. By the reasoning presented in Figure 1, they should not. If "tall" and "short" have different subjective meanings when applied to women versus men, the stereotype effect we have previously observed might be reduced or completely eliminated when respondents use a subjective (short to tall) rating scale to assess the heights of men and women. For example, a target who appears to be about 5 feet 8 inches might be labeled "tall" if a woman, but "rather short" if a man.

In Study 1, we asked subjects to make a series of height judgments, as in our previous work. One group of subjects judged height on a familiar scale of feet and inches, another group was asked how tall each target was in comparison to the average person, and the final group was invited to judge how tall each target was in comparison to the average man or woman (whichever norm was relevant). The latter two groups responded on subjective 7-point rating scales (*very short* to *very tall*) of the sort depicted in Figure 1. If end anchors are routinely shifted to accommodate the range of values that are typically encountered within a social category, this will normally *reduce* the stereotyping (assimilation) effect that is so often observed when respondents assess individual group members. That is, a subjective rating scale will, according to this reasoning, yield less evidence of stereotyping than a more firmly anchored objective scale, which is presumably fixed and invariant, regardless of the category membership of the targets.

In this study, then, we expected to demonstrate a strong stereotype effect (i.e., men judged taller than women) when judgments were made in feet and inches, and to find a reduction or reversal of this effect—women rated taller than men—when targets were judged in comparison to members of their own sex. Finally, we anticipated that when targets were compared with the average *person,* we would observe a rating pattern somewhere in between the other two; that is, men should be judged somewhat taller than women. We were particularly interested in comparing the "feet and inches" and "average person" conditions of this study, as the latter should reflect people's nonconscious or automatic use of different judgment standards for assessing men and women, *despite* instructions to use a common (average person) standard. The "own sex" condition is used to demonstrate the extreme case in which scales are deliberately adjusted to fit different target groups.

In Study 2, using similar logic, we explored related questions in connection with two other sex-linked judgment dimensions: weight and income. Finally, we expected that subjects would not shift their judgment standards when they evaluated men and women on attributes unrelated to gender. That is, when no sex stereotype regarding an attribute exists, judgments of male and female targets should be comparable, regardless of the type of response scale used. To test this hypothesis, Studies 2 and 3 also included judgment domains where sex stereotypes do not exist.

Study 1

Method

Subjects were 133 undergraduates at the University of Michigan, recruited through newspaper ads and postings around campus. Each of the 50 men and 83 women received $5 for their participation. Subjects completed questionnaires at their own pace while situated in a room with at least one other participant and the experimenter present.

The cover page of the questionnaire contained these instructions:

> We are interested in people's ability to estimate heights of others from photographs. You probably are used to thinking about height—it's an obvious feature we notice about other people. We've taken pictures of students at various points on campus. Your task is simply to look at the photographs on the following pages, and to estimate the heights of the individuals. Your judgment should be an assessment of the height of the person as pictured—wearing the shoes or boots shown.

After indicating their own sex and height,[1] subjects saw a series of 44 pictures, one per page, which were photocopied reproductions of 3.5 × 5 in black and white prints. The set of photos came from a larger sample that we collected by approaching college-aged subjects on campus in a more or less random fashion and asking them to serve as models. Each model posed for at least one sitting and one standing full-length photograph, in which a familiar reference object (a chair, desk, doorway, etc.) was situated nearby. The gender of each model was always readily identifiable. An effort was made to photograph models from different distances and angles, such that sex and posture varied independently of photographic image size. We measured models' heights on the spot, and included their footwear in those measurements.

In our total set of photographs, male heights ranged from 62 to 81 inches and female heights from 59 to 74.5 inches. For this study, however, we chose 22 male and 22 female pictures that had been matched for height. That is, for every 5 foot 9 inch man, we also included a 5 foot 9 inch woman, and so forth. The height range was 63 to 74.5 inches. Subjects were not informed of this matching. Equal numbers of sitting and standing targets of each sex were shown. Because we conceptualized the first four photographs (two male and two female) as practice trials, the final data we report are based on 40 height judgments—10 each of sitting men, standing men, sitting women, and standing women. Three different orderings of the picture stimuli were used; none of the results were affected by order of presentation.

The main manipulation of the study involved the type of scale on which subjects made their height judgments. Forty-two subjects responded in feet and inches, the objective scale we have used in our previous work (Nelson et al., 1990). Forty-seven subjects answered the following question in regard to each photo: "How tall is this person, compared to the average person?" Ratings were made on a 1 (*very short*) to 7 (*very tall*) scale, with only the end points labeled. Finally, using the same 1 to 7 judgment scale, 44 subjects responded to the question "How tall is this man/woman compared to the average man/woman?"

Results

To make the height judgment data comparable across question type, we calculated standardized scores, based on each subject's personal mean and variance in height estimates, across all the pictures that were presented. The key analysis was a repeated measures analysis of variance (ANOVA), with question type (feet and inches, average person, or average for sex) and sex of respondent as between-subject variables, and sex of target and posture (sitting or standing) as within-subject variables. For each respondent, there were four repeated measures based on the average standardized scores associated with their

[1] Subjects' own heights were unrelated to the key judgment variables.

judgments of sitting women, standing women, sitting men, and standing men. Each of these dependent variables was a mean height judgment, averaged across 10 photographed targets.

Of the possible main effects, only sex of target emerged as a significant predictor of height judgments, $F(1, 127) = 108.39$, $p < .0001$. Male targets ($M = .19$) were judged taller than female targets ($M = .18$).[2] This overall effect, which reflected respondents' reliance on stereotypes about the usual heights of men and women, was qualified, however, by two significant two-way interactions. One of these was between sex of target and posture, $F(1, 127) = 64.08$, $p < .0001$, such that the perceived difference in height between men and women was greatest when targets were sitting as opposed to standing. This replicates our previous work (Nelson et al., 1990), in which we argued that reliance on the height base rate is greatest when individuating information is ambiguous (i.e., when targets are sitting). When judging standing targets, respondents are more responsive to individuating height cues.

In answer to the main question of the present study—How do different response scales affect the judgments associated with individual female and male targets?—we also found a significant interaction between sex of target and question type, $F(2, 127) = 63.20$, $p < .0001$. This interaction is depicted in Figure 2, using standardized height judgments as the dependent variable. The pattern of results is striking. For respondents using an objective (feet and inches) scale, we found a strong stereotype effect, with men being judged quite a bit taller than women (mean difference between men and women = .98 in standard score units). Among subjects who judged these same targets in comparison to the average person, this effect was reduced by more than 80% ($p < .001$), although men were still judged to be significantly taller than women (mean difference = .20). Again, this difference is significantly reduced relative to the findings obtained with a feet and inches response scale, but the resulting stereotype effect is still reliable. Finally, when subjects were explicitly invited to judge target heights in comparison to the average man or woman, they rated the sexes as essentially equivalent in height (mean difference = −.01) as, in fact, *was* the case.[3] A comparison between only the two latter conditions in this study—the average person and average for sex conditions—also provides an important test of our model. In these two groups, subjects were explicitly provided with different judgment referents, yet responded using the same subjective scale. A test of the partial interaction between question type (average person vs. average sex) and gender of targets was also significant, $F(1, 79) = 11.31$, $p < .01$, as reflected in Figure 2. This suggests that a shift of standard was effected by question wording. None of the other two- or three-way interactions was significant.

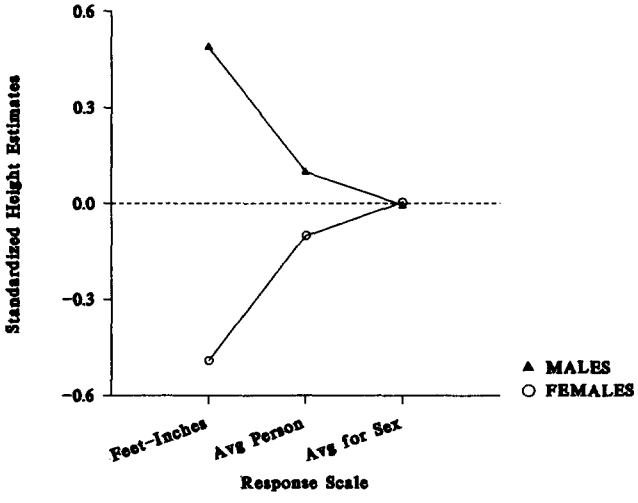

Figure 2. Interaction between sex of target and scale type on height judgments, Study 1.

To demonstrate the overall finding of an interaction between question type and target gender another way, we turned back to the nonstandardized data and calculated the average estimated height of each photograph in feet and inches and in terms of the average person response scale. (These two groups are of most interest because of our concern with objective versus subjective response scales and the inadvertent standard shift based on stereotypes that occurs in the average person condition). Corresponding to each photograph, then, was a mean judgment in feet and inches and a mean judgment on a subjective (1–7) scale, each based on 42 and 47 subjects, respectively. Figure 3 presents a simple scatter plot ($N = 40$ photographs) of these data. What seems to be happening is that a target who is said to be, say, "average" in height (subjective rating = 4.0) is thought to be taller on the objective scale (feet and inches) if the target is a man rather than a woman. Of nine male–female pairs that could be fairly closely matched in subjectively judged height, all nine revealed higher ratings on the objective scale for male than for female targets.[4] A sign test was significant at $p < .01$,[5] supporting the visual interpretation.

[2] A main effect for sex using standardized scores necessarily means that the sum of the female and male means will be zero. Here, .19 + −.18 = +.01 because of rounding error.

[3] Our response scales differ in ways other than objectivity and stability. The most notable other difference concerns the available *range* of response options—much more variation is possible in feet and inches than in 1 to 7 (short to tall) units. To test whether the interaction between scale type and target sex was due to this difference in variability, we converted respondents' judgments in feet and inches to 1 to 7 ratings by cutting the range of these judgments into seven equal-sized categories. For example, height ratings of 61 inches or less were labeled 1, and ratings of 76 inches or more were labeled 7. A second analysis of variance using these new dependent variables revealed the same pattern of interaction—a substantial reduction of the stereotyping effect on the *subjective* 1 to 7 scale as compared with the *objective* 1 to 7 scale. This was true whether we used the 1 to 7 responses in standardized or unstandardized form (both interaction Fs > 90.0, $p < .0001$). This suggests that the interaction between scale type and target sex is a substantive finding rather than a methodological (scale-range) artifact.

[4] Another way to test this pattern using all of the data is to calculate the partial correlation between sex of target and objectively judged height, controlling for subjectively judged height. This partial correlation was .79—male targets were judged taller than female targets in feet and inches, holding constant their perceived heights in subjective rating units.

[5] This pattern, as one might expect, is even more pronounced when feet and inches judgments are compared with the average for sex ratings.

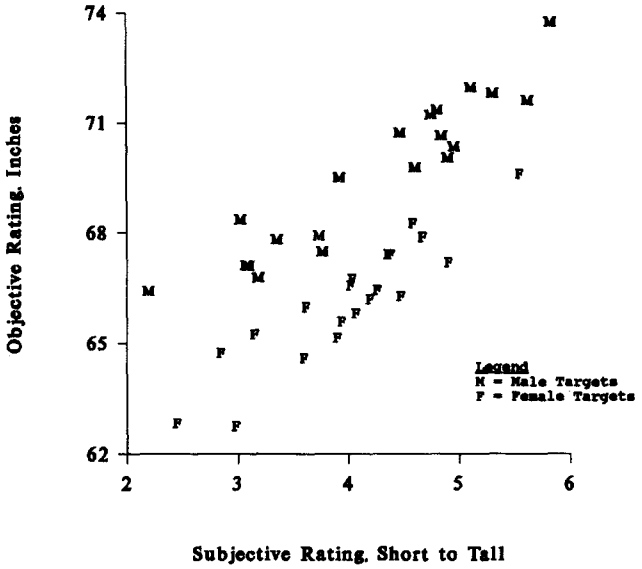

Figure 3. Scatter plot of average female and male height estimates in subjective units versus objective inches, Study 1.

Discussion

These data support our speculation that the application of different subjective judgment standards for assessing women and men may produce what appear to be small gender stereotype effects (or null effects), even when the respondents believe that male and female targets are substantially different with respect to the attribute being judged. We assume that the objective measurement scale, feet and inches, provided a relatively stable set of standards that was largely unaffected by the sex of the different targets. When this more stable scale was used, male targets were judged to be *substantially* taller than female targets. The data from subjects who compared men's and women's heights with the average man or woman ("average for sex") mapped neatly onto the schematic judgment model presented in Figure 1. Subjects believe that men are taller than women. Because of this (accurate) stereotype, and despite the fact that our set of male and female targets were equal in height, the targets were cognitively represented in stereotype-consistent terms (the men were seen as taller than the women, on average). Because we instructed these "average for sex" subjects to use separate rating scales for the two sexes, the divergent stereotype-based representations of the different targets were translated into *equivalent* ratings for the female and male models.

When we asked subjects to compare the targets with the average person, they should not have invoked different height referents for men and women. Nonetheless, our respondents appear to have "inadvertently" done just that, although to a lesser extent than they did in the intentional standard shift ("average for sex") condition. Our data from subjects who used the "average person" response scale correspond to a modified version of Figure 1, in which there is a somewhat reduced difference between the judgment standards applied to male and female targets. This resulted in our male targets being judged taller than the female targets, but not to the extent that they would have been if the respondents' subjective scales had remained *stable* (a condition that we assume to be approximated by the respondents who rendered their judgments in feet and inches).

Study 2 was designed to extend the results of Study 1 to domains other than height. Men and women are perceived to differ on a number of other attributes that can be measured both in objective and subjective ways. We asked subjects to judge male and female targets on two such attributes—weight and income—in addition to height, and on a non-sex-linked feature—age. Because we were most interested in subjects' inadvertent or automatic use of different standards when judging men and women, we compared objective scales—pounds, dollars per year, feet and inches, and years—with subjective comparisons involving the *average person.* We expected to replicate our height findings in all domains of this study except age, which would, we thought, be largely independent of sex in the minds of our respondents.[6]

Study 2

Method

Forty-three University of Michigan undergraduates served as subjects and were paid $5 for their efforts. All subjects were drawn from an introductory social psychology course and were run in a single testing session. Subjects were presented with a series of 44 individual color slides, projected on a large screen, each showing a man's or woman's face and shoulders. The targets ranged in age from 20 to 68 and were photographed at an outdoor market site, where they were recruited at random.[7]

Instructions, which appeared on the cover page of the response booklet, were as follows.

> This is an experiment in social sensitivity. I am going to show you a series of faces of men and women. You will then be asked to rate each person on the following four characteristics: height, age, financial status, and weight.

Presentation of the slides began after subjects read this introduction. Each slide appeared on the screen for approximately 20 s, during which time subjects rated the pictured individual on the four attributes listed above, with one questionnaire page corresponding to each slide. The first four slides served as practice trials and were not used in any of the analyses. Twenty-one of the subjects were given objective response scales—they estimated height in feet and inches, age in years, financial status in dollars earned per year, and weight in pounds, always in that order. The remaining 22 subjects rated the same attributes on subjective, 1 to 7 scales. The end points of each of the scales were labeled as follows: *very short* and *very tall, very young* and *very old, financially very unsuccessful* and *financially very successful,* and *very light* and *very heavy.* At the top of each page of these subjective ratings appeared the following statement: "Each of these ratings are to be made in comparison to the *average adult.*"

---

[6] We recognize that women do have longer life expectancies than men; nonetheless, we do not believe that this has created a stereotype that women in general are older than men.

[7] We gratefully acknowledge Veronica Fiske's provision of these slides.

The actual heights, weights, and incomes of the targets in this study had not been assessed, and thus we cannot compare the resulting judgments with objective reality. This means that differences between the ratings associated with female and male targets cannot confidently be attributed to a simple stereotype effect. We assume, however, that the widespread recognition of sex differences in height, weight, and income will lead most respondents to use different standards when assessing men and women with respect to subjective scales (such as *very short* to *very tall*), even when they are explicitly instructed to use the *average adult* as their basis for comparison.

Results

We conducted equivalent sets of analyses for each of the rated attributes. To make the ratings comparable across question type, standardization of the key variables was performed as described in Study 1 (i.e., we standardized each subject's responses by calculating his or her mean and variance in attribute estimates across all 40 slides). Following standardization, we analyzed the data using a between–within mixed design ANOVA. The two between-subject variables were question type (objective or subjective) and sex of subject; the within-subject variable was sex of target.[8]

Height judgments. Looking first to the height judgment data, we found a main effect of target sex, with men ($M = .68$) judged significantly taller than women ($M = -.65$), $F(1, 39) = 411.61$, $p < .0001$. No attempt had been made to match these targets for height, and thus we assume that the male targets were, in fact, taller than the female targets. It is unclear, however, whether our subjects exaggerated this difference. Furthermore, the photos were not full length and did not provide individuating height cues. In this sense, they are comparable to our sitting photos in Study 1, and they presumably encouraged use of global stereotypes or expectations about the usual heights of women and men.

The only other significant finding in the height data was the interaction between sex of target and question type, $F(1, 39) = 9.34$, $p < .01$. These results are shown in Figure 4. As in Study 1, the judged difference between men and women was greater when assessments were made with respect to an objective (feet and inches) as opposed to a subjective scale. Simple effects tests indicated that the difference between the ratings of male and female targets on the *subjective* scale was still significant, however, a finding also noted in Study 1.

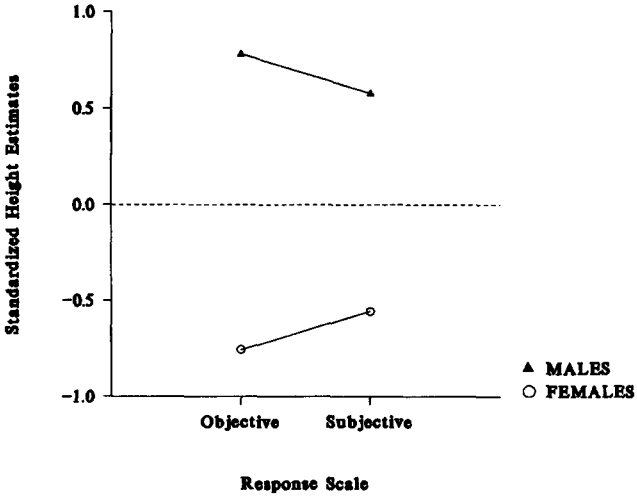

Figure 4. Interaction between sex of target and scale type on height judgments, Study 2.

Weight judgments. Turning now to the other physical judgment, weight, we note the same pattern of findings. Men ($M = .59$) were judged heavier than women ($M = -.57$), $F(1, 39) = 344.55$, $p < .0001$. This effect was qualified, however, by its interaction with question type, $F(1, 39) = 60.87$, $p < .0001$, which is graphically displayed in Figure 5. Again, judgments in "pounds" revealed greater differentiation between women and men, whereas judgments on the "light" to "heavy" scale revealed less, though still significant, sex differentiation.

Financial status judgments. Subjects also judged the financial status or incomes of our targets. In the analysis of these data, the effect of target sex was only marginally significant, $F(1, 39) = 1.97$, $p < .15$, with men ($M = .03$) judged somewhat more financially successful than women ($M = -.03$). The interaction between sex of target and question type, however, was highly significant, $F(1, 39) = 134.17$, $p < .0001$. As can be seen in Figure 6, this pattern of interaction was slightly different from the patterns noted for judgments of height and weight. On the objective scale (dollars per year), men were clearly rated as earning more than women. However, there was reversal of this difference when judgments were made on a subjective scale (*financially unsuccessful* to *financially successful*). Here, surprisingly, our female targets were rated as being *more* successful than the men, even though respondents in the objective scale condition had inferred that they made *less money* than the men. These observations were supported statistically in post hoc simple effects tests.

Age judgments. The final judgment domain was age. Because age is not a sex-linked attribute (i.e., men in general are perceived to be no more or less old than women), we did not expect to find an interaction between target sex and question type, for there is no reason why different subjective standards of judgment should be applied to male and female targets. Consistent with this thinking (and in contrast to our earlier results concerning height, weight, and income—all of which are widely recognized as sex linked), the interaction between sex of target and question type was far from significant ($F < 1$). We did, nonetheless, find a significant effect of sex of target, $F(1, 39) = 5.47$, $p < .05$, with the men (unexpectedly) judged to be older than the women. This was qualified by an equally unexpected interaction with sex of subject, $F(1, 39) = 12.32$, $p < .01$. Age was the only domain in either study in which the subjects' sex contributed significantly to judgments. For reasons that re-

[8] One might raise the objection, noted earlier, that our objective and subjective scales differ in potential variability as well as in objectivity. To address this criticism, we also analyzed these data by converting objective judgment ratings into 1 to 7 ratings, as described in Study 1 (see Footnote 3). As was the case in Study 1, the key interactions between scale type and sex of target were very similar whether the analysis used standardized or unstandardized versions of the 1 to 7 scales, or the standardized version described in the text, as dependent variables.

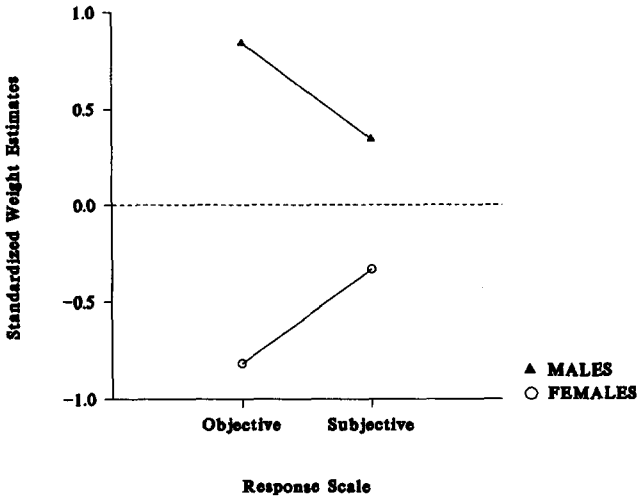

Figure 5. Interaction between sex of target and scale type on weight judgments, Study 2.

main unclear, our male subjects judged the female ($M = .026$) and male targets ($M = -.006$) to be nearly equivalent in age, whereas female subjects judged the female targets ($M = -.07$) to be significantly younger than the male targets ($M = .09$, $p < .01$).

As noted above, these significant results were quite unexpected, and they are difficult to interpret in meaningful theoretical terms. They may, indeed, prove to be nonreplicable. For our present purposes, however, these age data are most useful in demonstrating that in a judgment domain where sex stereotypes do not exist, the subject's assigned response scale (objective vs. subjective) does *not* influence the perceived differences between female and male targets.

Objective versus subjective response scales. In keeping with the analysis strategy used in Study 1, the data from Study 2 were reanalyzed, using the different target slides as our units of analysis. We calculated the mean judgments of height, weight, income, and age for each target individual ($N = 40$) on the objective scales and on the corresponding subjective scales. Figure 7 presents scatter plots of these data, with the mean objective ratings on the ordinate and subjective ratings on the abscissa. The height data (Figure 7a, upper left panel) show a pattern that is similar to the comparable data in Study 1 (see Figure 3). From the data depicted in Figure 7a, we could closely match six male and six female targets in terms of their mean *subjective* height ratings; in all picture cases, judged *objective* height was greater when the target was male as opposed to female, a clear replication of our results from Study 1. Making the same point but using all the data, the partial correlation between sex of target and judged height in feet and inches (removing the effect of subjectively judged height) was .85.

The scatter plot of the mean weight judgments (Figure 7b) vividly demonstrates the clear differentiation between men and women in estimated pounds. The visual and statistical effects are striking—targets who received, say, an "average" mean rating on the subjective scale (4.0) were perceived as weighing much more in *objective* pounds when male than when female. Once again, we calculated the partial correlation between sex of target and weight in pounds, holding constant judged weight in subjective units. This partial correlation was .95, again supporting our claim that subjects invoke different judgment standards when assessing members of different categories, even when instructed to use a common (average adult) judgment standard. To be labeled "heavy," a man must weigh much more than a similarly labeled woman.

On the ratings of financial success (Figure 7c), respondents' judgments were clearly assimilated to their sex stereotypes when they used an objective response scale, whereas their subjective ratings revealed an unexpected pattern of contrast (women were judged, on average, to be more financially successful than men). For example, when a target was judged to be "average" in financial success (subjective rating = 4.0), that target was estimated to have earned more money if male rather than female. Of 12 "pairs" of male–female targets, matched in terms of their mean *subjectively* judged financial status, 11 indicated higher judged income in dollars per year for the male than for the female target. A sign test supports the significance of this effect at $p < .001$. Lending further statistical support, the partial correlation between sex of target and income in dollars, removing the influence of subjectively rated financial success, was .78.

For comparison to the other attributes, a scatter plot of male and female objective and subjective *age* ratings is shown in Figure 7d. Judgment patterns for the male and female targets are equivalent, as one would expect on a feature that is independent of gender. That is, a *single* regression line can effectively represent the relation between subjective and objective assessments, regardless of the gender of the different targets.

Discussion

The results of Study 2 replicate and extend those of Study 1. In both cases, subjective rating scales failed to reveal the full

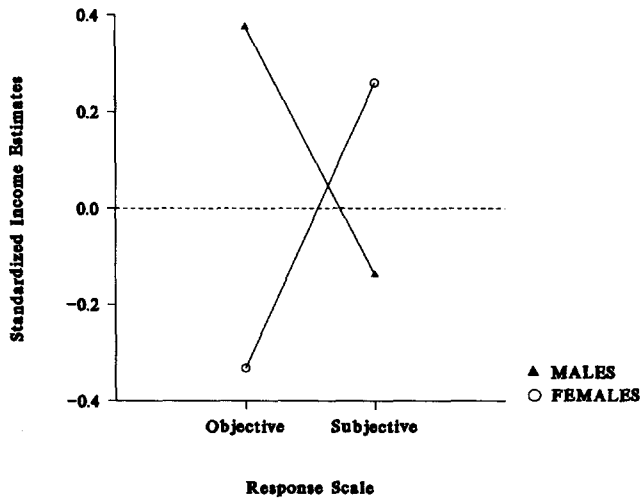

Figure 6. Interaction between sex of target and scale type on financial status judgments, Study 2.

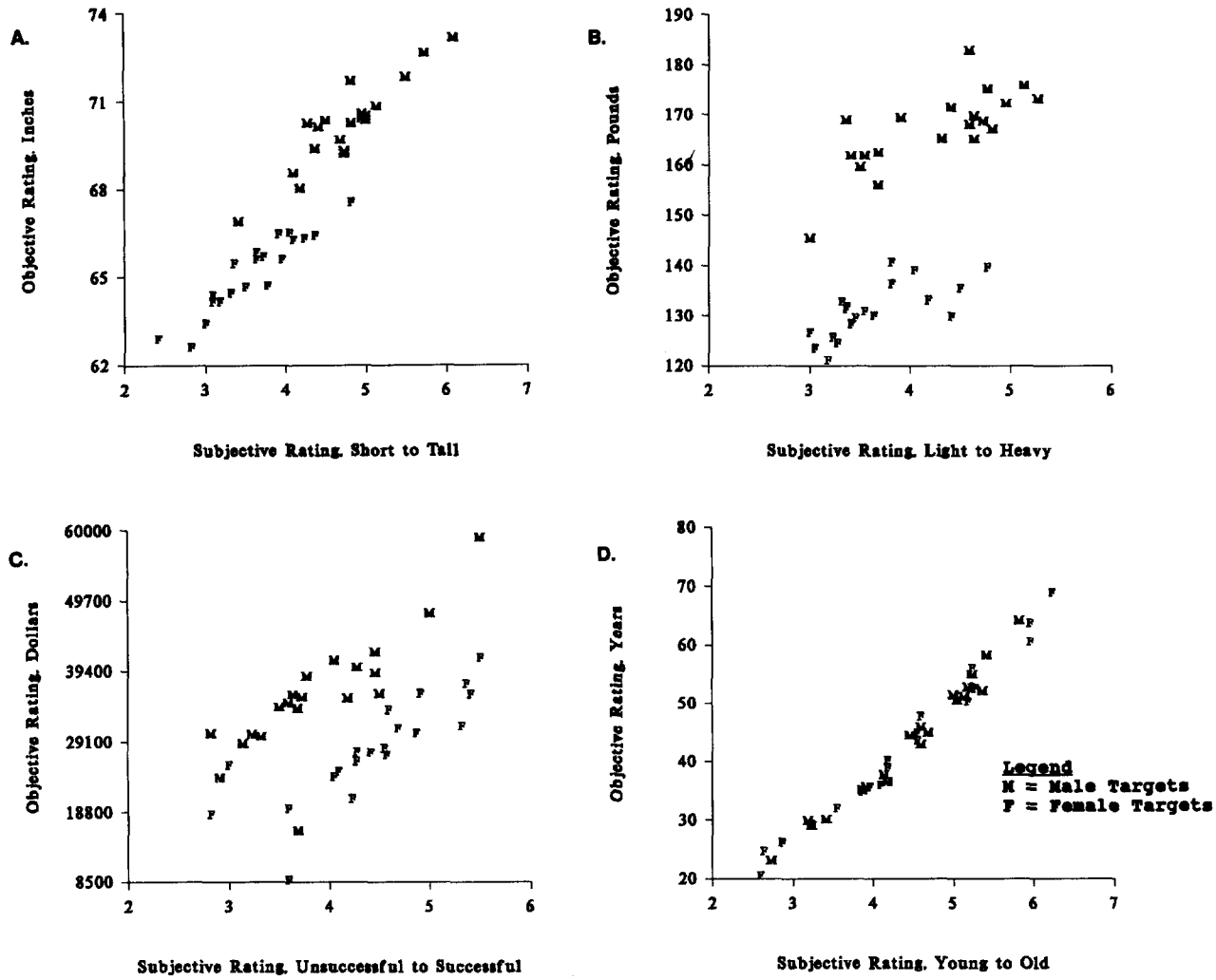

Figure 7. Scatter plots of average female and male attribute estimates in subjective versus objective units, Study 2.

extent to which our respondents' mental representations of different group members were affected by their stereotypes (expectations) regarding the group in question. By contrast, objective, consensually defined assessment procedures—involving such well-known metrics as feet and inches, pounds, and dollars per year—showed striking support for the proposition that the mental representations evoked by individual group members may be systematically affected by stereotyped expectations regarding the group as a whole.

We believe that these results derive from the shifting standards (end anchors) that respondents use when subjectively evaluating the members of different groups. In general, the subjective standards that are used derive from the range of variation that might plausibly be expected from the members of a given group. As a consequence, an individual target who seemed to be about 5 feet 8 inches in height might be subjectively regarded as "tall" if a woman, but as merely "average" (or perhaps below average) if a man. Similar shifts were observed when our respondents estimated the apparent weight and financial success of individual men and women. The results we obtained when targets' financial success was evaluated present a particularly striking example of respondents' inadvertent use of different subjective standards for assessing women and men. Briefly, as shown in Figure 7c, whereas the male targets were thought to earn more money per year than the female targets, these same men were regarded as being less successful than women when rated on a subjective scale.

As a final demonstration that subjective scale shifts depend on the differing expectations people have for various groups, the results of Study 2 showed no evidence of such a scale shift when respondents estimated the ages of individual men and women (see Figure 7d). Here, very similar results were obtained whether our respondents evaluated the different targets using a subjective (*very young* to *very old*) or an objective response scale (age in years). These results derive from the fact that men, as a group, are not expected to be any different from women with

respect to age. As a consequence, the end anchors of the respondents' subjective age scales were not affected by the sex of the different targets.

Study 3

The results of the first two studies are encouraging with regard to our main hypotheses; however, several remaining issues must be addressed. First, although we assumed the existence (or nonexistence) of sex stereotypes in regard to the various attributes investigated, we did not explicitly assess these beliefs. Second, the judgment domains we have used thus far involve demographics, rather than the kinds of traits or behaviors that are traditionally assumed to constitute stereotypes. We responded to both of these issues in the present study by explicitly asking subjects about their global stereotypes and by including several nondemographic judgment dimensions: time spent studying, movie-going behavior, and performance in math classes (only the last is conceptualized as a sex-linked attribute).

In our first two studies, we also may not have sufficiently dealt with the fact that our objective and subjective scales differ in ways other than objectivity and stability. One striking difference concerns the range of possible response alternatives: Much more variability is possible on an unbounded objective scale than on a 1–7 subjective response scale. In Study 3, we remove this problem by allowing the same number of response options in our subjective and objective conditions.

Finally, a plausible alternative explanation of our results exists and is easily tested in this third study. One could argue that our findings are based not on shifts in subjective judgment standards, but on differences in the difficulty, or cognitive effort, involved in making objective versus subjective judgments. Specifically, subjects may find it more difficult to rate targets in objective than in subjective units. Some evidence suggests that the more difficult or demanding a judgment task, the more likely one is to rely on global stereotypes as heuristics for dealing with heavy demands on information-processing capacity (Bodenhausen & Wyer, 1985; Wyer & Srull, 1989), and that such heuristic processing will be used to the extent that it allows subjects to "attain a sufficient degree of confidence" that they have satisfactorily accomplished their processing goals (Chaiken, Liberman, & Eagly, 1989, p. 221). In regard to the present studies, subjects forced to make the more difficult, objective ratings should be more likely to rely on heuristic processing and to show stereotype-assimilative effects than subjects making subjective ratings. This is, in fact, the pattern we have observed. To test the "effort" hypothesis, we asked our Study 3 subjects to rate the difficulty of the various judgment tasks.

Method

Subjects were 20 undergraduates (13 women and 7 men) at the University of Florida who volunteered to participate in the study during an otherwise cancelled class meeting of their introductory social psychology course.[9] Forty-two photographs from the same set used in Study 1 (using standing targets only) were presented in slide format during one experimental session. Subjects rated each target on the following attributes, always in this order: height ("How tall is this person?"), hours spent studying per week ("How much time does this student spend studying outside of class during a typical 5-day school week?"), number of movies seen per month ("How many movies per month does this student go out to see?"), and college math performance ("How poorly or well does this student perform in college math classes?"). After each height estimate, subjects were also asked to indicate how difficult they found the rating task on a 1 (*not at all difficult*) to 9 (*extremely difficult*) scale. Because we thought that subjects would find it too tedious to repeatedly answer individual difficulty questions, we did not ask for difficulty ratings after each judgment of studying, movies, or math. We did, however, ask *overall* difficulty questions at the end of the study, as described later.

Half of the subjects made their judgments in subjective rating units. Height (using the "average student" as a standard) was rated on a −9 (*very short*) to +9 (*very tall*) scale, with the zero-point labeled *average.* Studying time was also estimated on a −9 to +9 scale, with the end points labeled *very little time* and *very much time* and the zero-point marked *average.* Number of movies and math performance were each rated on −6 (*very few or no movies/very poorly*) to +6 scales (*very many movies/very well*), again with the zero-points labeled *average.*

The other half of the subjects estimated height, studying, movies, and math in objective units. In each case, the number of response category alternatives was equal to the number of alternatives allowed in the subjective condition. For example, 19 response categories, each corresponding to a 1-in. interval, were used for height judgments (thus matching the 19 levels of the −9 to +9 subjective scale). The first response category was marked "less than 5 feet tall," the second was marked "5 feet to 5 feet 1 inch," and so forth, with the 19th level marked "greater than 6 feet 5 inches."

Nineteen response alternatives were also offered for objective judgments of studying time. Options here included "less than 1 hour," "1–2½ hours," "2½–4 hours," and so forth, with the 19th level labeled "more than 26½ hours." Judgments of movies and math performance each had 13 response alternatives, corresponding to the −6 to +6 subjective scales. The movie response options ranged from "none" to "1 movie," and so forth, with the last category labeled "12 or more movies." The math performance response options were in the form of standard letter grades: A+, A, A–, and so forth, to E.

After judging each of the 42 slides on height, difficulty in estimating height, studying time, movies, and math, subjects were asked to indicate, *overall,* how difficult it was to make the various judgments (on 9-point *not at all difficult* to *extremely difficult* scales). Subjects then reported their own heights, studying hours, movies, math grades, and sex. The final sheet of the booklet asked subjects to judge the "average college male/female" on each of the attributes as well, with the order of these judgments (male or female first) counterbalanced. All subjects made these later judgments using the objective scales.

Results

Because objective and subjective ratings were made on scales with the same number of response alternatives, no standardization of the judgment variables was necessary for the key analyses. After dropping the two "practice" slides, we simply calculated subjects' mean judgments across the 20 female and the 20 male targets on each of the four attributes (height, studying, movies, and math). In each judgment domain, the main analysis took the form of a 2 (sex of subject) × 2 (scale type: objective

[9] We thank Teri Davis and Joe Alvarez for their help in running this study.

or subjective) × 2 (sex of target) ANOVA; sex of target was a within-subject variable.

Our main hypothesis concerns the interaction between scale type and sex of target: When a judgment domain is sex linked (i.e., when subjects have a gender stereotype involving an attribute), there should be a statistical interaction such that judgments in objective units reveal the stereotype, whereas judgments in subjective units mask it. In this study, we viewed height and math performance as sex-linked attributes, and studying time and movies as non-sex-linked attributes.

To be certain our subjects agreed with us, we looked for sex differences in their overall (base-rate) judgments of the average college man and woman on the four attributes. Subjects did rate the average college man ($M = 69.97$) as being significantly taller than the average college woman ($M = 65.5$), $t(18) = 10.53$, $p < .0001$; however, they did not show evidence of a sex stereotype concerning math performance, $t(18) = 1.53$, $p > .15$. Studying time and movie base rates also were not different for male and female targets. In sum, these data led us to slightly revise our predictions. Because math performance is apparently not perceived as being sex linked, we should find no evidence of a sex of target by scale type interaction in subjects' math judgments; we should only find such an interaction in judgments of height.

This is, in fact, precisely what our data indicated. The analysis of variance on height estimates, but not on any of the other judgment dimensions, produced a significant sex of target by scale type interaction, $F(1, 16) = 104.63$, $p < .0001$, which is depicted in Figure 8. In comparing the height judgment data across the three studies, the reader will note that Study 3 shows the first case of a *reversal* of the height sex stereotype (a crossover interaction) when judgments were made in subjective units. More will be said on this point later. The only other significant finding in the analysis of height judgments was a main effect of scale type, $F(1, 16) = 7.44$, $p < .05$. Average ratings in subjective units ($M = 10.63$) were higher than average ratings in objective units ($M = 8.84$).

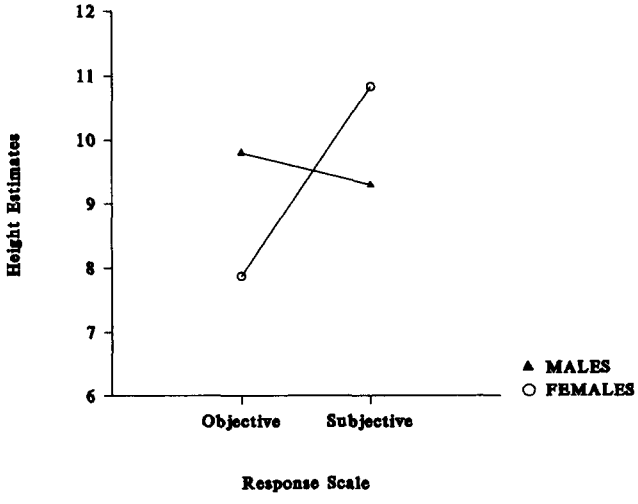

Figure 8. Interaction between sex of target and scale type on height judgments, Study 3.

Other findings of note in these analyses included a main effect of sex of target when studying was the dependent variable, with women ($M = 9.60$) rated higher than men ($M = 8.40$) $F(1, 16) = 22.11$, $p < .0005$. This was qualified, however, by an interaction with sex of subject, $F(1, 16) = 5.31$, $p < .05$. Female subjects reported that female targets studied significantly longer than male targets, whereas male subjects reported no difference between female and male targets. Finally, in the analysis of movie judgments, scale type interacted with sex of subject, such that male subjects making subjective ratings gave significantly higher estimates than male subjects making objective ratings, whereas this difference was attenuated among female subjects, $F(1, 16) = 5.42$, $p < .05$.

An alternative account of some of our findings has to do with the difficulty, or the amount of cognitive effort, involved in making objective versus subjective judgments. Specifically, it may be that judging an attribute on an objective scale is more difficult than judging that same attribute on a subjective scale. If stereotypes are more heavily relied upon when a judgment task is difficult (see Bodenhausen & Wyer, 1985; Chaiken et al., 1989; Wyer & Srull, 1989), subjects forced to make objective ratings should show more evidence of stereotyping than subjects making subjective ratings (the pattern of findings observed in all three studies).

To test this "cognitive effort" hypothesis, we simply asked subjects how difficult they found the judgment tasks to be. We did this in the form of single "overall difficulty" questions pertaining to the studying, movie, and math ratings, and with the overall as well as individual difficulty questions that followed each of the 40 height judgments. Did subjects find objective ratings more difficult than subjective ratings? The data clearly say "no." On the overall difficulty questions, the difference between subjective and objective ratings was nonsignificant (ts < 1) for height, studying, and movie estimates. The difference *was* significant for math judgments, $t(18) = 2.33$, $p < .05$, but the direction of the effect was the reverse of that expected: Subjective math ratings were judged more difficult to make than objective ratings (Ms = 8.22 and 6.5, respectively). Finally, we analyzed the 40 individual "height difficulty" questions using the 2 (sex of subject) × 2 (scale type) × 2 (sex of target) ANOVA format described earlier. The effect of scale type was not significant ($F < 1$); in fact, the only significant finding was a main effect of sex of target, $F(1, 16) = 5.68$, $p < .05$. Subjects found it more difficult, on average, to estimate the heights of male targets than of female targets (Ms = 4.48 and 4.27).[10]

As a final pass at the difficulty account of our findings, we reran the analyses of variance reported earlier, this time using the appropriate difficulty ratings as covariates. Covarying out

[10] We wondered whether subjects' difficulty ratings might be more meaningful if they involved an explicit comparison between subjective and objective ratings (i.e., if we manipulated scale type as a within-subject variable). In a replication of the present study, we ran 26 additional subjects who alternated between subjective and objective ratings from one slide to the next. Our analysis of the individual height difficulty questions again indicated no significant effect of scale type ($F < 1$). In fact, all of the analyses from this "scale type as a within-subject variable" study yielded results remarkably similar to those reported here.

difficulty did not change our results in any way. Of particular concern was the significant scale type by sex of target interaction on height judgments. Whether we used the overall difficulty or aggregated individual difficulty ratings as covariates, this interaction remained significant (Fs > 75.00 in both cases). In short, we find no support for the hypothesis that differences in difficulty between subjective and objective rating tasks can account for our pattern of findings.

Discussion

Study 3 ties up several loose ends from our earlier studies and lends considerable additional support to our hypothesis concerning stereotype effects and shifts in judgment standards. First, we removed a key difference between our subjective and objective rating scales—the range of possible response values—by providing the same number of response options for each task. We also included difficulty ratings to test whether the cognitive effort involved in making subjective versus objective judgments might account for our earlier findings. Finally, Study 3 incorporated three additional judgment dimensions to test the generalizability of our model. With these changes, we continue to find that when an attribute is part of prevailing sex stereotypes, subjects' judgments of individual targets on that attribute are influenced by the type of response scale used. Objective judgments reveal stereotype effects; subjective judgments do not. We suggest that firmly anchored objective scales do not allow subjects to shift standards as they naturally do when judging objects from different categories.

This finding does not seem to be due to the relative difficulty of making objective versus subjective judgments. However, the measure of cognitive effort (difficulty) we used in this study may not have been a valid indicator of subjects' judgment experiences. Self-reports of effort are problematic given that they concern a cognitive *process* to which people are unlikely to have much access (Ericsson & Simon, 1984; Nisbett & Wilson, 1977). This is particularly true in the present study, in which we asked for difficulty estimates after the fact, rather than concurrent with the judgment. In future work, a better test of the cognitive effort hypothesis will involve obtaining on-line measures of effort, such as thought protocols or response latencies for the objective versus subjective scales.

We had hoped that judgments of math performance would provide another domain where sex stereotypes operated. Our subjects, however, did not perceive a global difference in math performance between male and female college students. Anonymity of subjects was assured, so it seems unlikely that this lack of sex stereotyping was based on social desirability concerns. Furthermore, male subjects were no more likely than female subjects to report a sex difference in math performance. It seems more likely that college subjects simply do not perceive a relationship between sex and math performance, at least among college targets. Whatever the basis of their reports, these data signify the importance of asking subjects whether or not they espouse particular stereotypes. Furthermore, stereotypes may differ in the tenacity with which they are held. The sex stereotype concerning height is strong (perhaps because it is undoubtedly accurate), whereas others (i.e., sex and assertiveness) may be weak. Measuring stereotype strength may be one way of specifically identifying the judgment standards subjects invoke as they evaluate individual targets from different social categories.

General Discussion

Our general hypothesis was well-supported: Subjective scales appear to be adjusted to suit the range of values one expects to find in a particular target group. When evaluating men and women, different standards of height, weight, and financial success are used, even when respondents are explicitly instructed to make their judgments relative to the "average person." It seems plausible to conclude that scale adjustments of this type are quite common, and that they affect people's freely produced verbal assessments of targets in everyday life (e.g., when they characterize individuals from diverse segments of society) and in the laboratory, where they are provided with bounded rating scales while serving as subjects in our person perception experiments. This type of scale adjustment serves to *reduce* the apparent effect of group stereotypes on one's assessments of individuals and may explain the absence of significant stereotype effects in studies by Locksley et al. (1980, 1982) and others. In short, inadvertent scale adjustments may mask, or at least diminish, the effects that stereotypes actually exert on one's mental representations of individuals.

As cases in point, we have applied this reasoning to several recent research findings. One example comes from the work of Linville and Jones (1980), who reported that hypothetical law school applicants with the same credentials were judged more extremely if they were Black than if they were White. Thus, when reading a strong application, White subjects rated a Black applicant more favorably than a comparable White applicant; when reading a weak application, they rated a Black applicant less favorably than a comparable White applicant.

From this study, as well as Linville's other work (Linville, 1982, 1985, 1987; Linville, Salovey, & Fischer, 1986), it seems clear that the complexity-extremity hypothesis provides an extremely viable and elegant account of this type of finding. We are only speculating on one possible additional mechanism behind these judgments—the use of different standards of judgment for Black as opposed to White students. If Whites believe that Blacks are generally weaker than Whites in academic ability, the mental representation associated with a particular applicant might produce different evaluations, depending on that applicant's race. This possibility is graphically depicted in Figure 9, where it is assumed that a given applicant (e.g., one with an A– grade point average) might be *mentally* represented as having higher academic ability if she or he is White rather than Black. Nonetheless, as Figure 9 suggests, whereas our hypothetical respondent assumes that the Black applicant with the A– average has less academic ability than the White applicant with comparable credentials, the Black might be given a more favorable rating than the White, because her or his perceived standing would place her or him well above the rater's expectations for Blacks. By comparison, the White applicant might be evaluated against a more demanding set of standards, and hence would receive a less favorable judgment.

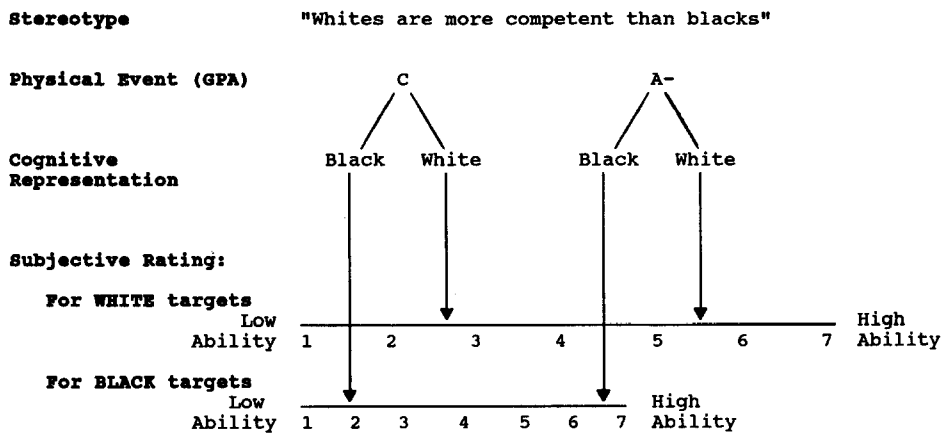

Figure 9. Schematic model of judgment standard and scale shifts applied to Linville and Jones (1980).

A similar analysis can be applied in the case of a less-qualified (C grade point average) applicant. Here, however, the Black applicant receives a lower evaluation than the White with those same C credentials. Figure 9 depicts a situation in which, because of the operation of group stereotypes, the Black applicant is again regarded (at a representational level) as having less competence or ability than the White applicant. Although these cognitive representations are labeled with respect to rather different subjective standards that depend on the target's racial heritage, in this case the predicted ratings are consistent with the representation or stereotype (the Black C applicant is rated as less promising than the White applicant).

Why do these disparate labeling systems (for White and Black targets) yield one set of results for "good" applicants and a very different pattern for "weak" applicants? Figure 9 suggests that at the low end of the "ability" continuum, respondents' subjective scales are anchored at about the same point, whether the target is White or Black. At the high end of the continuum, in contrast, the end anchor of the scale for White targets is more extreme than it is for Blacks. The location of these end anchors is consistent with evidence that "contemporary stereotypes involve differential association of positively valued characteristics to whites, but not negatively valued traits to blacks" (Gaertner & Dovidio, 1986, p. 83). This conclusion stems from reaction time studies, in which subjects respond more quickly to positive traits following a "White" versus a "Black" prime, but do not differentially respond to negative traits following those primes (Dovidio, Evans, & Tyler, 1984). Similar results have been reported in trait rating studies, in which race is differentially related to mean positive evaluative ratings, but not to mean negative ratings (Dovidio, 1984; Gaertner & McLaughlin, 1983; see also Gaertner & Dovidio, 1986). Thus, although Blacks, as a group, are not expected to be "more stupid" than Whites, they are thought to have relatively few group members with outstandingly positive intellectual attributes. These differing expectations, in turn, are assumed to underlie the location of the end anchors in Figure 9.

Again, our account is not an attempt to refute the impressive literature that has accumulated in support of the complexity-extremity hypothesis. It certainly cannot explain the strong correlations typically found between complexity scores and evaluative extremity, the data on self-complexity, or Linville's (1982) young–old studies (because no general belief exists that one age group is more likable than the other). The present analysis only suggests that the application of different judgment standards, based on group stereotypes, might also affect evaluations of members of those stereotyped groups.

An important next step in research on this topic, suggested by the above analysis, is to develop a method for precisely measuring the subjective standards subjects use in judgment tasks. Can one explicitly identify the end anchors subjects refer to as they judge targets from different social categories? In Study 3, we assessed subjects' beliefs about the *average* female and male targets' standing on the attributes of interest. This allowed us to predict the presence or absence of a sex of target by scale type interaction, but it did not allow us to predict the *pattern* of that interaction. For example, we know from our base rate questions that our Study 3 subjects perceived a 4½-in. difference in height between the average male and the average female college student. We predicted, in turn, that the corresponding stereotype effect would be strong in the objective judgment condition, as we had observed in our earlier studies. What we could not predict is the reversal of the stereotype effect in the subjective judgment condition (see Figure 8); Studies 1 and 2 indicated a reduction, not a reversal, in the stereotype effect.

Because the stimulus materials were essentially the same in Studies 1 and 3, what accounts for the observed difference between the interaction patterns (between target sex and scale type) in the two studies? Some obvious methodological differences must be mentioned. In Study 1, subjects judged only the heights of targets, made their subjective judgments on 1 to 7 scales using the "average person" as a standard of comparison, and worked at their own pace through a questionnaire booklet. In Study 3, subjects judged height along with three other attributes, made both their subjective and objective judgments on 19-point scales (using the "average student" as a standard of comparison in the former case), and responded to slides flashed for limited amounts of time during a group testing session. Aside from these differences, we can only speculate that subjects in the two studies used slightly different standards of height for women and men.

To more satisfactorily address questions of this sort, a better

method is needed for determining precisely what those standards are. In our analysis of the Linville and Jones (1980) study, we used evidence from Gaertner and Dovidio's lab to identify the "end points" of subjects' distributions of Blacks and Whites across a general negative to positive continuum. The range of expected values for Black and White targets apparently begins at the same point, but ends at a lower level for Blacks than for Whites. This information allows us to calculate the specific pattern of results likely to emerge (in this case, the more extreme ratings for Black than for White targets).

One other option for assessing judgment standards is to measure the perceived overlap in subjects' distributions of members of different categories across levels of the attribute of interest. For example, what range and distribution of heights do subjects perceive for female versus male adults? As Figure 1 suggests, the smaller the perceived overlap in scale values anticipated for female and male targets, the more likely one is to find *reverse* stereotype effects in subjective ratings. Our financial status data from Study 2 (specifically the reversal of the stereotype effect when judgments were made in subjective units) are consistent with the speculation that subjects perceive very little overlap in the distribution of incomes for women and men. Thus, a woman who "looked" financially successful easily surpassed a comparable man when she was rated on a subjectively defined scale. Clearly, it would be helpful to explicitly test the "distribution overlap" hypothesis and to use such a measure of judgment standards as a means of predicting when and how much bias in subjective judgments will occur.

Although our discussion has focused primarily on biases that occur at the point of response generation, we also recognize that encoding processes play an important role when subjects are asked to make judgments of members of stereotyped groups. In fact, we assume that it is the on-line, relatively automatic encoding of targets by sex (or race, for example) that influences the use of differential judgment standards. Our schematic model presented in Figure 1 reflects the importance of encoding in its assumption that at the heart of the judgment process is a stereotyped belief, spontaneously applied to category members and manifested either directly in objective response units or indirectly (due to standard shifts) in subjective responses.

We must also raise one additional issue concerning the judgment model we have proposed. Our assumption has been that bias appears in *subjective* ratings. It is, however, possible that, instead, our *objective* ratings are the more subject to bias and scale range artifacts. Consider, for example, a model in which we assume individuals cognitively represent the world in qualitative (rather than quantitative) terms, and that these representations are reflected in overt responses on external qualitative scales. Quantitative judgments, on the other hand, may be distortions of internal representations, chosen to reflect stereotypic beliefs. In such a model, subjective (i.e., qualitative) judgments are the more accurate reflections of subjects' mental constructs.

We acknowledge this model as a possible alternative to our own, but suggest that our account is the more plausible for several reasons. First, in our studies, objective judgments approximated reality more closely than did subjective judgments (e.g., men *do* earn more money on average than do women, a fact subjects readily recognized). Second, our data from the objective conditions more closely resembled data we have collected using paired comparison methods. For example, in previous studies (Nelson et al., 1990), and at the end of Study 1 (although not reported here), we asked subjects to choose "Who's taller?" in pairs of men and women who had been matched for height. We assume that paired comparisons provide us with a relatively direct index of the perceived heights of individual men versus women, and therefore are a good "criterion" measure of subjects' internal representations. In all of our work using this methodology, we have discovered a clear stereotyping effect, with men chosen significantly more often as the taller of mixed-sex, matched-height pairs. The fact that we also find consistent stereotype effects in our *objective* single-judgment studies suggests, then, that such data provide a closer approximation to the criterion (paired comparison) results than do subjective ratings, on which patterns of findings are inconsistent across studies.

Finally, in Study 1, objective ratings more than subjective ratings were consistently related to paired comparisons *across respondents.* For each subject, we calculated a "stereotype score" —the difference between the subject's average male and female height assessments—and correlated this score with his or her probability of choosing the male photograph as the taller in a series of 16 pairs. Among subjects in the "feet and inches" judgment condition, this correlation was significant at $r = .40$ ($n = 42$, $p < .01$), but the comparable correlations among the "average person" and "average for sex" subjects were nonsignificant at $rs = .06$ ($n = 47$) and .11 ($n = 44$), respectively. That is, the objective ratings more closely matched what we assume to be our direct index of mental representation, the paired comparison data. Of course, our data cannot definitively confirm the form of subjects' internal representations, but our model appears to be a reasonable account of the representation–judgment process.

Final Comments

We have attempted to identify what we believe to be a very general phenomenon: the implicit use of different judgment standards when evaluating individuals drawn from diverse social categories. We have proposed, moreover, that these differing standards (i.e., different end anchors) are based on global stereotypes concerning the range or variation that might plausibly be expected among the members of a given group. It seems ironic to recognize that these shifting judgment standards may underlie some cases in which stereotype effects appear to be minimal (e.g., Locksley et al., 1980, 1982).

Previous discussions regarding the deficiencies of rating scales due to the shifting locations of relevant end anchors have focused mainly on between-subject differences based on divergent recent experiences. For example, a judge may be assigned to one of two contrasting *context* conditions, involving exposure to large versus small animal names (Herr, Sherman, & Fazio, 1983). These divergent "local contexts," in turn, may determine the subjective representations associated with the relevant end anchors (e.g., when the sizes of other animals are judged). The present work, on the other hand, highlights the shifting nature of people's everyday judgment standards as they encounter

individuals from diverse social groups. These studies raise a cautionary flag about experimental work on person perception—inadequate subjective scaling methods may mask substantial stereotype effects. Indeed, as shown in our study of financial status (see Figure 6), the results obtained with subjective scales may provide a very misleading picture of respondents' mental representations.

To overcome some of the deficiencies associated with subjective rating scales, three suggestions seem worthy of further exploration:

1. Where it is possible, researchers would do well to occasionally consider the use of objective (consensually defined) judgment scales such as dollars, feet and inches, and perhaps grade point average (as an externally anchored index of academic performance).

2. Scaling methods that bypass the necessity for constructing subjectively defined rating categories should also be used. For example, Thurstone's various paired comparison procedures focus on the *direct comparison* of individual stimuli, and presumably provide a fairly close approximation to subjects' internal representations of targets. Techniques of this sort may help future researchers overcome some of the assessment problems that have been documented in the present experiments (see e.g., Dawes, 1972; Dawes & Smith, 1985; Manis, Nelson, & Shedler, 1988; Manis & Paskewitz, 1984).

3. Subjective response scales will undoubtedly continue to be popular, particularly when researchers ask for judgments of the rather abstract traits (e.g., laziness, assertiveness) contained in many stereotypes. To reduce the extent of the standard shift phenomenon and to provide a better context for interpreting resulting data, researchers should always provide respondents with explicit standards of reference when subjective judgment scales are used.

References

Bar-Hillel, M. (1980). The base rate fallacy in probability judgments. *Acta Psychologica, 44,* 211–233.

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

Bodenhausen, G. V., & Wyer, R. S., Jr. (1985). Effects of stereotypes on decision making and information-processing strategies. *Journal of Personality and Social Psychology, 48,* 267–282.

Broverman, I. K., Vogel, S. R., Broverman, D. M., Clarkson, F. E., & Rosenkrantz, P. S. (1972). Sex-role stereotypes: A current appraisal. *Journal of Social Issues, 28,* 59–78.

Chaiken, S., Liberman, A., & Eagly, A. H. (1989). Heuristic and systematic information processing within and beyond the persuasion context. In J. S. Uleman & J. A. Bargh (Eds.), *Unintended thought: Limits of awareness, intention, and control* (pp. 212–252). New York: Guilford.

Dawes, R. M. (1972). *Fundamentals of attitude measurement.* New York: Wiley.

Dawes, R. M., & Smith, T. L. (1985). Attitude and opinion measurement. In G. Lindzey & E. Aronson (Eds.), *The handbook of social psychology* (Vol. 2, pp. 509–566). New York: Random House.

Dovidio, J. F. (1984). *Attributions of positive and negative characteristics to Blacks and Whites.* Unpublished manuscript, Colgate University, Hamilton, NY.

Dovidio, J. F., Evans, N., & Tyler, R. (1984). *Racial stereotypes as prototypes.* Unpublished manuscript, Colgate University, Hamilton, NY.

Eiser, J. R. (1971). Enhancement of contrast in the absolute judgment of attitude statements. *Journal of Personality and Social Psychology, 17,* 1–10.

Eiser, J. R., & Stroebe, W. (1972). *Categorization and social judgment.* San Diego, CA: Academic Press.

Ericsson, K. A., & Simon, H. A. (1984). *Protocol analysis: Verbal reports as data.* Cambridge, MA: MIT Press.

Gaertner, S. L., & Dovidio, J. F. (1986). The aversive form of racism. In J. F. Dovidio & S. L. Gaertner (Eds.), *Prejudice, discrimination, and racism* (pp. 61–90). San Diego, CA: Academic Press.

Gaertner, S. L., & McLaughlin, J. P. (1983). Racial stereotypes: Associations and ascriptions of positive and negative characteristics. *Social Psychology Quarterly, 46,* 23–30.

Ginossar, Z., & Trope, Y. (1980). The effects of base rates and individuating information on judgments about another person. *Journal of Experimental Social Psychology, 16,* 228–242.

Ginossar, Z., & Trope, Y. (1987). Problem solving in judgment under uncertainty. *Journal of Personality and Social Psychology, 52,* 464–474.

Gravetter, F., & Lockhead, G. R. (1973). Criterial range as a frame of reference for stimulus judgment. *Psychological Review, 80,* 203–216.

Herr, P. M., Sherman, S. J., & Fazio, R. H. (1983). On the consequences of priming: Assimilation and contrast effects. *Journal of Experimental Social Psychology, 19,* 323–340.

Higgins, E. T. (1977). The varying presuppositional nature of comparatives. *Journal of Psycholinguistic Research, 6,* 203–222.

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

Huttenlocher, J., & Higgins, E. T. (1971). Adjectives, comparatives, and syllogisms. *Psychological Review, 78,* 487–504.

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

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review, 80,* 237–251.

Kreuger, J., & Rothbart, M. (1988). Use of categorical and individuating information in making inferences about personality. *Journal of Personality and Social Psychology, 55,* 187–195.

Linville, P. W. (1982). The complexity-extremity effect and age-based stereotyping. *Journal of Personality and Social Psychology, 42,* 193–211.

Linville, P. W. (1985). Self-complexity and affective extremity: Don't put all of your eggs in one cognitive basket. *Social Cognition, 3,* 94–120.

Linville, P. W. (1987). Self-complexity as a cognitive buffer against stress-related depression and illness. *Journal of Personality and Social Psychology, 52,* 663–676.

Linville, P. W., & Jones, E. E. (1980). Polarized appraisals of outgroup members. *Journal of Personality and Social Psychology, 38,* 689–703.

Linville, P. W., Salovey, P., & Fischer, G. W. (1986). Stereotyping and perceived distributions of social characteristics: An application to ingroup-outgroup perception. In J. 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: An instance of the base-rate fallacy. *Journal of Experimental Social Psychology, 18,* 23–42.

Manis, M. (1967). Context effects in communication. *Journal of Personality and Social Psychology, 5,* 326–334.

Manis, M. (1971). *An introduction to cognitive psychology.* Belmont, CA: Brooks/Cole.

Manis, M., Dovalina, I., Avis, N. E., & Cardoze, S. (1980). Base rates can affect individual predictions. *Journal of Personality and Social Psychology, 38,* 231–248.

Manis, M., Nelson, T. E., & Shedler, J. (1988). Stereotypes and social judgment: Extremity, assimilation, and contrast. *Journal of Personality and Social Psychology, 55,* 28–36.

Manis, M., & Paskewitz, J. R. (1984). Specificity in contrast effects: Judgments of psychopathology. *Journal of Experimental Social Psychology, 20,* 217–230.

McKee, J. P., & Sherriffs, A. C. (1957). The differential evaluation of males and females. *Journal of Personality, 25,* 356–371.

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.

Nisbett, R. E., & Ross, L. (1980). *Human inference: Strategies and shortcomings of social judgment.* Englewood Cliffs, NJ: Prentice-Hall.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review, 84,* 231–259.

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

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. 188–252). New York: Praeger.

Spence, J. T., & Helmreich, R. (1978). *Masculinity and femininity.* Austin: University of Texas Press.

Upshaw, H. S. (1965). The effects of variable perspectives on judgments of opinion statements for Thurstone scales. *Journal of Personality and Social Psychology, 2,* 60–69.

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. R. Rohrer & M. Sherif (Eds.), *Social psychology at the crossroads* (pp. 273–294). New York: Harper.

Wyer, R. S., Jr., & Srull, T. (1989). *Memory and cognition in its social context.* Hillsdale, NJ: Erlbaum.

Zukier, H., & Jennings, D. L. (1984). Nondiagnosticity and typicality effects in prediction. *Social Cognition, 2,* 187–198.

Received May 30, 1990
Revision received October 23, 1990
Accepted October 23, 1990 ■