Psychological Bulletin
1989, Vol. 105, No. 3, 409–429

Joan McKay Versus John McKay: Do Gender Stereotypes Bias Evaluations?

Janet Swim, Eugene Borgida, and Geoffrey Maruyama
University of Minnesota

David G. Myers
Hope College

Examines research using a classic, influential experiment conducted by Goldberg (1968), showing that women were likely to rate male authors (e.g., John T. McKay) more favorably than female authors (e.g., Joan T. McKay) of identical articles. Although replications of this study have been inconclusive, Goldberg's research is still frequently cited as demonstrating an evaluative bias against women. A quantitative meta-analysis of research using Goldberg's experimental paradigm shows that the average difference between ratings of men and women is negligible. Furthermore, although the effect sizes are not homogeneous, the difference remains negligible when other factors such as sex of subject or year of publication are taken into consideration. Several explanations for the heterogeneity of effect sizes and the inconsistency of findings are discussed.

In 1968, social psychologist Philip Goldberg published his influential experimental study on prejudice against women (Goldberg, 1968). The study was simple yet compelling. Goldberg gave female subjects identical booklets containing six different articles. For each article, however, half the subjects were told that the author was a woman (e.g., Joan T. McKay) and the other half were told that the author was a man (e.g., John T. McKay). On a number of ratings, such as evaluations of the author's competence, the male author was given higher ratings than the female author.

Specifically, Goldberg reported that men received higher ratings on 44 of 54 measures. However, after forming an index by summing across the dependent variables, he reported six *t* tests, only three of which were significant. Two of the three significant findings were for articles about traditional masculine fields (law and city planning), and one was for a sex-neutral field (linguistics). Two of the three nonsignificant findings were for the articles on traditionally feminine fields (dietetics and education), and the third was for a second sex-neutral field (art history). In effect, then, Goldberg seemed to have found qualified support for gender bias in evaluations.

Perhaps not surprisingly, Goldberg's article is frequently cited as evidence of discrimination against women and of women holding prejudicial beliefs about their own gender (e.g., Cash & Trimer, 1984; Gornick & Moran, 1971; Lips & Colwill, 1978; Paludi & Strayer, 1985; Ruble & Ruble, 1982). Although it is well established that discrimination occurs (e.g., Hewlett, 1986) and that Goldberg's findings certainly seem to illustrate sex bias, many authors in the scientific and popular literature who cite his article in fact misrepresent the strength of the results Goldberg reported. In addition, the robustness of Goldberg's findings is uncertain because of numerous inconsistent conceptual replications.

One frequent misrepresentation of Goldberg's study has been to distort the nonsignificant results. For instance, Paludi and Bauer (1983) stated, "Results indicated that women rated the articles (even those in fields considered sex appropriate for women) more favorably when they were attributed to a male rather than a female author" (p. 387). In fact, Goldberg does not report any significant differences on ratings of the authors on the traditionally feminine fields. Further, Goldberg did not demonstrate discrimination in all of the comparisons tested.

Other authors do not misstate the conclusion but implicitly distort the findings by failing to mention the nonsignificant findings. For instance, Lips and Colwill (1978) stated, "This stereotype is seen in action in studies that look at the way people evaluate male and female performance. Goldberg (1968) showed that female college students evaluated articles supposedly written by women lower than the identical articles attributed to male authors" (p. 188). In a more recent review paper, Wallston and O'Leary (1981) also distorted the findings by stating, "In the landmark study of competency bias favoring men (Goldberg, 1968), college women rated professional articles for value, persuasiveness, profundity, writing style and competence. Higher ratings were given to identical papers when the author of the article was portrayed as a male rather than a female" (p. 19). Omitting the nonsignificant effects creates the

Janet Swim, who is now in the Department of Psychology at Pennsylvania State University, was supported in part by a fellowship from the Graduate School of the University of Minnesota. The authors are grateful to John Claypool for his research assistance and to Nancy Brekke for her insights during the formative stages of the research. Allen Omoto, Mark Snyder, Cynthia Thomsen, and especially Alice Eagly provided helpful comments on an earlier version of the manuscript. Portions of this research were presented at the 1986 meetings of the American Psychological Association, Washington, DC.

Correspondence concerning this article should be addressed either to Janet Swim, Department of Psychology, 515 Moore Building, Pennsylvania State University, University Park, Pennsylvania 16802, or to Eugene Borgida, Department of Psychology, University of Minnesota, 75 East River Road, Minneapolis, Minnesota 55455.

impression that the Goldberg effects were more robust than they actually were.

The strength of Goldberg's findings have also been distorted by disregarding failures to replicate. For instance, in his well-respected introductory psychology text, Gleitman (1981) described Goldberg's study but did not mention any of the conceptual replications. In a review paper, Unger (1976) referenced both Goldberg (1968) and Pheterson, Kiesler, and Goldberg (1971) as evidence for "the persistent tendency of females to devalue the work of other females" (p. 5), thus leaving the impression that Goldberg's findings have been supported by replications.

More recent reviews of this literature, such as those by Wallston and O'Leary (1981) and Basow (1986), have acknowledged the presence of studies that fail to replicate Goldberg's findings. These studies examined additional independent variables (e.g., status of the person), different samples (e.g., male subjects), and different stimulus materials (e.g., job applications). The conclusions of these reviews suggest that these empirical inconsistencies can be explained by various qualifying conditions. For instance, Basow stated, "From the vast amount of research in this area since 1968, it is clear that prejudice still exists in both men and women, but it usually exists in interaction with situational factors, and more often in subtle, as opposed to obvious ways" (pp. 234–235).

Reviews of this literature to date have primarily been qualitative reviews (e.g., Basow, 1986; Wallston & O'Leary, 1981); such reviews are constrained by reviewers' biases and limitations in synthesizing information. Qualitative reviews typically require reviewers to subjectively evaluate studies, which can lead to bias resulting from the reviewers' beliefs or expectations. Perhaps more importantly, qualitative reviews are characterized by the inherent difficulty of summarizing large numbers of studies and by the neglect by reviewers of large amounts of data in the original reports (see Cooper & Rosenthal, 1980; Glass, 1976). As an alternative approach, quantitative methods such as meta-analysis have been developed to examine results across a body of conceptually interrelated studies. These methods can eliminate some of the biases associated with qualitative reviews. For instance, quantitative reviews tend to be more objective and are more likely to include most, if not all, relevant studies (see Bangert-Drowns, 1986; Green & Hall, 1984).

In light of the questionable strength of Goldberg's original findings and because of the limitations associated with qualitative reviews, the validity of even qualified conclusions, such as those of Basow (1986), is uncertain. The purpose of our study, then, is to use quantitative methods in the form of meta-analysis techniques (a) to assess the strength of the tendency to discriminate against women in studies using variations of the paradigm originally used by Goldberg and (b) to identify and examine variables that might influence the strength of this tendency. In the first section, hypotheses that have been tested by other investigators are reviewed. These same hypotheses are also examined in the quantitative meta-analysis presented later in the article. Additional hypotheses about methodological differences across studies and about characteristics associated with the publication of the articles are presented in the second and third sections, respectively.

Variables Examined in Studies Subsequent to Goldberg's Original Study

Subject characteristics. Three subject characteristics tested in previous research are examined in this review. Although Goldberg (1968) only studied female subjects, other studies frequently tested differences between male and female subjects. In these studies it was generally expected that male subjects would be more likely than female subjects to rate female target persons lower than male target persons. A related subject variable is sex role orientation. It was expected that masculine or feminine subjects would be more discriminatory than androgynous subjects. A third subject variable is age of the subjects. Like subjects with masculine or feminine sex role orientations, older people were expected to discriminate more than younger people.

Target person characteristics. One characteristic of the target person that has been studied is his or her physical attractiveness. People who are more attractive are likely to be rated more favorably than less attractive people (Berscheid, 1985; Dion, Berscheid & Walster, 1972). However, attractiveness may be more important for women than for men (Bar-Tal & Saxe, 1976; Wallston & O'Leary, 1981). Thus, more discrimination might be predicted for unattractive women, in comparison with unattractive men.

A second characteristic of the target person is his or her race. To the extent that evaluations of women reflect general impressions based upon their lower status in society (Unger, 1976), other visible and salient status variables such as race should also affect evaluations. Insofar as effects of status variables are cumulative, minority women may be judged particularly unfavorably. Moreover, the interplay between race and sex cues may be more complex, resulting in greater differences between non-White women and men than between White women and men. Thus it seems important to examine directly status cues based on race in order to understand better how evaluations of men and women may differ.

Two opposing conclusions have been proposed regarding a third characteristic of the target person, that of competency or success of the target person. First, because being a woman is typically seen as a lower marker of status than is being a man (Unger, 1976), competent or successful women would violate gender stereotypes. If so, competent or successful women might receive greater discriminatory treatment by evaluators than would incompetent or unsuccessful women. In support of this perspective, Nieva and Gutek (1980) stated, "While females are evaluated less favorably than males when they are highly qualified or perform well, females are evaluated more favorably than males when both are not well qualified or are unsuccessful performers" (pp. 273–274). On the other hand, biases may exist when women have not displayed competency or success, but biases may not exist when women have displayed competency or success. For instance, Wallston and O'Leary (1981) stated, "The results indicate that women are likely to be evaluated as being as competent as men when their performance is deemed exceptional" (p. 20). These opposite conclusions may not be inconsistent but rather may be a function of an additional variable, namely, whether the competency or success was displayed in a traditionally female or male occupation (see Basow, 1986; Wallston & O'Leary, 1981).

Stimulus material characteristics. Some studies have varied characteristics of the stimulus materials in addition to different characteristics of the target person. Quality of the stimulus material has been manipulated in a manner similar to the way competency or success of a target person has been manipulated. For instance, some studies varied the success of the target person by indicating whether he or she had a college degree, and others varied the quality of the stimulus material by indicating whether the article being rated had been accepted for publication. Thus, opposing predictions could be made for quality of the stimulus material in the same way as they are made for competency or success of a target person.

A second characteristic of the stimulus material is its sex role orientation. As Goldberg (1968) concluded, biases against women may be stronger in fields incongruent with their gender.

Methodological Differences Across Studies

Research design. Strength of findings in studies may vary as a result of different characteristics of the study designs. For instance, whether subjects rated both male and female target persons or just a male or just a female target person may affect ratings. That is, rating both male and female target persons might sensitize subjects to the purpose of the study, which could decrease the likelihood that they would differentially rate the target persons. On the other hand, rating both male and female target persons may cause subjects to contrast the target persons and thus perceive greater differences between male and female target persons.

Stimulus material. A second methodological characteristic is the amount of information given to the subjects about the target person. Nieva and Gutek (1980) stated, "The more task-related information provided about the 'evaluatee' and the greater the clarity about the criteria to be used in the evaluation situation, therefore, the less likely it is that 'actuarial prejudice' will operate" (p. 273; see also Futoran & Wyer, 1986). Similarly, Locksley, Borgida, Brekke, and Hepburn (1980) found that sex stereotypes, conceptualized as probabilistic base rates, were less likely to be used when specific information in addition to gender was presented. Thus, more information about the target person might yield less discriminatory ratings.

The different behaviors or products that were used as stimulus material can be placed into general categories such as behaviors or applications for jobs. Differences in ratings of men and women may be more pronounced in one category than in another. For instance, subjects may be less cautious about using their gender stereotypes when the stimulus materials are essays than when they are job applications.

Dependent variables. The extent to which stereotypes are used when making judgments may differ depending on the particular dependent variables in use (e.g., Bodenhausen & Wyer, 1985). For instance, Whitley and Frieze (1986) found differences in attributions for achievement based on question wording. One difference in question wording in the literature reviewed here involves whether subjects rated the target person or the stimulus materials. Rating the target person might make the gender of the target person more salient than would rating the stimulus material. In the same manner that having less information about a target may result in more stereotypic responses, increasing the salience of the target person's gender might yield a greater tendency to discriminate against women. On the other hand, greater saliency might increase the tendency to give a socially desirable response and therefore decrease the likelihood of a discriminatory response.

Quality of research. A more general methodological difference in the studies that might influence the strength of the findings is the quality of the research. Meta-analyses have been criticized for giving equal weight to both well-conducted and poorly conducted studies (e.g., Eysenck, 1978; Slavin, 1986). Like all reviews, meta-analysis results are influenced by the quality of the primary sources. For instance, previous meta-analyses have shown that the strength of the summary statistics is influenced by the inclusion or exclusion of studies that only allow the reviewer to assign nominal levels of significance to effects (e.g., Eagly & Carli, 1981). Other deficiencies in reporting—such as completeness, accuracy, and clarity—have also been shown to influence meta-analyses (Orwin & Cordray, 1985).

Publication of Articles

Sex of author. The last set of variables that might influence the strength of discriminatory responses involve characteristics of the articles themselves. One variable associated with the articles is the sex of the authors. Eagly (1978) and Eagly and Carli (1981), in reviews of gender differences in influenceability, concluded that male authors were more likely than female authors to find that women were more easily influenced than men. This conclusion has not, however, been reached in other meta-analyses of gender differences (e.g., Eagly & Crowley, 1986; Eagly & Steffen, 1986).

Publication year. A second difference in the articles is the year in which they were published. Publication year has been shown to be related to findings regarding gender differences in cognitive abilities (Rosenthal & Rubin, 1986), influenceability (Eagly & Carli, 1981), and helping behavior (Eagly & Crowley, 1986). In the case of helping behavior, however, publication year was found to be correlated with a number of other study characteristics (Eagly & Crowley, 1986), and for aggressive behavior, the correlation with publication year was not significant (Eagly & Steffen, 1986). Given the apparent advances in women's rights and the decrease in people's willingness to endorse unequal treatment of men and women, it seems important to examine the relationship between ratings of target people and year of publication for the present research literature as well.

In summary, variables examined in previous research were reviewed, and for each variable, pertinent hypotheses were identified for purposes of this meta-analysis. Furthermore, methodological factors that bear on the robustness of the Goldberg (1968) phenomenon were reviewed and are examined in the meta-analysis presented in the next section. Finally, different characteristics of the articles were identified as possible sources of variation in the strength of the effects.

Method

Literature Search

The data base of articles was compiled by locating articles referenced in the Social Science Citation Index, by a computer search of the Psychological Abstracts (PsycINFO), and by cross-referencing various articles and reviews of the literature on sex discrimination. From 1968 to 1985, the Social Science Citation Indices contained 322 citations that referenced P. or P. A. Goldberg (1968), making this study a "citation classic" by *Current Content*'s standards. The PsychINFO computer search of abstracts from 1968 to 1985, using the document title *Goldberg Replication*, yielded 103 abstracts. Clearly, this is a well-known and influential study. The articles and chapters that reviewed the experimental literature were Arvey (1979), Nieva and Gutek (1980), Wallston and O'Leary (1981), Basow (1986), and Tosi and Einbender (1985). Additionally, the reference sections of the articles included in this meta-analysis were examined.

The following selection criteria were used in order to form a relatively homogeneous sample that represented conceptual replications of Goldberg's original study:

1. The vast majority of empirical papers were published in North American English-language journals. Thus, only studies conducted in North America and published from 1968 to 1985 were included.

2. Only published articles were included. Although this may produce a biased sample, this limitation was imposed primarily because most reviews indicate that unpublished articles are less likely to report significant findings (Glass, McGaw, & Smith, 1981; Green & Hall, 1984; Greenwald, 1975; Lane & Dunlap, 1978). This tendency was supported in the unpublished papers we obtained through correspondence with a few investigators. Thus, exclusion of unpublished studies would yield findings that maximize the size of effects in the present meta-analysis. Alternatively, it could be argued that there was a bias against publishing differential evaluations of men and women. However, such a publication bias seemed unlikely for this topic because articles and reviews published during the time period covered tended to emphasize significant findings.

3. Only studies in which subjects rated hypothetical adult target persons were included. This excluded studies in which subjects rated children, actual adults such as subjects' supervisors or teachers, occupational categories, sex stereotypical traits, or men and women "in general," and excluded salary comparisons, differential rates of publications, and self-ratings. In addition to keeping the sample homogeneous, this also insured that there were not actual differences among the target persons.

4. The subject had to have been given information about a behavior produced by the target person, such as a job performance or articles written by the target person. That is, studies were excluded when subjects rated a target person on the basis of only a name or a name and an occupational title. If a subject was given a job application, resumé, or similar set of materials, these documents were assumed to include indicators of past behaviors, and so these studies were included. Additionally, this criterion excluded studies that had subjects decide if hypothetical psychotherapy clients were psychologically healthy. However, this criterion did not exclude studies that had subjects evaluate hypothetical counselors' or clinicians' job performances.

5. Studies primarily exploring different topic areas other than differential evaluations of men and women were excluded. In particular, this excluded persuasion, attribution, and attraction studies that were not directly related to our focus on the Goldberg paradigm. If a study primarily dealt with evaluations but included a few attributional or attraction-dependent variables, it was included, but the dependent variables measuring attribution or attraction were excluded.

The search and selection criteria resulted in 123 studies from 106 articles published primarily between 1974 and 1979 (see Appendix A). The trend in publishing the replications resembles a normal distribution (see Figure 1), perhaps reflecting the established tendency for certain research themes in social psychology to generate considerable research over a relatively brief period of time (Jones, 1985).

Coding of Studies

Both main effects and simple effects reflecting the comparisons between male and female target persons from two-way interactions were coded for each dependent variable. Higher order interactions were not included. Mean effect sizes (d) and voting scores (a tally of the number of significant and insignificant results reported) were calculated for both the main effects and the simple effects. (Appendix B contains a more detailed description of our assumptions, and Appendix C contains a more detailed description of the calculations.) A negative value for the effect sizes represents a more favorable rating given to men, and a positive value represents a more favorable rating given to women.

Effect sizes were weighted inversely by their variances. This gives more weight to studies with smaller variances and thus yields a more precise estimate of d (see Hedges & Olkin, 1985). Because the weighted d is a more precise estimate, only the weighted ds are reported. However, it should be noted that because the majority of the effect sizes were small, there was little difference between the weighted and the unweighted effect size.

In addition to the main effects, the following variables were coded:

Subject variables. These included (a) sex of subjects (women only, men only, or both men and women), (b) the Bem Sex Role Inventory, and (c) age of the subjects (high school, undergraduates or graduates, older than college age, more than one age group represented in the study). If the sample contained more than one age group and differences between age groups were reported, separate results were recorded for each age group. These two results were coded as two different studies. If the differences were not significant between age groups and if separate data for the age groups were not reported, identical effect sizes and voting scores were recorded for each finding.

Target persons. (d) Competence of the target person (high, medium, or low) was determined by the level reported in the article. If the study only had two levels, no value was recorded for the medium competence level. (e) Attractiveness of the target person was determined by the level reported in the article. If the study only had two levels, no value was recorded for the medium attractiveness level. Although there are no absolute standards of competence and attractiveness, it was assumed that using the levels reported would yield distributions of high, medium, and low levels whose means would differ in the same order. (f) Race of the target person (White or non-White) was also coded.

Stimulus materials. (g) Quality of the stimulus material was determined by the level reported in the article and again was assumed to represent a distribution of high, medium, and low levels. If the study only had two levels, no value was recorded for the medium quality level. (h) Sex role orientation of the stimulus material (feminine, masculine, or sex neutral) was also coded.

Methodological characteristics. These included (i) number of target persons evaluated and (j) amount of information about the target person given to the subjects (only name, one sentence, one paragraph, more than one paragraph, or resumé/application). Also coded was whether the subjects saw a picture of the target person, viewed a film or video of the target person, or heard a tape recording of the target person. In the analyses, if the target person was represented in these forms, the amount of information about the target person was considered to be *more than one paragraph*. Additionally, the category of *resumé/application* was collapsed with the category of *more than one paragraph*. Thus, only four levels of amount of information about the target person were analyzed. (k) The number of independent variables manipulating differences in the target person across conditions was also used as a second method of approximating the amount of information about the target person. (l) Six categories were formed for choice of the stimulus material: written works (e.g., an article, memo, speech, or artwork); behaviors; job applications or resumés; biographical descriptions; applications and essays; and other. (m) Question wording was defined as whether the dependent variable asked for ratings of the target person or ratings of the stimulus material or both. (n) The variable *journal quality* was used as an indirect measure of the quality of the research. Journal ratings were obtained by averaging ratings of these journals by five social psychologists at the University of Minnesota.

Study characteristics. These included (o) sex of the first author, (p) percent of female authors, and (q) publication year.

Analyses

Effect sizes were treated in three different ways. First, the mean weighted *d*s were calculated across all the findings. Second, the mean weighted *d* was calculated after excluding studies in which effect sizes and voting scores were assigned a value of zero because of insufficient information. Third, mean effect sizes, including all findings, were calculated within a study, and then a mean of these mean effect sizes was calculated.

Findings with insufficient information were from studies that (a) did not mention the significance of the finding or (b) reported the finding as nonsignificant but did not report sufficient information to calculate the *d* or to estimate the *z* score. When included, these findings are assumed to be randomly distributed around a mean of zero. It should be noted that studies that reported findings as significant but did not report other information needed for calculations were included in this second calculation (see Appendix C). It could be argued that including all findings (i.e., those with sufficient and insufficient information) results in a bias toward nonsignificant findings and that excluding findings assumed to be equal to zero results in a bias toward more significant findings. Thus, the results including findings assumed to be equal to zero can be considered a lower limit, and those excluding these findings can be considered an upper limit.

The first two methods of combining *d*s accords more weight to studies with more dependent variables, whereas the third method weights studies equally by weighting each finding inversely by the number of findings per study. The unit of analysis in the first two approaches is findings, and the unit of analysis in the third approach is studies. Because results reported within a study are not independent, Rosenthal and Rubin (1986) recommended that the study be used as the unit of analysis by combining the findings, with the correlation between the findings within the studies taken into account. However, in this meta-analysis, only seven studies reported these correlations. Furthermore, the calculations assume homogeneous correlations between measures that did not seem likely in the present research. Therefore, finding a mean *d* for each study seemed to be the most reasonable way to adjust for the interdependence between measures in this set of studies.

Because the distribution of effect sizes is not likely to meet assumptions required for conventional statistical tests, alternative tests recommended by Hedges and Olkin (1985) were used. Specifically, Hedges and Olkin provided formulas for calculating the variance of the effect sizes and confidence intervals. Additionally, they provided formulas for calculating the homogeneity of the effect sizes. The homogeneity tests allow one to search for outliers and to test whether the effect sizes are uniform. If effect sizes are heterogeneous, they may be able to be categorized into smaller groups that are homogeneous. Furthermore, once ho-

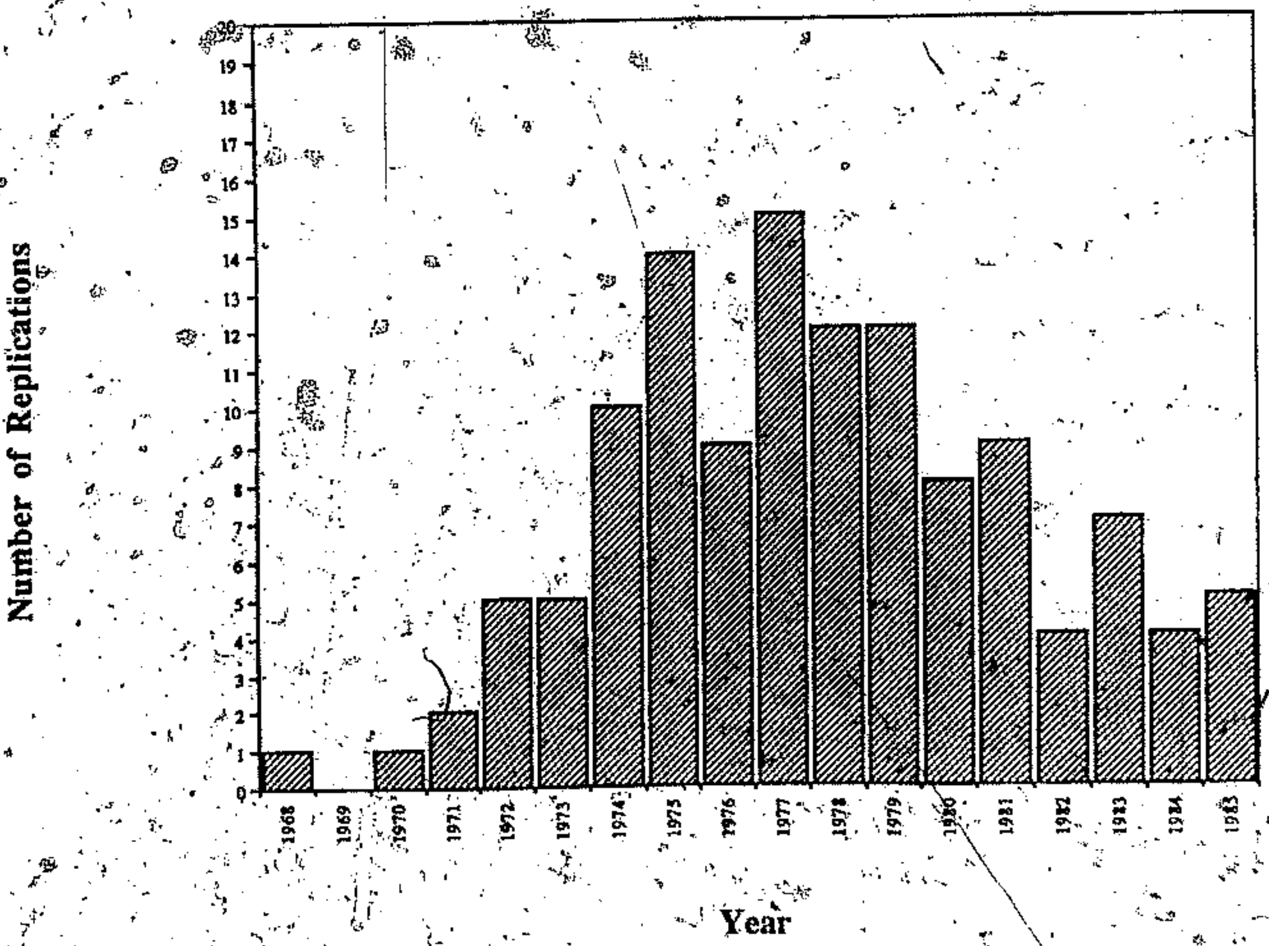

Figure 1. Number of replications per year.

mogeneous groups are found, these groups can be compared to see if their average effect sizes differ.

On the basis of the homogeneity test, three findings were determined to be outliers and were removed from the data. These three findings came from results reported by Rosen and Jerdee (1974) and Rosen, Jerdee, and Prestwich (1975). Although the findings were large (d = −.85, −.74, and .62), they were not the largest effect sizes. The findings were outliers not only because the effect sizes were among the largest, but also because the study had a large sample size (N = 1,476). Because of the large number of findings in this review, including or excluding these three findings does not appreciably change the mean weighted effect size. However, it does increase the likelihood that the findings will be more homogeneous. It should be noted that other studies using similar stimulus material were included in this review. Some of the findings from these studies were also fairly large; however, they were not outliers, and therefore they were not excluded from the data.[1]

In addition to calculating mean weighted *d*s and voting scores for the main effects and simple effects reported in the articles, correlations between the mean weighted main effect *d* and continuous variables were examined. Differences in the mean weighted main effect *d*s were examined across noncontinuous variables, as well as across continuous variables put into the form of categorical data, to test for nonlinear relationships. Significance of the correlations is calculated according to the recommendations outlined in Hedges and Olkin (1985).

Results

First, descriptions are given of the articles, subjects, and procedures reviewed. Second, the mean weighted main effect *d*s and simple effect *d*s and the voting scores are reported. Finally, correlations and other analyses with the mean main effect *d*s are reported.

Articles

Of the 106 articles reviewed with a total of 595 findings, 12 of the articles reported two experiments, 1 reported three experiments, and 1 reported four experiments. This yields 123 studies. Seventeen percent of the studies were published in "high-quality" journals, 58% in "medium-quality" journals, and 25% in "low-quality" journals. Both the median and the mode of the publication year are 1977. The distribution resembles a normal distribution. Gender of the first author could not be determined for two studies. Of the remaining 121 studies, 55% of first authors were women. The percent of female authors was indeterminate for four studies. In the remaining studies, 38% had less than 50% female authors, 22% had 50% female authors, and 40% had more than 50% female authors. Thirty-two percent had only male authors, and 35% had only female authors.

Subjects

These 123 studies engaged 21,379 subjects with the number of subjects per study ranging from 20 to 3,261. When we excluded three studies that had more than 1,000 subjects, the mean number of subjects was 173, and the median was 96. Nine percent of the studies had only female subjects, 17% had only male subjects, and 72% had both male and female subjects. Of the 121 studies in which the age of the subjects was reported or could be inferred, 4% of the studies involved subjects who were younger than college students, 74% involved college-aged subjects, and 21% involved subjects older than college students.

Procedures

Forty-five percent of the studies had subjects evaluate only a male or a female target person, and the remaining 55% had subjects rate more than one target person. In addition to the stimulus material supposedly produced by the target person, the amount of information given to subjects about the target person ranged from only a name (24%); to a one-sentence description (5%); to a paragraph of information (11%); to more than a paragraph, a resumé, or an application, showing a picture, a video, or the actual person or hearing an audio tape (60%). Twenty-three percent of the studies had stimulus materials that could be characterized as written work or art work, 30% as behaviors, 11% as resumés or applications, and 11% as biographies. Six studies had both an application and an essay, and seven studies had stimulus materials categorized as *other* (6%).

Of the 595 findings across the studies, 81% were ratings of the target person, 14% were ratings of the stimulus material, and 4% reported results by combining the dependent variables, making it impossible to determine how many variables were ratings of the target person or of the stimulus materials.

Main and Simple Effects

Main effects. Table 1 presents the results for the main effect of ratings of male versus female target persons. Four studies, representing 20 findings, did not analyze for the main effect; in these cases, no assumptions about the effect sizes could be made to accurately estimate the *d* and voting scores for the main effects (in contrast to the simple effects discussed later). Thus, the results for the main effect are based on 119 studies and 575 findings. The number of findings in which the *d* and the voting score were assigned to zero were 360 (63%). When these values were excluded, 215 findings remained to be analyzed.

Some studies had only male or female subjects or only masculine, feminine, or sex-neutral stimulus materials. The main effects from these studies would be similar to the simple effects from studies that had sex of subject or sex role of stimulus material as independent variables. Hence, the main effects were calculated including and excluding these studies. The results were essentially identical. The main effects presented in Table 1 include these studies.

It may be seen that the main effects were all negative and all very small. A negative value indicates a lower rating of female than male target persons. The values for the homogeneity test were compared to chi-square distributions with degrees of free-

[1] The first two effect sizes (−.85 and −.75) come from subjects' ratings of recommended procedures for an employee who was offered a job in a different company. The two procedures were "Try to convince operator to remain with organization" and "Don't try to influence operator." The negative effect sizes indicate that the responses were more favorable for the male target person than the female target person. The third effect size (.62) comes from subjects rating a target person's request for leave of absence because of family obligations as being more appropriate for female target people than for male target people.

Table 1
Main Effect: Ratings of Male Target Persons Versus Ratings of Female Target Persons

Summary statistic	Study as unit of analysis	Finding as unit of analysis	Excluding findings assumed equal to zero
Effect size weighted by variance			
N	119	575	215
M ± 95% CI	−.07 ± .03	−.05 ± .01	−.08 ± .01
Homogeneity test	303*	1,394*	1,315*
Voting score			
% female	7	8	20
% male	20	17	45
% neither	73	76	35

Note. Negative effect size indicates lower evaluations of women. CI = confidence interval.
* $p < .05$, that is, effect sizes are heterogeneous.

Table 2
Effect Size Weighted by Variance for Simple Effects From Interaction: Sex of Subject × Sex of Target Person

Summary statistic	Study as unit of analysis	Finding as unit of analysis	Excluding findings assumed equal to zero
Female subjects			
N	91	410	145
M ± 95% CI	−.02 ± .04	−.02 ± .02	−.06 ± .03
Homogeneity test	82	385	370
Male subjects			
N	99	479	188
M ± 95% CI	−.06 ± .03	−.04 ± .01	−.07 ± .02
Homogeneity test	151*	888*	853*

Note. Negative effect size indicates lower evaluations of women. CI = confidence interval.
* $p < .05$, that is, effect sizes are heterogeneous.

dom equal to $n - 1$. The significance of these tests indicated that the main effect *d*s were heterogeneous.

An accepted rule of thumb is that a *d* of .80 is considered large, .50 is moderate, and .20 is small (Eagly, 1987). As a comparison, average *d*s for sex differences have been found to range from small *d*s such as .09 to .32 for influenceability (Eagly & Carli, 1981) to large *d*s such as 2.18 for sex differences in the motor performance of throwing velocity (Thomas & French, 1985).

When we compared the different ways of combining the results across studies, the main effect *d* was smallest when findings were the unit of analysis (−.05), largest when *d*s assumed to equal zero were excluded (−.08), and intermediate when calculated with study as the unit of analysis (−.07).

The voting scores, the tally of significant and nonsignificant findings, showed that when study or finding with values assumed to be equal to zero is the unit of analysis, more than 7 in 10 results were nonsignificant. When the findings were significant, they were more likely to indicate that the man had been rated more favorably. Sixty-three percent of the findings had to be assumed to be equal to zero. When these values are excluded, 45% of the results indicated that male target persons were preferred to female target persons. It should be noted again that although this excluded studies in which the only information reported about the finding was that it was nonsignificant, it included studies in which the only information reported about the finding was that it was significant.

Simple effects. Tables 2 through 15 present the results for the simple effects. The number of findings and studies per simple effect decreased from Table 2 to Table 15. Smaller *n*s decrease the reliability and power of the analysis (Strube & Miller, 1986). As for the main effect *d*s, the size of the effect sizes and voting scores for the simple effects tended to be smallest when finding was the unit of analysis, to be intermediate when study was the unit of analysis, and to be largest when values assumed to be equal to zero were excluded. The voting scores showed that the majority of findings and studies reported nonsignificant differences. When we excluded values assumed to be equal to zero, the majority of simple effects indicated that men were rated higher than women. Unlike the main effect *d*, some of the simple effects were homogeneous.

Sex of subject. Contrary to predictions of greater bias among male subjects, little difference was found between male and female subjects' ratings of male and female target persons (see Table 2). However, the findings for male subjects were substantially more heterogeneous than for female subjects.

Sex role orientation of stimulus material. As was mentioned earlier, some studies only had one type of stimulus material. Such studies kept the sex role orientation of the stimulus material constant across subjects by presenting all subjects with only masculine, only feminine, or only sex-neutral stimulus material. The simple effects for sex role orientation of the stimulus material were calculated both excluding these studies and including these studies. The results were essentially identical whether or not these studies were included. Table 4 presents the results with these studies included.

Table 3
Voting Score for Simple Effects From Interaction: Sex of Subject × Sex of Target Person

Summary statistic	Study as unit of analysis	Finding as unit of analysis	Excluding findings assumed equal to zero
Female subjects			
N	91	410	145
% female	9	9	25
% male	15	13	36
% neither	76	78	39
Male subjects			
N	99	479	188
% female	7	8	21
% male	22	17	42
% neither	70	75	37

Table 4
Effect Size Weighted by Variance for Simple Effects From Interaction: Sex Role Orientation of Stimulus Material × Sex of Target Person

Summary statistic	Study as unit of analysis	Finding as unit of analysis	Excluding findings assumed equal to zero
Feminine sex role			
N	44	202	73
M ± 95% CI	−.01 ± .05	−.03 ± .03	−.08 ± .04
Homogeneity test	75*	403*	391*
Masculine sex role			
N	46	226	81
M ± 95% CI	−.12 ± .05	−.10 ± .02	−.25 ± .04
Homogeneity test	75*	424*	299*
Sex-neutral sex role			
N	28	106	45
M ± 95% CI	−.13 ± .06	−.15 ± .03	−.32 ± .04
Homogeneity test	64*	301*	185*

Note. Negative effect size indicates lower evaluations of women. CI = confidence interval.
* $p < .05$, that is, effect sizes are heterogeneous.

All the methods of combining results showed, consistent with predictions, that a greater difference in ratings of male and female target persons (with men being rated more favorably) occurred when the stimulus materials were masculine in comparison with when the stimulus materials were feminine. However, the largest effect size occurred when the stimulus material was sex neutral. Furthermore, it should be kept in mind that the *ds* were still small. The distribution of the effect sizes was heterogeneous for the feminine, masculine, and sex-neutral sex roles.

Table 5
Voting Score for Simple Effects From Interaction: Sex Role Orientation of Stimulus Material × Sex of Target Person

Summary statistic	Study as unit of analysis	Finding as unit of analysis	Excluding findings assumed equal to zero
Feminine sex role			
N	44	202	73
% female	22	16	45
% male	16	12	33
% neither	62	72	22
Masculine sex role			
N	46	226	81
% female	7	8	22
% male	36	21	59
% neither	57	71	18
Sex-neutral sex role			
N	28	106	45
% female	1	4	9
% male	23	27	64
% neither	75	69	27

Table 6
Effect Size Weighted by Variance for Simple Effects From Interaction: Status of Target Person × Sex of Target Person

Summary statistic	Study as unit of analysis	Finding as unit of analysis	Excluding findings assumed equal to zero
High competence			
N	30	95	32
M ± 95% CI	−.08 ± .07	−.05 ± .04	−.21 ± .08
Homogeneity test	56*	149*	130*
Medium competence			
N	6	14	5
M ± 95% CI	−.18 ± .20	−.13 ± .13	−.34 ± .22
Homogeneity test	4	7	1
Low competence			
N	29	94	31
M ± 95% CI	−.07 ± .06	−.04 ± .04	−.19 ± .08
Homogeneity test	21	74	57

Note. Negative effect size indicates lower evaluations of women. CI = confidence interval.
* $p < .05$, that is, effect sizes are heterogeneous.

Competence of target person. Six studies (14 findings) included medium-competent target persons in addition to high- and low-competent target persons (see Table 6). Because of the small number of studies and findings with a target person of medium competence, it was more meaningful to examine the differences between evaluations of male and female target persons who were high and low in competence.

The *ds* were essentially identical whether the target person was of high or of low competence (see Table 6). However, the *ds* were heterogeneous for target persons of high competence and

Table 7
Voting Score for Simple Effects From Interaction: Status of Target Person × Sex of Target Person

Summary statistic	Study as unit of analysis	Finding as unit of analysis	Excluding findings assumed equal to zero
High competence			
N	30	95	32
% female	8	7	22
% male	17	13	41
% neither	75	80	41
Medium competence			
N	6	14	5
% female	53	0	0
% male	47	43	100
% neither	0	57	0
Low competence			
N	29	94	31
% female	22	5	16
% male	2	15	45
% neither	76	80	39

Table 8
Effect Size Weighted by Variance for Simple Effects From Interaction: Status of Stimulus Material × Sex of Target Person

Summary statistic	Study as unit of analysis	Finding as unit of analysis	Excluding findings assumed equal to zero
High quality			
N	10	38	10
M ± 95% CI	−.02 ± .11	.01 ± .05	−.16 ± .17
Homogeneity test	30*	55*	51*
Medium quality			
N	2	7	3
M ± 95% CI	−.24 ± .34	−.19 ± .18	−.87 ± .37
Homogeneity test	5	17*	0
Low quality			
N	10	37	14
M ± 95% CI	−.05 ± .11	−.06 ± .05	−.22 ± .10
Homogeneity test	6	34	22

Note. Negative effect size indicates lower evaluations of women. CI = confidence interval.
* $p < .05$, that is, effect sizes are heterogeneous.

Table 10
Effect Size Weighted by Variance for Simple Effects From Interaction: Physical Attractiveness of Target Person × Sex of Target Person

Summary statistic	Study as unit of analysis	Finding as unit of analysis	Excluding findings assumed equal to zero
High physical attractiveness			
N	12	62	11
M ± 95% CI	−.15 ± .12	−.07 ± .05	−.32 ± .12
Homogeneity test	9	38	16
Medium physical attractiveness			
N	5	19	4
M ± 95% CI	−.11 ± .17	−.09 ± .09	−.27 ± .16
Homogeneity test	2	19	12
Low physical attractiveness			
N	12	62	12
M ± 95% CI	−.05 ± .12	−.02 ± .05	−.08 ± .11
Homogeneity test	7	40	38*

Note. Negative effect size indicates lower evaluations of women. CI = confidence interval.
* $p < .05$, that is, effect sizes are heterogeneous.

homogeneous for target persons of low competence. And again, the voting scores indicated that the majority of findings and studies found no significant differences and that the *ds* were small.

Only two studies (seven findings) included medium-quality stimulus material (see Table 8). Again, it was more meaningful to examine just the high- and low-quality conditions. As with the simple effects for competence, the sizes of the *ds* for high and low quality were quite similar. Also, the *ds* for the high-quality stimulus materials were more heterogeneous than the *ds* for the low-quality stimulus materials.

Physical attractiveness of target persons. Only five studies (19 findings) had a medium level of physical attractiveness. Hence, comparisons were only made between target persons who were high and who were low in physical attractiveness (see Tables 10 and 11). All but one of the results indicated that the findings were homogeneous. Contrary to the prediction that gender

Table 9
Voting Score for Simple Effects From Interaction: Status of Stimulus Material × Sex of Target Person

Summary statistic	Study as unit of analysis	Finding as unit of analysis	Excluding findings assumed equal to zero
High quality			
N	10	38	10
% female	10	3	10
% male	12	10	40
% neither	78	87	50
Medium quality			
N	2	7	3
% female	0	0	0
% male	50	43	100
% neither	50	57	0
Low quality			
N	10	37	14
% female	0	0	0
% male	12	10	29
% neither	88	90	71

Table 11
Voting Score for Simple Effects From Interaction: Physical Attractiveness of Target Person × Sex of Target Person

Summary statistic	Study as unit of analysis	Finding as unit of analysis	Excluding findings assumed equal to zero
High physical attractiveness			
N	12	62	11
% female	1	2	10
% male	38	18	91
% neither	61	81	0
Medium physical attractiveness			
N	5	19	4
% female	3	5	25
% male	20	16	75
% neither	77	79	0
Low physical attractiveness			
N	12	62	12
% female	10	5	25
% male	25	11	58
% neither	65	84	17

Table 12
Effect Size Weighted by Variance for Simple Effects From Interaction: Bem Sex Role Inventory of Subjects × Sex of Target Person

Summary statistic	Study as unit of analysis	Finding as unit of analysis	Excluding findings assumed equal to zero
Feminine subjects			
N	1	14	3
M ± 95% CI	−.16 ± .64	−.15 ± .17	−.76 ± .38
Homogeneity test	0	12	0
Androgenous subjects			
N	2	24	8
M ± 95% CI	−.17 ± .33	−.16 ± .10	−.35 ± .15
Homogeneity test	0	25	14

Note. Negative effect size indicates lower evaluations of women. CI = confidence interval.

would influence evaluations when the target persons were unattractive but not when they were attractive, there was a tendency for the effect sizes to be larger for more attractive target persons. Again, the effects were small and similar for target persons who were high and who were low in physical attractiveness.

Bem Sex Role Inventory. The number of studies and findings assessing the effect of sex role orientation of the subjects was quite small, and hence the reliability of these results is questionable. No studies had masculine subjects, and only one study (one finding) had undifferentiated (i.e., neither feminine, masculine, or androgenous) subjects. Hence results were only reported for the feminine and androgenous subjects. Contrary to expectations, feminine subjects were more likely to give more favorable ratings to female target persons, and androgenous subjects were more likely to give more favorable ratings to male target persons. However, these effect sizes were still quite small and came from a limited number of studies. So, although they were homogeneous, the results may not be reliable.

Table 13
Voting Score for Simple Effects From Interaction: Bem Sex Role Inventory of Subjects × Sex of Target Person

Summary statistic	Study as unit of analysis	Finding as unit of analysis	Excluding findings assumed equal to zero
Feminine subjects			
N	1	14	3
% female	14	14	67
% male	0	0	0
% neither	86	86	33
Androgenous subjects			
N	2	24	8
% female	0	0	0
% male	25	21	62
% neither	75	80	38

Table 14
Effect Size Weighted by Variance for Simple Effects From Interaction: Race of Target Person × Sex of Target Person

Summary statistic	Study as unit of analysis	Finding as unit of analysis	Excluding findings assumed equal to zero
White			
N	7	14	4
M ± 95% CI	−.003 ± .14	−.01 ± .11	−.03 ± .17
Homogeneity test	16*	18	18
Non-White			
N	7	14	4
M ± 95% CI	−.11 ± .12	−.08 ± .09	−.16 ± .13
Homogeneity test	4	6	4

Note. Negative effect size indicates lower evaluations of women. CI = confidence interval.
* $p < .05$, that is, effect sizes are heterogeneous.

Race of target person. The number of findings and studies was also quite small for the simple effects associated with race of the target person (see Tables 14 and 15). There appeared to be a slight tendency for comparisons between non-Whites to yield larger effect sizes than comparisons between Whites. Yet again, these effect sizes were small, and the findings were based on a limited number of studies.

Correlations

Correlations of the main effect *ds* with continuous variables are presented in Table 16. These correlations are presented in the form of standardized regression coefficients (see Hedges & Olkin, 1985). Positive coefficients indicate that gender bias toward women decreases as the other variables increase. None of the coefficients were significant, and none were large.

Breakdown of Effect Sizes

Main effect *ds* were calculated for different categorical levels of a series of independent variables. The breakdown of the

Table 15
Voting Score for Simple Effects From Interaction: Race of Target Person × Sex of Target Person

Summary statistic	Study as unit of analysis	Finding as unit of analysis	Excluding findings assumed equal to zero
White			
N	7	14	4
% female	14	7	25
% male	36	21	75
% neither	50	71	0
Non-white			
N	7	14	4
% female	14	7	25
% male	21	14	50
% neither	64	79	25

Table 16
Standardized Regression Coefficients for Relating Variables to Weighted Effect Sizes

	Study as unit of analysis			Finding as unit of analysis			Excluding findings assumed equal to zero		
Variable	beta	N^a	t^b	beta	N^a	t^b	beta	N^a	t^b
% female authors	.00	115	0	.07	553	.16	.07	200	.16
Publication year	.00	119	0	.06	575	.15	.10	215	.20
Number of independent variables associated with target person	.00	119	0	.08	575	.18	.14	215	.24
Number of target persons rated	.00	119	0	−.06	575	.16	−.11	215	−.21

Note. Negative effect sizes indicate lower evaluations of women.
[a] Sample size. [b] *t* test.

weighted *d*s by the type of question wording was only done when finding was the unit of analysis. Both within-group and between-groups homogeneity tests were calculated. This allows one to examine whether the effect sizes within each category can be considered identical and whether the effect sizes across categories were different. These results are presented in Tables 17, 18, and 19.

All of the effect sizes within the different categories remained small, and most were heterogeneous. Even though the effect sizes were heterogeneous, a few of the between-groups homogeneity tests were significant, indicating that there were differences between a few of the categories. However, none of the effect sizes were large.

The categories for age of subject differed, but the effect sizes within the categories were heterogeneous. The pattern of results differed depending on whether finding or study was the unit of analysis. When finding was the unit of analysis, with or without zero values, a greater difference in ratings was found for younger subjects; when study was the unit of analysis, greater differences were found for older subjects.

Like the age of the subjects, the categories for percentage of female authors differed, but the effect sizes were heterogeneous within categories. The trend was for greater differences in ratings when there were more than 50% female authors. Similarly, greater differences in ratings were found when the first author was a woman. However, though the categories differed when finding was the unit of analysis, the effect sizes were again heterogeneous within categories for all combinatorial methods.

The categories for year of publication were only heterogeneous when values assumed to equal zero were excluded and tended to be heterogeneous within a category. The trend was to find fewer differences in ratings in the middle years than in the earlier or later years. The categories for journal quality were heterogeneous, but the effect sizes within categories were also heterogeneous. The trend was to find greater differences in ratings when the journals were of either high or low quality.

The categories differ for the amount of information and a few of the categories were homogeneous. The trend was to find larger differences when less information was given. The number of independent variables was also used as a measure of amount of information. The same trend of greater differences with less information was found here. However, the categories are heterogeneous, and as mentioned previously, the correlation between number of independent variables and effect size is small.

The number of target persons rated had little effect on ratings. The categories for type of stimulus material differed, though most of the effect sizes within categories were heterogeneous. There tended to be greater differences when applications were used as the stimulus material. The question wording had little effect on the effect sizes.

Discussion

To assess the robustness of findings from Goldberg's (1968) original study, we reviewed studies that were paradigmatically similar. The task situation was typically one in which subjects were given some evidence about a hypothetical target person's abilities, for example, in the form of an essay purportedly written by the target person or a target person's job application. The subject was asked to evaluate the target person, the product that the target person had generated, or both. Some subjects were only told that the target person was a man; others were told that the target person was a woman. In other studies, subjects evaluated more than one target person, some of whom were men and some of whom were women.

One hundred and six published articles using North American subjects and reporting on 123 studies that used this experimental paradigm were selected and reviewed. About three-fourths of these studies were published between 1974 and 1981. There were approximately equal numbers of male and female authors. Most of the studies used college students as well as both male and female subjects. About half of the studies had subjects evaluate only male or only female target persons, and the other half had subjects evaluate both male and female target persons. The stimulus materials were approximately equally likely to be written work or art work, behavior, résumés, applications for jobs, or short biographies of the person.

Differences between the average ratings of men and women were analyzed across studies by tallying voting scores and by combining effect sizes. The unit of analysis was either the study or the finding. The study is the best unit of analysis because within each study the dependent variables are likely to be interdependent. An upper limit on size of the difference was obtained by using finding as the unit of analysis and excluding those findings in which no difference between men and women had to be assumed because of limited reporting in the studies.

We found that the size of the difference in ratings between female and male target persons was extremely small (−.07, accounting for less than 1% of the variance, with study as the unit of analysis, and −.05, accounting for less than 1% of the variance, with finding as the unit of analysis). Excluding those findings for which we had insufficient information (i.e., values were assumed to be equal to zero), the effect size only reached a value of −.08, clearly less than 1% of the variance. This difference is even smaller than those typically reported for gender differences in social behavior (e.g., Eagly & Carli, 1981; Hyde & Linn, 1986; see Deaux, 1985). Furthermore, when merely examining

Table 17
Breakdown of Main Effect Size Weighted by Variance (Study as Unit of Analysis)

Variable	*M* ± 95% CI	*N*	Homogeneity within group	Homogeneity between groups
Age of subject				
Younger than college age	−.01 ± .15	5	5	7**
College age	−.07 ± .03	85	254*	
Older than college age	−.11 ± .05	23	42*	
% female authors				
<50%	−.08 ± .04	42	82*	16**
50%	−.01 ± .05	26	73*	
>50%	−.12 ± .04	47	132*	
Sex of first author				
Female	−.09 ± .04	65	147*	2
Male	−.06 ± .04	52	154*	
Publication year				
1968–1973	−.10 ± .08	14	12	5
1974–1978	−.10 ± .04	60	198*	
1979–1985	−.04 ± .04	45	88*	
Journal quality				
High	−.14 ± .06	21	38*	8**
Medium	−.04 ± .03	68	128*	
Low	−.11 ± .05	30	129*	
Amount of information about target person				
Only name	−.12 ± .06	27	38	37**
One sentence	−.05 ± .13	6	5	
One paragraph	−.001 ± .06	13	16	
More than one paragraph	−.08 ± .03	73	207*	
Number of independent variables associated with target person				
One	−.13 ± .04	54	148*	10
More than one	−.04 ± .03	65	145*	
Number of target persons rated				
One	−.07 ± .04	54	152*	0
More than one	−.08 ± .03	65	151*	
Type of stimulus material				
Written work	−.04 ± .05	28	28	45**
Behavior	−.07 ± .05	34	155*	
Application	−.10 ± .06	29	19	
Biography	+.03 ± .08	13	26*	
Application and essay	−.04 ± .09	6	10*	
Other	−.16 ± .11	7	20*	

Note. Negative effect sizes indicate lower evaluations of women. CI = confidence interval.
* $p < .05$, that is, effect sizes are heterogeneous within category.
** $p < .05$, that is, average effect sizes are significantly different between categories.

the number of studies that found significant and nonsignificant results, the majority of studies and findings were nonsignificant.

We also found, however, that the distribution of the effect sizes was heterogeneous; that is, although the mean effect size was small, there was a lot of variability among the effect sizes. Although the potential moderators of effect size that we considered do not fully account for this variability, other moderators not analyzed in this review may be able to partition the effect sizes into homogeneous categories.

Variables that had been manipulated in the studies reviewed, differences in characteristics associated with the methodologies used in these studies, and differences associated with the articles themselves were examined in an attempt to identify variables that moderated the size of the effects. Even when one takes into consideration these variables, however, the effect sizes remain extremely small. For instance, even when the stimulus material tapped a traditionally masculine sex role orientation as opposed to a traditionally female sex role orientation, the effect size was still only around −.10. Similar conclusions were found for other variables such as sex of subject and competence of the target person.

However, like the effect sizes for the main effect, the distribution of weighted effect sizes is heterogeneous for male subjects, sex-neutral and masculine stimulus materials, highly competent target people, and high-quality stimulus material. Thus, it appears that further research will be necessary to clarify the

Table 18
Breakdown of Main Effect Size Weighted by Variance (Finding as Unit of Analysis)

Variable	M ± 95% CI	N	Homogeneity within group, $df = k - p$	Homogeneity between groups, $df = p - 1$
Age of subject				
Younger than college age	−.06 ± .08	13	15	10**
College age	−.06 ± .01	413	880*	
Older than college age	−.03 ± .02	143	489*	
% female authors				
<50%	−.04 ± .01	230	620*	62**
50%	−.01 ± .03	115	180*	
>50%	−.09 ± .02	208	532*	
Sex of first author				
Female	−.06 ± .02	310	665*	20**
Male	−.04 ± .01	250	709*	
Publication year				
1968–1973	−.07 ± .05	40	31	5
1974–1978	−.004 ± .01	351	954*	
1979–1985	−.06 ± .02	184	406*	
Journal quality				
High	−.12 ± .03	121	309*	33**
Medium	−.03 ± .01	341	829*	
Low	−.07 ± .03	113	223*	
Amount of information about target person				
Only name	−.13 ± .04	74	152*	28**
One sentence	−.02 ± .09	14	10	
One paragraph	−.01 ± .04	47	78*	
More than one paragraph	−.05 ± .01	440	1126*	
Number of independent variables associated with target person				
One	−.14 ± .02	233	664*	111**
More than one	−.01 ± .01	342	619*	
Number of target persons rated				
One	−.03 ± .02	269	433*	4
More than one	−.05 ± .01	306	957*	
Type of stimulus material				
Written work	−.03 ± .03	86	82	121**
Behavior	.00 ± .02	216	518*	
Application	−.13 ± .02	198	379*	
Biography	−.06 ± .05	41	142*	
Application and essay	−.18 ± .06	16	40*	
Other	−.11 ± .05	18	112*	
Question wording				
Rate of target person	−.05 ± .01	472	1252*	2
Rate of stimulus material	−.04 ± .03	81	124*	
Rate of both	−.01 ± .07	22	16	

Note. Negative effect sizes indicate lower evaluations of women. CI = confidence interval.
* $p < .05$, that is, effect sizes are heterogeneous within category.
** $p < .05$, that is, average effect sizes are significantly different between categories.

conditions under which these variables will be associated with large versus small effect sizes.

Correlations with and breakdowns of the effect size were used to test for the influence of methodological differences and characteristics of the articles themselves. All of the correlations were small and nonsignificant, and none of the breakdowns of the mean effect size yielded large effect sizes.

There was some indication, however, that women will be rated less favorably than men when less information is presented. The finding that the amount of information provided may influence effect size is consistent with Tosi and Einbender's (1985) quantitative review of a smaller set of these studies. Tosi and Einbender used voting scores to assess the effect of the amount of information provided on sex bias, and they used a continuous variable that represented the number of independent variables manipulated and a subjective assessment of all

Table 19
Breakdown of Main Effect Size Weighted by its Variance (Finding as Unit of Analysis Excluding Findings Assumed Equal to Zero)

Variable	*M* ± 95% CI	*N*	Homogeneity within group	Homogeneity between groups
Age of subject				
Younger than college age	−.14 ± .12	6	12	104**
College age	.00 ± .02	131	720*	
Older than college age	−.04 ± .02	76	479*	
Percent female authors				
<50%	−.06 ± .02	105	602*	147**
50%	−.02 ± .05	23	180*	
>50%	−.06 ± .04	72	386*	
Sex of first author				
Female	−.20 ± .03	95	527*	90**
Male	−.05 ± .02	108	698*	
Publication year				
1968–1973	−.28 ± .10	11	9	32**
1974–1978	−.07 ± .02	131	922*	
1979–1985	−.15 ± .03	73	352*	
Journal quality				
High	−.26 ± .04	54	216*	117**
Medium	−.05 ± .01	133	816*	
Low	−.23 ± .05	28	166*	
Amount of information about target person				
Only name	−.38 ± .06	25	64	152**
One sentence	−.04 ± .12	8	10	
One paragraph	−.02 ± .07	14	78*	
More than one paragraph	−.08 ± .01	168	1011*	
Number of independent variables associated with target person				
One	−.23 ± .02	91	512*	193*
More than one	−.02 ± .02	124	610*	
Number of target persons rated				
One	−.11 ± .04	79	402*	3
More than one	−.08 ± .02	136	910*	
Type of stimulus material				
Written work	−.14 ± .06	20	47*	238**
Behavior	.00 ± .02	88	509*	
Application	−.25 ± .03	69	253*	
Biography	−.11 ± .06	19	136*	
Application and essay	−.23 ± .16	9	30*	
Other	−.17 ± .06	10	102*	
Question wording				
Rate of target person	−.08 ± .01	184	1186*	3
Rate of stimulus material	−.13 ± .06	26	110*	
Rate of both	−.02 ± .14	5	16*	

Note. Negative effect sizes indicate lower evaluations of women. CI = confidence interval.
* $p < .05$, that is, effect sizes are heterogeneous within category.
** $p < .05$, that is, average effect sizes are significantly different between categories.

the information provided. They reported stronger support for the hypothesized relationship than what we found here. This finding is also consistent with research by Locksley et al. (1980), who found that when no information was presented about a target person, subjects made stereotypical judgments. However, when behavior relevant to the trait being judged was presented, subjects did not appear to use their stereotypes (also see Rasinski, Crocker, & Hastie, 1985). Our review included only studies in which the target had ostensibly produced a behavior or product. If we had included studies that provided subjects with only the name of a target person, perhaps larger effect sizes would have been obtained.

There was also some indication of greater bias when the stimulus material was a resumé or application. There may be something unique to studies in a job context; the three findings that were excluded because they were outliers came from a job context. The significance of this finding should not be overestimated, however; although there are indications of small differences in the magnitude of effect sizes between categories, it must be stressed that the effect sizes are uniformly very small within categories.

Given the lack of subsequent support for Goldberg's original conclusion that women will be evaluated less favorably than men even for identical work, why have Goldberg's study and subsequent replications been so frequently cited as evidence for this conclusion?[2] Several explanations may be advanced, any or all of which are plausible. First, unlike quantitative reviews, qualitative reviews are limited in their ability to summarize a large number of studies, thereby restricting their accuracy. For instance, readers may have been overly influenced by the significance of the findings and simply unaware that the size of the effects across studies was not large. Second, researchers may have lacked knowledge about other replications because of incomplete literature reviews or because of the diverse array of journals that contain such studies. Third, literature reviews often selectively emphasize significant results and trends that, if due to chance, would cause inflated perceptions of the strength of the evidence supporting the tendency to discriminate against women. In fact, nearly two thirds of the findings for the main effect had to be assumed to be equal to zero because their nonsignificant findings were not accompanied with sufficient statistical information.

Fourth, it could be argued that in the early years of the women's movement, there were no clear experimental demonstrations of sex discrimination in social psychology. Goldberg's study filled this gap by providing straightforward evidence. As Deaux (1985) suggested, at the time when Goldberg's study was conducted, there was "satisfaction with simple (yet dramatic) demonstrations of differential evaluation and reliance on perfunctory reference to gender stereotypes" (p. 66). A fifth and related explanation is that textbook writers generally prefer to use compelling illustrations such as Goldberg's experiment to buttress their arguments. They may have felt that this experiment in particular would be interesting to students because it was simple to understand and had obvious implications. Also some authors may have been unaware of or may have felt that it was too complicated to address the numerous replications that followed Goldberg's original study.

What implications does the lack of empirical support for Goldberg's conclusions have for research on gender stereotyping? One obvious but erroneous implication might be that people do not hold gender stereotypes or use such stereotypes when thinking about others. However, recent research continues to demonstrate that people are willing to express gender stereotypes (e.g., Deaux & Lewis, 1983; Martin, 1986) and that people still follow traditional sex roles (e.g., Lueptow, 1980).

One could also argue that the paradigm we reviewed is not ideal for revealing gender stereotypes. First, subjects in psychology experiments are likely to be trying to present themselves as unbiased individuals. Second, the widespread citing and describing of Goldberg's findings in introductory psychology texts, in conjunction with society's attention to sex discrimination, may well sensitize most subjects to the particular paradigm used. In effect, it is possible that the approaches we reviewed are neither blatant enough to command stereotypic evaluations nor subtle enough to expose any real biases that subjects may harbor. Such biases may emerge most clearly only when the array of available information is such that there are other plausible explanations for subjects' judgments and behaviors (and subjects even may be unaware that they are using stereotypic information in their judgments). Note that the somewhat stronger effects for the job resumé paradigm in this review could be interpreted as supporting such a position. On the other hand, because the studies we reviewed do not vary greatly in the amount of information available (e.g., virtually none matched the complexity of real-life hiring or job evaluation decisions), the previously mentioned explanation awaits a more conclusive empirical test.

A more substantive interpretation of our meta-analytic findings is that people's evaluations of men and women are a complex function of a set of factors that influence the process of gender stereotyping. In the studies reviewed in this meta-analysis, factors that might have influenced subjects' use of global gender stereotypes (but were not systematically examined in these studies) included (a) the content of gender stereotypes (e.g., the specific components and subtypes that have been activated in a situation [Deaux & Lewis, 1984; Deaux, Winton, Crowley, & Lewis, 1985]); (b) the information presented about the target person or the stimulus material (e.g., the perceived diagnosticity of the stimulus material for the evaluation [Ginossar & Trope, 1980]); (c) the interaction between the stereotype and the information presented (e.g., whether the stimulus material even activates a gender stereotype [Rothbart and John, 1985]); (d) the goals or motivations of the subjects (e.g., the subjects' level of involvement in the evaluation [Ginossar & Trope, 1987; Neuberg & Fiske, 1987]); and (e) the task demands (e.g., the complexity of the task [Bodenhausen & Lichtenstein, 1987]), or whether the task activates a more clinical/narrative judgmental orientation or a more scientific/paradigmatic orientation [Zukier, 1985; Zukier & Pepitone, 1984]).

How such factors might account for the small effect sizes found in the present meta-analysis can be illustrated by considering recent gender stereotyping research on levels of categorization, along with research on the integration of prior gender beliefs and specific case information.

Predictions about the evaluation of a specific target person may be based on more specific categories than global gender stereotypes about men and women. For instance, Deaux et al. (1985) found that subjects' descriptions of subtypes of women and men such as *housewife* or *businesswoman*, and *blue-collar working man* or *businessman* were as rich as the broader categories of woman and man. The activation of these more specific constructs may not be assessed in studies in which the dependent measures were designed to detect only the effects of more general stereotypes. In addition, social roles or other social categories such as race or occupation may also have been activated in the judgment setting and may have been more influential than gender categories (e.g., Taylor, 1980). In the studies reviewed, subtypes may have matched the target information more closely than broader stereotypes about men and women.

[2] Even Roger Brown (1987), in his insightful social psychology text, concluded his chapter on stereotyping by commenting that "the general result, varying with the date of the study, the nature of the job, is that (no surprise) discrimination in favor of men exists and (some surprise) exists whether the evaluators are themselves male or female" (p. 601).

Furthermore, there may not have been evaluative differences between the subtypes. For example, subjects may have subtypes about high-status men and high-status women, but the content of these subtypes may not differ (Deaux & Lewis, 1984).[3]

Even though male and female subtypes with different implications may be activated, subjects' judgments may be influenced more by the case information presented than by the prior beliefs associated with the subtypes. Locksley et al. (1980) demonstrated that subjects disregarded global gender stereotypes when individuating information was presented. Locksley et al. found that case information nondiagnostic of gender stereotypes yielded greater use of gender stereotypes than did diagnostic case information (see also Ginossar & Trope, 1980, 1987; Locksley, Hepburn, & Ortiz, 1982). Consistent with such findings, this review found that sex-neutral sex role stimulus material reflected greater gender bias than either masculine or feminine sex role stimulus material. In addition, this review generally found that there was a slight tendency for studies in which subjects were provided with less information to yield larger effect sizes. However, the effect sizes remained relatively small even when sex-neutral stimulus material was used and when only minimal amounts of information about the target were provided. It should be noted that this review does not represent a strong test of the Locksley et al. (1980) findings, despite the amount of information about the target person that subjects were given, because all subjects received a description of some evidence of the person's abilities. Thus, the present research still leaves open the possibility that the presence of individuating information accounts for people's disregard of their stereotypes.

In summary, little evidence was found for the simple prediction that subjects differentially evaluate men and women in the context of the classic paradigm that Goldberg introduced in 1968. This contradicts previous qualitative reviews of the literature that have in part perpetuated misconceptions about the data base on gender-biased evaluations. Gender-biased evaluations indeed occur, but as research in social cognition and, more specifically, as recent research in gender stereotypes suggest, the complexity of the conditions under which such evaluations occur and the flexibility of social perceivers' thinking must be taken into consideration.

[3] The frequency of occurrence of different subtypes of men and women may perhaps explain why discrimination is more likely to occur outside an experimental setting. For instance, within a given situation, a man is more likely to have a higher status position than a woman. Thus, comparisons between men and women are likely to be confounded by status level (cf. Eagly, 1987; Eagly & Steffen, 1984).

References

Arvey, R. D. (1979). Unfair discrimination in the employment interview: Legal and psychological aspects. *Psychological Bulletin, 86,* 736–765.

Bangert-Drowns, R. (1986). Review of developments in meta-analytic method. *Psychological Bulletin, 99,* 388–399.

Bar-Tal, D., & Saxe, L. (1976). Physical attractiveness and its relationship to sex-role stereotyping. *Sex Roles, 2,* 123–148.

Basow, S. A. (1986). *Gender stereotypes: Traditions and alternatives.* Monterey, CA: Brooks/Cole.

Berscheid, E. (1985). Interpersonal attraction. In G. Lindzey and E. Aronson (Eds.), *Handbook of social psychology* (Vol. 2, 3rd ed., pp. 413–484). Hillsdale, NJ: Erlbaum.

Bodenhausen, G. V., & Lichtenstein, M. (1987). Social stereotypes and information-processing strategies: The impact of task complexity. *Journal of Personality and Social Psychology, 52,* 871–880.

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.

Brown, R. (1987). *Social psychology* (2nd ed.). New York: Free Press.

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

Cooper, H. M., & Rosenthal, R. (1980). Statistical versus traditional procedures for summarizing research findings. *Psychological Bulletin, 87,* 442–449.

Deaux, K. (1985). Sex and gender. *Annual Review of Psychology, 36,* 49–81.

Deaux, K., & Lewis, L. L. (1983). Assessment of gender stereotypes: Methodology and components. *Psychological Documents, 13,* 25. (Ms. No. 2583)

Deaux, K., & Lewis, L. L. (1984). The structure of gender stereotypes: Interrelationships among components and gender labels. *Journal of Personality and Social Psychology, 46,* 991–1004.

Deaux, K., Winton, W., Crowley, M., & Lewis, L. L. (1985). Level of categorization and context of gender stereotypes. *Social Cognition, 3,* 145–167.

Dion, K. L., Berscheid, E., & Walster, E. (1972). What is beautiful is good. *Journal of Personality and Social Psychology, 24,* 285–290.

Eagly, A. H. (1978). Sex differences in influenceability. *Psychological Bulletin, 85,* 86–116.

Eagly, A. H. (1987). *Sex differences in social behavior: A social-role interpretation.* Hillsdale, NJ: Erlbaum.

Eagly, A. H., & Carli, L. L. (1981). Sex of researchers and sex-typed communications as determinants of sex differences in influenceability: A meta-analysis of social influence studies. *Psychological Bulletin, 90,* 1–20.

Eagly, A. H., & Crowley, M. (1986). Gender and helping behavior: A meta-analytic review of the social psychological literature. *Psychological Bulletin, 100,* 283–308.

Eagly, A. H., & Steffen, V. J. (1984). Gender stereotypes stem from the distribution of women and men into social roles. *Journal of Personality and Social Psychology, 46,* 735–754.

Eagly, A. H., & Steffen, V. J. (1986). Gender and aggressive behavior: A meta-analytic review of the social psychological literature. *Psychological Bulletin, 100,* 309–330.

Eysenck, H. J. (1978). An exercise in mega-silliness. *American Psychologist, 33,* 517.

Futoran, G. C., & Wyer, R. S. (1986). The effects of traits and gender stereotypes on occupational suitability judgments and the recall of judgment-relevant information. *Journal of Experimental Social Psychology, 22,* 475–503.

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.

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher, 5,* 3–8.

Glass, G. V., McGaw, B., & Smith, M. L. (1981). *Meta-analysis in social research.* Beverly Hills, CA: Sage.

Gleitman, H. (1981). *Psychology.* New York: Norton.

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

Gornick, N., & Moran, B. K. (1971). *Woman in sexist society: Studies of power and powerlessness.* New York: Basic Books.

Green, B. F., & Hall, J. A. (1984). Quantitative methods for literature reviews. *Annual Review of Psychology, 35,* 37–53.

Greenwald, A. G. (1975). Consequences of prejudices against the null hypothesis. *Psychological Bulletin, 82,* 1–20.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis.* Orlando, FL: Academic Press.

Hewlett, S. A. (1986). *A lesser life: The myth of women's liberation in America.* New York: William Morrow.

Hunter, J. E., Schmidt, F. L., & Jackson, G. B. (1982). *Meta-analysis: Cumulating research findings across studies.* Beverly Hills, CA: Sage.

Hyde, J. S., & Linn, M. C. (1986). *The psychology of gender: Advances through meta-analysis.* Baltimore, MD: Johns Hopkins University Press.

Johnson, R. (1980). *Elementary statistics.* North Scituate, MA: Duxbury Press.

Jones, E. E. (1985). Major developments in social psychology during the past five decades. In G. Lindzey and E. Aronson (Eds.), *Handbook of social psychology* (Vol. 2, 3rd ed., pp. 47–107). Hillsdale, NJ: Erlbaum.

Lane, D. M., & Dunlap, W. P. (1978). Estimating effect size: Bias resulting from the significance criterion in editorial decisions. *British Journal of Mathematical and Statistical Psychology, 31,* 107–112.

Lips, H. M., & Colwill, N. L. (1978). *The psychology of sex differences.* Englewood Cliffs, NJ: Prentice-Hall.

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.

Lueptow, L. B. (1980). Social change and sex-role change in adolescent orientations toward life, work, and achievement: 1964–1975. *Social Psychology Quarterly, 43,* 48–59.

Martin, C. L. (1986). A ratio measure of sex stereotyping. *Journal of Personality and Social Psychology, 52,* 489–499.

Neuberg, S. L., & Fiske, S. T. (1987). Motivational influences on impression formation: Outcome dependency, accuracy-driven attention, and individuating processes. *Journal of Personality and Social Psychology, 52,* 431–450.

Nieva, V. F., & Gutek, B. A. (1980). Sex effects on evaluation. *Academy of Management Review, 5,* 267–276.

Orwin, R. G., & Cordray, D. S. (1985). Effects of deficient reporting on meta-analysis: A conceptual framework and reanalysis. *Psychological Bulletin, 97,* 133–147.

Paludi, M. A., & Bauer, W. D. (1983). Goldberg revisited: What's in an author's name. *Sex Roles, 9*(3), 387–396.

Pheterson, G. I., Kiesler, S. B., & Goldberg, P. A. (1971). Evaluation of the performance of women as a function of their sex, achievement and personal history. *Journal of Personality and Social Psychology, 19,* 114–118.

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.

Rosen, B., & Jerdee, T. H. (1974). Sex stereotyping in the executive suite. *Harvard Business Review, 52*(March–April), 45–58.

Rosen, B., Jerdee, T. H., & Prestwich, T. L. (1975). Dual-career marital adjustment: Potential effects of discriminatory managerial attitudes. *Journal of Marriage and the Family, 37,* 565–573.

Rosenthal, R., & Rubin, D. B. (1986). Meta-analytic procedures for combining studies with multiple effect sizes. *Psychological Bulletin, 99,* 400–406.

Rothbart, M., & John, O. P. (1985). Social categorization and behavioral episodes: A cognitive analysis of effects of intergroup contact. *Journal of Social Issues, 41,* 81–104.

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.

Slavin, R. E. (1986). Best evidence synthesis: An alternative to meta-analysis and traditional reviews. *Educational Researcher, 15,* 5–11.

Strube, M. J., & Miller, R. H. (1986). Comparison of power rates for combined probability procedures: A simulation study. *Psychological Bulletin, 99,* 407–415.

Taylor, S. E. (1980). A categorization approach to stereotyping. In D. L. Hamilton (Ed.), *Cognitive processes in stereotyping and intergroup behavior* (pp. 83–114). Hillsdale, NJ: Erlbaum.

Thomas, J. R., & French, K. E. (1985). Gender differences across age in motor performance: A meta-analysis. *Psychological Bulletin, 98,* 260–282.

Tosi, H. L., & Einbender, S. W. (1985). The effects of the type and amount of information in sex discrimination research: A meta-analysis. *Academy of Management Journal, 28,* 712–723.

Unger, R. K. (1976). Male is greater than female: The socialization of status inequality. *Counseling Psychologist, 6,* 2–9.

Wallston, B., & O'Leary, V. (1981). Sex makes a difference: Differential perceptions of women and men. *Review of Personality and Social Psychology, 2,* 9–41.

Whitley, B. E., & Frieze, I. H. (1986). Measuring causal attributions for success and failure: A meta-analysis of the effects of question-wording style. *Basic and Applied Social Psychology, 7,* 35–51.

Zukier, H. (1985). The paradigmatic and narrative modes in goal-guided inference. In R. M. Sorrentino & E. T. Higgins (Eds.), *Handbook of motivation and cognition* (pp. 465–502). Hillsdale, NJ: Erlbaum.

Zukier, H., & Pepitone, A. (1984). Social roles and strategies in prediction: Some determinants of the use of base-rate information. *Journal of Personality and Social Psychology, 47,* 349–360.

(Appendixes follow on next page)

Appendix A

Articles Reviewed

Abramson, P., Goldberg, P. A., Greenberg, J. H., & Abramson, L. M. (1977). The talking platypus phenomenon: Competency ratings as a function of sex and professional status. *Psychology of Women Quarterly, 2,* 114–124.

Anderson, R., & Nida, S. (1978). Effect of physical attractiveness on opposite- and same-sex evaluations. *Journal of Personality, 46,* 401–413.

Bartol, K. M., & Butterfield, A. D. (1976). Sex effects in evaluating leaders. *Journal of Applied Psychology, 61,* 446–454.

Baruch, G. K. (1972). Maternal influences upon college women's attitudes toward women and work. *Developmental Psychology, 6,* 32–37.

Basow, S. A., & Distenfeld, M. S. (1985). Teacher expressiveness: More important for male teachers than female teachers? *Journal of Educational Psychology, 77,* 45–52.

Basow, S. A., & Howe, K. G. (1979). Sex bias and career evaluations by college women. *Perceptual and Motor Skills, 49,* 705–706.

Beattie, M. Y., & Diehl, L. A. (1979). Effects of social conditions on the expression of sex-role stereotypes. *Psychology of Women Quarterly, 4*(2), 241–255.

Bigoness, W. J. (1976). Effect of applicant's sex, race, and performance on employers' performance ratings: Some additional findings. *Journal of Applied Psychology, 61,* 80–84.

Brawley, L. R., Landers, D. M., Miller, L., & Kearns, K. F. (1979). Sex bias in evaluating motor performance. *Journal of Sport Psychology, 1,* 15–24.

Brief, A. P., & Wallace, M. J., Jr. (1976). The impact of employee sex and performance on the allocation of organizational rewards. *Journal of Psychology, 92,* 25–34.

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

Cann, A., Siegfried, W. D., & Pearce, L. (1981). Forced attention to specific applicant qualifications: Impact on physical attractiveness and sex of applicant biases. *Personnel Psychology, 34,* 65–75.

Cash, T. F., Gillen, B., & Burns, S. (1977). Sexism and "beautyism" in personnel consultant decision making. *Journal of Applied Psychology, 62,* 301–310.

Cash, T. F., & Kehr, J. (1978). Influence of nonprofessional counselors' physical attractiveness and sex on perceptions of counselor behavior. *Journal of Counseling Psychology, 25,* 336–342.

Cash, T. F., & Trimer, C. A. (1984). Sexism and beautyism in women's evaluations of peer performance. *Sex Roles, 10*(1/2), 87–98.

Chobot, D. S., Goldberg, P. A., Abramson, L. M., & Abramson, P. R. (1974). Prejudice against women: A replication and extension. *Psychological Reports, 35,* 478.

Clifford, M. M., & Walster, E. (1972). The effect of sex on college admission, work evaluation, and job interviews. *Journal of Experimental Education, 41*(2), 1–5.

Cline, M. E., Holmes, D. S., & Werner, J. C. (1977). Evaluations of the work of men and women as a function of the sex of the judge and type of work. *Journal of Applied Social Psychology, 7*(1), 89–93.

Cohen, S. L., & Bunker, K. A. (1975). Subtle effects of sex stereotypes on recruiters' hiring decisions. *Journal of Applied Psychology, 60*(5), 566–572.

Cohen, S. L., Bunker, K. A., Burton, A. L., & McManus, P. D. (1978). Reactions of male subordinates to the sex-role congruency of immediate supervision. *Sex Roles, 4*(2), 297–311.

Cohen, S. L., & Leavengood, S. (1978). The utility of the WAMS: Shouldn't it relate to discriminatory behavior. *Academy of Management Journal, 21,* 742–748.

Deaux, K., & Taynor, J. (1973). Evaluations of male and female ability: Bias works two ways. *Psychological Reports, 32,* 261–262.

Dell, D. M., & Schmidt, L. D. (1976). Behavioral cues to counselor expertness. *Journal of Counseling Psychology, 23,* 197–201.

Dipboye, R. L., Arvey, R. D., & Terpstra, D. E. (1977). Sex and physical attractiveness of raters and applicants as determinants of resumé evaluations. *Journal of Applied Psychology, 62,* 288–294.

Dipboye, R. L., Fromkin, H. L., & Wiback, K. (1975). Relative importance of applicant attractiveness, and scholastic standing in evaluation of job applicant resumes. *Journal of Applied Psychology, 60,* 39–43.

Dipboye, R. L., & Wiley, J. W. (1977). Reactions of college recruiters to interviewee sex and self-presentation style: *Journal of Vocational Behavior, 10,* 1–12.

Dipboye, R. L., & Wiley, J. W. (1978). Reactions of male raters to interviewee self-presentation style and sex: Extensions of previous research. *Journal of Vocational Behavior, 13,* 192–203.

Ekstrand, L. E., & Eckert, W. A. (1981). The impact of candidates sex on voter choice. *Western Political Quarterly, 34,* 78–87.

Etaugh, C., & Kasley, H. C. (1981). Evaluating competence: Effects of sex, marital status, and parental status. *Psychology of Women Quarterly, 6*(2), 196–203.

Etaugh, C., & Riley, S. (1983). Evaluating competence of women and men: Effects of marital and parental status and occupational sex-typing. *Sex Roles, 9*(9), 943–952.

Etaugh, C., & Rose, S. (1975). Adolescents' sex bias in the evaluation of performance. *Developmental Psychology, 11,* 663–664.

Etaugh, C., & Sanders, S. (1974). Evaluation of performance as a function of status and sex variables. *The Journal of Social Psychology, 94,* 237–241.

Etaugh, C., & Stern, J. (1984). Person perception: Effects of sex, marital status, and sex-typed occupation. *Sex Roles, 11,* 413–424.

Fidell, L. S. (1970). Empirical verification of sex discrimination in hiring practices in psychology. *American Psychologist, 25,* 1094–1098.

Francesco, A. M., & Hakel, M. D. (1981). Gender and sex as determinants of hireability of applicants for gender-typed jobs. *Psychology of Women Quarterly, 5,* 747–757.

Frank, F. D., & Drucker, J. (1977). The influence of evaluatee's sex on evaluations of a response on a managerial selection instrument. *Sex Roles, 3*(1), 59–64.

Gerdes, E. P., & Kelman, J. H. (1981). Sex discrimination: Effects of sex-role incongruence, evaluator sex, and stereotypes. *Basic and Applied Social Psychology, 2*(3), 219–226.

Gilbert, L. A., Lee, R. N., & Chiddix, S. (1981). Influence of presenter's gender on students' evaluations of presenters discussing sex fairness in counseling: An analogue study. *Journal of Counseling Psychology, 28,* 258–264.

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

Gordan, T., & Draper, T. W. (1982). Sex bias against male workers in day care. *Child Care Quarterly, 11*(3), 215–217.

Gross, M. M., & Geffner, R. A. (1980). Are the times changing? An analysis of sex-role prejudice. *Sex Roles, 6*(5), 713–722.

Gruber, K. J., & Gaebelein, J. (1979). Sex differences in listening comprehension. *Sex Roles, 5*(3), 299–310.

Gutek, B. A., & Stevens, D. A. (1979). Effects of sex of subjects, sex of stimulus cue, and androgyny level on evaluations in work situations which evoke sex role stereotypes. *Journal of Vocational Behavior, 14,* 23–32.

Haccoun, D. M., Haccoun, R. R., & Sallay, G. (1978). Sex differences in the appropriateness of supervisory styles: A nonmanagement view. *Journal of Applied Psychology, 63,* 124–127.

Haefner, J. E. (1977). Race, age, sex, and competence as factors in employer selection of the disadvantaged. *Journal of Applied Psychology, 62,* 199–202.

Hagen, R. L., & Kahn, A. (1975). Discrimination against competent women. *Journal of Applied Social Psychology, 5*(4), 362–376.

Hall, F. S., & Hall, D. T. (1976). Effects of job incumbents' race and sex on evaluations of managerial performance. *Academy of Management Journal, 19*(3), 476–481.

Hamner, W. C., Kim, J. S., Baird, L., & Bigoness, W. J. (1974). Race and sex as determinants of ratings by potential employers in a simulated work-sampling task. *Journal of Applied Psychology, 59,* 705–711.

Harris, M. B. (1975). Sex role stereotypes and teacher evaluations. *Journal of Educational Psychology, 67,* 751–756.

Harris, M. B. (1976). The effects of sex, sex-stereotyped descriptions, and institution on evaluations of teachers. *Sex Roles, 2,* 15–21.

Heilman, M. E., & Guzzo, R. O. (1978). The perceived cause of work success as a mediator of sex-discrimination in organizations. *Organizational Behavior and Human Performance, 21,* 346–357.

Heilman, M. E., & Saruwatari, L. R. (1979). When beauty is beastly: The effects of appearance and sex on evaluations of job applicants for managerial and nonmanagerial jobs. *Organizational Behavior and Human Performance, 23,* 360–372.

Heneman, H. E. (1977). Impact of test information and applicant sex on applicant evaluations in a selection simulation. *Journal of Applied Psychology, 62,* 524–526.

Heppner, P. P., & Pew, S. (1977). Effects of diplomas, awards, and counselor sex on perceived expertness. *Journal of Counseling Psychology, 24,* 147–149.

Hodgins, D. C., & Kalin, R. (1985). Reducing sex bias in judgments of occupational suitability by the provision of sex-typed personality information. *Canadian Journal of Behavioral Science, 17,* 346–358.

Isaacs, M. B. (1981). Sex role stereotyping and the evaluation of the performance of women: Changing trends. *Psychology of Women Quarterly, 6*(2), 187–195.

Kaschak, E. (1978). Sex bias in student evaluations of college professors. *Psychology of Women Quarterly, 2,* 235–243.

Kaschak, E. (1981). Another look at sex bias in students' evaluations of professors: Do winners get the recognition that they have been given? *Psychology of Women Quarterly, 5,* 767–772.

Kryger, B. R., & Shikiar, R. (1978). Sexual discrimination in the use of letters of recommendation: A case of reverse discrimination. *Journal of Applied Psychology, 63,* 309–314.

Labovitz, S. (1974). Some evidence of Canadian ethnic, racial, and sexual antagonism. *Review of Canadian Sociology and Anthropology, 11*(3), 247–254.

Lab, R. C., Upchurch, W. H., Corwin, B. J., & Grossnickle, W. F. (1975). Biased attitudes toward females as indicated by ratings of intelligence and likeability. *Psychological Reports, 37,* 1315–1320.

Larwood, L., Rand, P., & Hovanessian, A. D. (1979). Sex differences in response to simulated employee discipline cases. *Personnel Psychology, 32,* 539–550.

Lee, D. M., & Alvares, K. M. (1977). Effects of sex on descriptions and evaluations of supervisory behavior in a simulated industrial setting. *Journal of Applied Psychology, 62,* 405–410.

Lee, D. Y., Hallberg, E. T., Jones, L., & Haase, R. F. (1980). Effects of counselor gender on perceived credibility. *Journal of Counseling Psychology, 27,* 71–75.

Lenny, E., Mitchell, L., & Browing, C. (1983). The effect of clear evaluations criteria on sex bias in judgments of performance. *Psychology of Women Quarterly, 7,* 313–328.

Levenson, H., Burford, B., Bonno, B., & Davis, L. (1975). Are women still prejudiced against women? A replication and extension of Goldberg's study. *The Journal of Psychology, 89,* 67–71.

Lewin, A. Y., & Duchan, L. (1971). Women in academia. *Science, 173,* 892–895.

Linsenmeier, J. A. W., & Wortman, C. B. (1979). Attitudes toward workers and toward their work: More evidence that sex makes a difference. *Journal of Applied Social Psychology, 4,* 326–334.

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

London, M., & Stumpf, S. A. (1983). Effects of candidate characteristics on management promotion decisions: An experimental study. *Personnel Psychology, 36,* 241–259.

Mai-Dalton, R. R., Feldman-Summers, S., & Mitchell, T. R. (1979). Effect of employee gender and behavioral style on the evaluations of male and female banking executives. *Journal of Applied Psychology, 64,* 221–226.

Miller, G. R., & McReynolds, M. (1973). Male chauvinism and source competence: A research note. *Speech Monographs, 40,* 154–155.

Mischel, H. N. (1974). Sex bias in the evaluation of professional achievements. *Journal of Educational Psychology, 66,* 157–166.

Morrow, W. R., Lowenberg, G., Larson, S., Redfearn, M., & Schoone, J. (1983). Evaluations of business memos: Effects of writer sex and organizational position, memo quality, and rater sex. *Personnel Psychology, 36,* 73–85.

Muchinsky, P. M., & Harris, S. L. (1977). The effect of applicant sex and scholastic standing on the evaluation of job applicant resumes in sex-typed occupations. *Journal of Vocational Psychology, 11,* 95–108.

Noel, R. C., & Allen, M. J. (1976). Sex and ethnic bias in the evaluation of student editorials. *Journal of Psychology, 94,* 53–58.

Norton, S. D., Gustafson, D. P., & Foster, C. E. (1977). Assessment for management potential: Scale design and development, training effects and rater/ratee sex effects. *Academy of Management Journal, 20*(1), 117–131.

Paludi, M. A., & Bauer, W. D. (1983). Goldberg revisited: What's in an author's name. *Sex Roles, 9*(3), 387–396.

Paludi, M. A., & Strayer, L. A. (1985). What's in an author's name? Differential evaluations of performance as a function of author's name. *Sex Roles, 12*(3/4), 353–361.

Panek, P. E., Deitchman, R., Burkholder, J. H., Speroff, T., & Haude, R. H. (1976). Evaluation of feminine professional competence as a function of level of accomplishment. *Psychological Reports, 38,* 875–880.

Peck, T. (1978). When women evaluate women, nothing succeeds like success: The differential effects of status upon evaluations of male and female professional ability. *Sex Roles, 4*(2), 205–213.

Pheterson, G. I., Kiesler, S. B., & Goldberg, P. A. (1971). Evaluation of the performance of women as a function of their sex, achievement, and personal history. *Journal of Personality and Social Psychology, 19,* 114–118.

Piacente, B. S. (1974). Women as experimenters. *American Psychologist, 29,* 527–529.[A1]

Piacente, B. S., Penner, L. A., Hawkins, H. L., & Cohen, S. L. (1974). Evaluation of the performance of experimenters as a function of their sex and competence. *Journal of Applied Social Psychology, 4*(4), 321–329.

Renwick, P. A., & Tosi, H. (1978). The effects of sex, marital status, and educational background on selection decisions. *Academy of Management Journal, 21*(1), 93–103.

Rhue, J. W., Lynn, S. R., & Garske, J. A. (1984). The effects of compe-

[A1] Although Piacente (1974) and Piacente, Penner, Hawkins, and Cohen (1974) reported the same study, Piacente et al. (1974) is more complete; hence this reference was used as the primary source for the data.

tent behavior on interpersonal attraction and task leadership. *Sex Roles, 10,* 925–937.

Rose, G. L., & Andiappan, P. (1978). Sex effects on managerial hiring decisions. *Academy of Management Journal, 21*(1), 104–112.

Rose, G. L., & Stone, T. H. (1978). Why good job performance may (not) be rewarded: Sex factors and career development. *Journal of Vocational Behavior, 12,* 197–207.

Rosen, B., & Jerdee, T. H. (1973). The influence of sex-role stereotypes on evaluations of male and female supervisory behavior. *Journal of Applied Psychology, 57,* 44–48.

Rosen, B., & Jerdee, T. H. (1974a). Effects of applicant's sex and difficulty of job on evaluations of candidates for managerial positions. *Journal of Applied Psychology, 59,* 511–512.

Rosen, B., & Jerdee, T. H. (1974b). Influence of sex role stereotypes on personnel decisions. *Journal of Applied Psychology, 59,* 9–14.

Rosen, B., & Jerdee, T. H. (1974c). Sex stereotyping in the executive suite. *Harvard Business Review, 52,* 45–58.[A2]

Rosen, B., Jerdee, T. H., & Prestwich, T. L. (1975). Dual-career marital adjustment: Potential effects of discriminatory managerial attitudes. *Journal of Marriage and the Family, 37,* 565–572.

Sanders, G. S., & Schmidt, T. (1980). Behavioral discrimination against women. *Personality and Social Psychology Bulletin, 6*(3), 484–488.

Schmitt, N., & Lappin, M. (1980). Race and sex as determinants of the mean and variance of performance ratings. *Journal of Applied Psychology, 65,* 428–435.

Sharp, C., & Post, R. (1980). Evaluation of male and female applicants for sex-congruent and sex-incongruent jobs. *Sex Roles, 6,* 391–401.

Sigelman, L., & Sigelman, C. K. (1982). Sexism, racism, and ageism in voting behavior: An experimental analysis. *Social Psychology Quarterly, 45*(4), 263–269.

Soto, D. H., & Cole, C. (1975). Prejudice against women: A new perspective. *Sex Roles, 1*(4), 385–393.

Starer, R., & Denmark, F. (1974). Discrimination against aspiring women. *International Journal of Group Tensions, 4*(1), 65–70.

Stumpf, S. A., & London, M. (1981). Capturing rater policies in evaluating candidates for promotion. *Academy of Management Journal, 24,* 752–766.

Tanner, L. R. (1977). Sex bias in children's response to literature. *Language Arts, 54*(1), 48–50.

Taylor, S. E., & Falcone, H. T. (1982). Cognitive bases of stereotyping: The relationship between categorization and prejudice. *Personality and Social Psychology Bulletin, 8,* 426–432.

Terborg, J. R., & Ilgen, D. R. (1975). A theoretical approach to sex discrimination in traditionally masculine occupations. *Organizational Behavior and Human Performance, 13,* 352–376.

Toder, N. L. (1980). The effect of the sexual composition of a group on discrimination against women and sex-role attitudes. *Psychology of Women Quarterly, 5*(2), 92–310.

Valasek, D., Avolio, B. J., & Forbringer, L. R. (1979). Effects of sex-stereotyping in evaluating males in a traditional female role. *Psychological Reports, 44,* 1196–1198.

Walsh, R. P., & Connor, C. L. (1979). Old men and young women: How objectively are their skills assessed? *Journal of Gerontology, 34,* 561–568.

Wiley, M. G., & Eskilson, A. (1982). The interaction of sex and power base on perceptions of managerial effectiveness. *Academy of Management Journal, 25,* 671–677.

Wiley, M. G., & Eskilson, A. (1985). Speech style, gender stereotypes, and corporate success: What if women talk more like men? *Sex Roles, 12*(9/10), 993–1007.

[A2] Some of the findings from Rosen and Jerdee (1974b) are also reported in Rosen, Jerdee, and Prestwich (1975). Hence, when there were overlaps, the data were taken from Rosen et al. (1975). It should be noted that studies were included in the present meta-analysis only if they were close approximations to the original Goldberg paradigm.

Appendix B

General Assumptions

1. For all results, negative values represent a more favorable rating given to men, and positive values represent a more favorable rating given to women. If the study reported that a main effect or simple effect was significant, but the authors did not report the direction of the findings or report the means, then the findings were recorded as nonsignificant. For instance, some studies reported the means for each cell without reporting the sample size per cell. Without the sample sizes the main effect means could not be calculated nor could the simple effects be calculated for one variable summed across a second variable. Thus, it was impossible to tell if the finding was more favorable for men or for women.

2. If the dependent variable did not clearly represent a rating of favorability, it was not included in the analyses. For instance, ratings of masculinity or femininity were not included.

3. If a study only reported effects for a male (female) target person in one condition (e.g., high competence) versus a male (female) target person in a second condition (e.g., low competence), it was assumed that the other comparisons between male and female target persons were not significant (e.g., that between a highly competent man and a highly competent woman and that between a less competent man and a less competent woman were assumed to be nonsignificant.)

4. If both multivariates and univariates were reported, the univariates were used in calculations to make the results across the studies comparable. Only four studies reported multivariate F values.

5. If nonsignificant interactions were reported but follow-up tests of the simple effects resulted in significant findings, nonsignificant results were recorded for the simple effects.

6. Results from nonparametric tests were assigned nominal levels of significance.

Appendix C

Calculations

Voting scores and effect sizes were assigned to the findings in each study. Additionally, z scores were calculated and used in the calculation of effect sizes. The voting score is a tally of the number of significant and nonsignificant results. The effect size measures the magnitude of an effect independent of the significance level and the sample size.

Voting score. If a finding was significant and favored men, it was assigned a value of -1. If it favored women, it was assigned a value of $+1$. If the finding was nonsignificant or if no result was reported, it was assigned a value of 0. When study is the unit of analysis, the voting score represents the mean of the percentage of findings within each study that found significantly more favorable ratings of female target persons and male target persons and no significant differences.

Z scores. If only probability levels were reported, z scores for these levels were assigned. Z scores were assigned a value of 1.96 if the study only reported that it was significant. If the study reported that the level of significance was set at certain levels (e.g., $p \leq .05$ or $p \leq .10$), then the z scores for significant findings were reported at comparable values (e.g., 1.96 or 1.645, respectively). If the study reported that the effect was nonsignificant or did not report any effect, it was assigned a z score of zero.

If chi-squares were used, nominal levels of significance were used unless proportions were reported. In the latter case the following formulas were used (Johnson, 1980): (a) For tests in which subjects must make a choice between a man and a woman,

$$z = (p_1 - p^*)/(p^*q^*/n_1)^{1/2},$$

where p_1 = proportion of men or proportion of women chosen, n_1 = number of men or number of women chosen, and $p^* = q^* = .5$. (b) For tests in which the subject chooses a man or chooses a woman but does not make a choice between a man and a woman,

$$z = (p_1 - p_2)/[(p^*q^*(1/n_1 + 1/n_2)]^{1/2},$$

where p_1 = proportion of men chosen, n_1 = number of men chosen, p_2 = proportion of women chosen, n_2 = number of women chosen, $p^* = (p_1 + p_2)/(n_1 + n_2)$, and $q^* = 1 - p^*$.

Effect size calculations. When means and standard deviations were reported, these were used to calculate main and simple effect sizes (Hedges & Olkin, 1985; Hunter, Schmidt, & Jackson, 1982).

$$\text{Effect size} = (X_1 - X_2)/S_p,$$

where X_1 = mean for rating of female target person, X_2 = mean for rating of male target person, and S_p = pooled standard deviation. The standard deviations were pooled in the same manner for between-groups and within-group comparisons. This causes the effect size calculated in this manner for within-group comparisons to be larger because variances tend to be smaller for this design.

When only z scores, F tests, or t tests were available, these were used to calculate the effects (Hunter et al. 1982). When gender of the target person was analyzed as a between-subjects variable,

$$\text{Effect size} = M*(1/n_1 + 1/n_2)^{1/2},$$

where M = z score, t test, or the square root of the F test; n_1 = sample size for those rating male target persons; and n_2 = sample size for those rating female target persons. If not mentioned, it was assumed that there were an equal number of subjects per condition. When gender of the target person was analyzed as a within-subject variable,

$$\text{Effect size} = M*(1/n)^{1/2},$$

where M = z score, t test, or the square root of the F test; and n = number of subjects rating each pair of male and female target persons.

If both means and test statistics were reported, effect sizes were calculated from the means and standard deviations. If exact values of the z score, t tests, or F tests were not reported, the assigned z score was used in the calculations. If F tests were reported as being less than 1, they were assumed to equal 0. Test statistic values reported were assumed to be accurate, although in a few cases this did not seem likely. For instance, a few studies used between-groups tests when they should have used within-group tests. Using between-groups tests rather than within-group tests causes the error terms to be smaller and the resulting test statistics to be larger. This overestimate also causes the effect sizes to be overestimated.

If main effects were reported as significant and the interactions as nonsignificant, then the effect size assigned to the main effect was also assigned to the simple effects.

Weighting of effect sizes by variances, confidence intervals of the effect sizes, tests of the beta weights in regressions, and homogeneity tests were calculated by methods recommended by Hedges and Olkin (1985).

Received November 9, 1987
Revision received September 12, 1988
Accepted September 13, 1988 ■