Bulletin of the Psychonomic Society
1978, Vol. 12 (3), 179-181

Reliability of "the eye of the beholder": Effects of sex of the beholder and sex of the beheld

NORBERT L. KERR
University of California at San Diego, La Jolla, California 92093

and

SUSAN TURNER KURTZ
University of North Carolina, Chapel Hill, North Carolina 27514

The effects of the sex of the judge and the sex of the stimulus person on the intra- and interjudge reliabilities of physical attractiveness judgments were examined. All judges were more consistent in their judgments of female stimuli, while there was greater agreement between male judges than between female judges for all stimuli. The methodological and theoretical significance of these results are discussed.

Although the effects of physical attractiveness on social perception and interaction have been examined extensively (Berscheid & Walster, 1974), relatively little systematic work has been devoted to the measurement of physical attractiveness. Most investigators have relied upon consensual validation of their attractiveness manipulations (Berscheid & Walster, 1974), that is, those persons, photographs, silhouettes, etc., that could be reliably discriminated by several judges were assumed to differ on the physical attractiveness dimension. This approach assumes that (1) judges have standards of physical attractiveness which they apply consistently and (2) there exists some consensus between judges on these standards. The high interjudge reliabilities typically reported (e.g., Berscheid, Dion, Walster, & Walster, 1971) are consistent with these assumptions. This paper further explores the validity of the assumptions by examining the effects of the sex of the judge and the sex of the stimulus person on the intra- and interjudge reliabilities of physical attractiveness ratings. These issues are important both methodologically and theoretically. Methodologically, one would like to identify systematic sources of error variance in the scaling of attractiveness stimuli to more effectively manipulate attractiveness. Theoretically, sex differences in the reliabilities of physical attractiveness judgments may be informative for the perceptual process underlying such judgments.

There are several indications of sex differences in the perception of attractiveness. There has been a good deal of speculation and some research contending that females are more concerned about their appearance than are males (e.g., Wagman, 1967). Males consider physical attractiveness more important in making a dating choice than do females (e.g., Coombs & Kenkel, 1966), and physical attractiveness correlates more strongly with a female's dating popularity than with a male's (e.g., Berscheid et al., 1971). Bar-Tal and Saxe's (1976) literature review led them to conclude that "physical attractiveness is a more important factor in the evaluation of females than in the evaluation of males" (p. 131). Relatively greater emphasis seems to be placed on feminine appearance in the mass media, and it also appears that greater effort and expense are devoted to preserving and enhancing feminine beauty (e.g., cosmetics and cosmetic surgery). Such observations suggest that standards of physical attractiveness are better defined, more widely shared, and more salient for judgments of female than male stimuli. With this in mind, we tested the following hypotheses: (1) Judges are more consistent in their judgments of the physical attractiveness of females than of males (i.e., test-retest reliabilities are significantly higher for judgments of female stimuli). (2) There is greater between-judge agreement on the attractiveness ratings of female stimuli than of male stimuli (i.e., interjudge reliabilities are significantly higher when females are judged than when males are judged).

METHOD

Subjects and Design

The subjects were 241 students in an introductory social psychology class (114 males, 127 females). The basic design was a 2 (sex of judge) by 2 (sex of stimulus) factorial. The latter was a within-subjects factor; all subjects judged 10 male and 10 female stimuli.

Stimuli

The stimuli were slides made from recent college yearbook photographs. All of the persons pictured were Caucasian college seniors and were photographed from the shoulders up. The stimuli were chosen to span a wide range of physical attractiveness in the judgment of the experimenters. No photos were included which generally would be judged as physically grotesque or ugly.

Procedure

In a regular classroom meeting, the students were asked to rate the physical attractiveness of several persons as part of a classroom demonstration. The students were asked to refrain from any overt response to the stimuli that might influence the judgments of classmates. All 20 stimuli were then projected upon a large screen at the front of the room in a random order. A new slide was shown every 20 sec and exposed for 10 sec. The purpose of this initial exposure was to familiarize the subjects with the entire set of stimuli so that they could reasonably define the endpoints of the rating scale; no ratings were made during this exposure series. A second exposure series with a new random order followed, during which subjects privately rated each of the stimuli on "how physically attractive you find the individual shown on the screen" along a 7-point scale. Approximately 5 min after this exposure series, the subjects rerated the 20 stimuli in a third random order. No clues were given on the response sheet as to the location of any single stimulus in the order of exposure for the two ratings series; therefore, subjects could not simply copy their first response during the second rating. The interstimulus interval was shortened to 5 sec during the second rating series because subjects had made their judgments very rapidly during the previous series and appeared to be restless with the longer intervals. After completing the ratings, each subject provided his/her sex, age, marital status, frequency of dating, and perception of self-attractiveness along the same 7-point scale. Only the first of these responses are included in the data analyses reported here.

RESULTS

To achieve a balanced design, 13 randomly selected female subjects were dropped from the analyses, leaving 114 males and 114 females.

Preliminary Analyses

Preliminary analyses of variance were performed on mean attractiveness ratings and on the range of subjects' responses. This was done to probe for effects that might spuriously affect the analyses of reliability coefficients. Three such effects were obtained. The first was a significant Sex of Judge by Testing (i.e., first vs. second rating) by Stimulus photo interaction on mean attractiveness ratings; the effect of testing depended upon both the particular stimulus photo and the sex of the judge. Thus, sex of judge and/or sex of stimulus effects on test-retest reliabilities may be caused by systematic shifts in ratings due to testing rather than to true differences in the reliability of judgment. The other two significant effects of interest were a sex of stimulus main effect and a Sex of Stimulus by Sex of Judge interaction on the range of responses. (Identical results were obtained when the standard deviations of mean ratings were analyzed instead of the ranges.) The range of female photo ratings was greater, on average, than the range for male photos, and, within each sex of stimulus condition, the judges' ratings of the opposite sex displayed a wider range than the judges' ratings of the same sex. Such variations in range could also lead to spurious results in the comparison of reliability coefficients. Corrections for these problems are described below.

Analysis of Test-Retest Reliabilities

Test-retest reliability coefficients were computed separately for the male and female photos for each of the 228 subjects. To correct for between-conditions differences in the range of judgments, the following procedure was used. Each subject's two ratings of each photo were averaged. The variance of these mean ratings was computed separately for male and female photos. The variance was used to estimate what the subject's test-retest reliabilities would have been had the variance of the ratings been constant by using the formula (Guilford & Fruchter, 1973, p. 420)

$$r' = 1 - \sigma^2(1 - r)/\sigma'^2, \qquad (1)$$

where r' is the corrected reliability coefficient, σ^2 is the observed variance in ratings, r is the uncorrected test-retest correlation, and σ'^2 is the fixed variance. For this analysis and the following one on interjudge reliabilities, $\sigma'^2 = 3.0$ ($\cong$3.045, the average of the 456 variance statistics). (Other defensible corrections for systematic sources of error that might also have been applied did not alter the results.[1]) The corrected coefficients were then analyzed in a 2 (sex of judge) by 2 (sex of stimulus) analysis of variance with repeated measures on the second factor.

The cell means are presented in Table 1. As predicted, the ratings of female photos were more consistent across ratings (average $r = .954$) than ratings of male photos (average $r = .940$) [$F(1,226) = 25.77$, $p < .001$]. No other effects were significant.

Analysis of Interjudge Reliabilities

The intraclass correlation coefficient was not used to summarize the interjudge reliabilities because it did not control for the observed fluctuations in the range of judgment across the judge- and stimulus-sex conditions. An alternative method of analysis was devised. The 114 male judges were randomly grouped into 57 pairs, as were the 114 female judges. The two ratings of each stimulus were averaged for each judge. The average ratings were then used to compute separate interjudge

Table 1
Mean Reliabilities

	Test-Retest Reliabilities		Interjudge Reliabilities	
	Sex of Judge		Sex of Judges	
Sex of Stimulus	Male	Female	Male	Female
Male	.941	.939	.889	.854
Female	.958	.949	.879	.852

reliability coefficients for the male and female photos for each pair of judges. The correlations were corrected for differences in variance using Equation 1. The resulting coefficients were then analyzed in a 2 (sex of judge) by 2 (sex of stimulus) analysis of variance, with repeated measures on the second factor. The cell means for this analysis are reported in Table 1. Contrary to prediction, there was not greater consensus for ratings of female photos than for male photos ($F < 1$). Unexpectedly, there was a significant main effect for sex of judge [$F(1,112) = 6.77$, $p < .05$]; there was greater agreement between male judges (average $r = .884$) than between female judges (average $r = .853$). The interaction was not significant ($F < 1$).

DISCUSSION

The first experimental hypothesis was confirmed: Test-retest reliabilities were significantly higher for judgments of female photos than male photos. Further research must determine if this is due to better-defined standards of feminine attractiveness and/or to more consistent application of attractiveness standards to female stimuli. There was no support in the data for the second experimental hypothesis; interjudge reliabilities were not higher for judgments of female than male photos. An unexpected finding was that male judges showed significantly greater agreement with one another than did female judges, regardless of the sex of the stimuli. This effect cannot be attributed to lower consistency in ratings by female judges; the sex-of-judge effect on test-retest reliabilities was not significant. This effect might result if more complex criteria for attractiveness were applied by female judges; however, this would also lead one to expect a sex-of-judge effect on test-retest reliabilities. One plausible explanation is that male judges might have relied somewhat more upon culturally shared attractiveness stereotypes than did female judges, who may have applied more individualistic standards.

While the obtained effects were statistically reliable, they were small. The conditions of testing must be remembered when considering the size of the effects. The very short interval between tests suggests that subjects' memory of their first response may have artificially inflated the test-retest coefficients. Threats to independent responding, which can arise when assessment occurs in larger groups (e.g., copying of responses, exclamations like whistles), may have artificially inflated the interjudge reliabilities. If the inflation of these coefficients was so great as to result in ceiling effects, the strength of the obtained effects may have been sharply attenuated. The rather high reliability coefficients (mean corrected test-retest $r = .947$, mean corrected interjudge $r = .868$) are consistent with this reasoning. If ceiling effects obscured the true magnitudes of the obtained effects, as we suggest, one grows more confident that the results have practical as well as statistical significance.

Methodologically, the results suggest that attention to the sex of the judges and stimuli is required in the prescaling of physical attractiveness stimuli. When one expects physical attractiveness to produce a strong effect on some behavior, the apparently small effects of sex of judge and stimulus on the reliability of judgment found here are probably unimportant. However, if rather weak attractiveness effects are anticipated, the results suggest that error variation in prescaling can be reduced, and, hence, stronger manipulations can be achieved using female stimuli and male judges/subjects. Instances in the research literature, where attractiveness effects obtain for male subjects responding to female stimuli but not for female subjects responding to male stimuli (e.g., Efran, 1974), should be reexamined with this in mind. Where extremely weak attractiveness effects are likely, individualized prescaling of stimuli for each subject is advisable, especially when the subjects are females.

REFERENCES

Bar-Tal, D., & Saxe, L. Physical attractiveness and its relationship to sex-role stereotyping. *Sex Roles*, 1976, **2**, 123-133.

Berscheid, E., Dion, K., Walster, E., & Walster, G. Physical attractiveness and dating choice: A test of the matching hypothesis. *Journal of Experimental Social Psychology*, 1971, **7**, 173-189.

Berscheid, E., & Walster, E. Physical attractiveness. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 7). New York: Academic Press, 1974.

Coombs, R., & Kenkel, W. Sex differences in dating aspirations and satisfaction with computer-selected partners. *Journal of Marriage and the Family*, 1966, **28**, 62-66.

Efran, M. The effect of physical appearance on the judgment of guilt, interpersonal attraction, and severity of recommended punishment in a simulated jury task. *Journal of Research in Personality*, 1974, **8**, 45-54.

Guilford, J., & Fruchter, J. *Fundamental statistics in psychology and education* (5th ed.). New York: McGraw-Hill, 1973.

Wagman, M. Sex differences in types of daydreams. *Journal of Personality and Social Psychology*, 1967, **7**, 329-332.

NOTE

1. To correct for the sex of judge by testing by stimulus photo effect on attractiveness ratings, we carried out tests of simple main effects for the testing factor for each of the 20 photographs. This was done separately for male and female judges. The observed difference between mean first and second ratings was added to every judge's second rating for every photograph that had produced a significant testing simple main effect. This procedure removed any reliable changes in ratings of specific photos across testing that might spuriously affect test-retest reliabilities. The other correction we applied was a transformation of the reliability coefficients using Fisher's r-to-z transformation to help satisfy the normality assumption of the analyses of variance. Applying these two controls, singly or jointly, for the analysis of test-retest or interjudge reliabilities did not affect the results of the analyses of variance or the nature of the effects. To simplify the description of the analyses, the corrections were not applied for the analyses reported in the text.

(Received for publication May 22, 1978.)