Emotion
2004, Vol. 4, No. 2, 201–206

1528-3542/04/$12.00 DOI: 10.1037/1528-3542.4.2.201

BRIEF REPORTS

Gender Differences in Judgments of Multiple Emotions From Facial Expressions

Judith A. Hall
Northeastern University

David Matsumoto
San Francisco State University

The authors tested gender differences in emotion judgments by utilizing a new judgment task (Studies 1 and 2) and presenting stimuli at the edge of conscious awareness (Study 2). Women were more accurate than men even under conditions of minimal stimulus information. Women's ratings were more variable across scales, and they rated correct target emotions higher than did men.

Women's greater accuracy in judging the emotional meanings of nonverbal cues is well established (Hall, 1978, 1984; Hall, Carter, & Horgan, 2000). The overall gender difference corresponds to a Cohen's *d* (Cohen, 1988) of about .40 and a point-biserial correlation (*r*) of about .20. The tasks used in this literature encompass wide variation in stimuli and response options such as whether observers are asked to identify emotions, situations, or interpersonal relationships (Costanzo & Archer, 1989; Hall, 1978, 1984; Nowicki & Duke, 1994; Rosenthal, Hall, DiMatteo, Rogers, & Archer, 1979). Despite methodological variety, it has been virtually universal for such research to present stimuli for long-enough durations so that they are clearly within judges' conscious awareness, use a categorical (multiple choice) response judgment format, and compare men and women on *hits*—the number or proportion of items judged accurately.

In this article, we explore gender differences by presenting stimuli extremely quickly and by using a judgment format that allows examination of the patterning of multiple scalar ratings as opposed to a single fixed choice of emotion category. These methodological changes may provide insight into the nature of gender differences in judgments. Katsikitis, Pilowsky, and Innes (1997) demonstrated the potential utility of this type of judgment task by asking observers to rate drawings of smiling and neutral faces on a 9-point scale on which low values meant *definitely not a smile* and high values meant *definitely a smile*. Female observers gave higher ratings to smiles and lower ratings to neutral expressions than men did. Katsikitis et al. (1997) interpreted this more polarized rating pattern among women to indicate a higher level of accuracy.

In the present studies, we reanalyzed data from two previous studies that used multiple scalar ratings (Matsumoto et al., 2000; Yrizarry, Matsumoto, & Wilson-Cohn, 1998). Both studies used the Japanese and Caucasian Facial Expressions of Emotion (JACFEE; Matsumoto & Ekman, 1988), which contains 56 expressions of seven emotions (anger, contempt, disgust, fear, happiness, sadness, and surprise) depicted by 56 individuals. The expressions were verified as showing prototypical emotion expressions using Ekman and Friesen's (1978) Facial Action Coding System and have produced high agreement in categorical emotion judgments in many cultures (Biehl et al., 1997). In Study 1, each expression was presented for 10 s; in Study 2, each was presented for .20 s or less, barely within conscious awareness. The methodology of Study 1 was described in Matsumoto (1986), Matsumoto and Ekman (1989), and Yrizarry et al. (1998), and the methodology of Study 2 was described in Matsumoto et al. (2000); none of the present gender difference results was previously reported. Below we offer a condensed methodology, focusing on aspects that are relevant to the questions under consideration.

We thank Michelle Weissman for data analysis and Marianne Schmid Mast for her helpful comments on an earlier version of this article.

Correspondence concerning this article should be addressed to Judith A. Hall, Department of Psychology, Northeastern University, Boston, MA 02115, or to David Matsumoto, Department of Psychology, San Francisco State University, 1600 Holloway Avenue, San Francisco, CA 94132. E-mail: hall1@neu.edu or dm@sfu.edu

Study 1

Method

Participants were 69 male and 27 female undergraduates from the University of California, Berkeley, participating in partial fulfillment of class requirements. They viewed each of the JACFEE (Matsumoto & Ekman, 1988) expressions one at a time for 10 s in a random order. For each expression, they rated the presence or absence of seven emotions—anger, contempt, disgust, fear, happiness, sadness, and surprise—using 9-point rating scales where 0 = *not at all,* 1 = *a little,* 4 = *moderately,* and 8 = *a lot.*

Results

We first examined the variability in men's versus women's ratings of each item, predicting that the variability would be greater for women (consistent with Katsikitis et al., 1997). We computed within-participant standard deviations across the seven rating scales for each expression and then averaged these across all expressions for each participant. The mean standard deviation for women was 2.57 compared with 2.39 for men, $t(104) = 2.90$, $p < .01$. Thus, women differentiated among the rating scales more than men did.

To further understand the judgment process used by men versus women, we examined the ratings that participants gave to specific expressions. For each participant and item, we calculated the target rating (i.e., the rating given to the emotion the face was selected to display) and the mean of the nontarget ratings (i.e., the ratings given to the other six emotions). We then computed separate 2 (participant gender) × 2 (rating type: target vs. nontarget ratings) analyses of variance for each emotion. The interactions were significant for disgust, happiness, sadness, and surprise (see Table 1).

One way to interpret these interactions is based on the residual pattern of means after main effects have been subtracted (Rosnow & Rosenthal, 1989). This method suggests that women saw relatively more of the correct (target) emotion and relatively less of the nontarget emotions than men did. The term *relatively* conveys attention to the interaction pattern rather than to the means per se, which contains information about main effects as well as interaction.

A second way to interpret the significant interactions would be to use the raw cell means and test the simple effects of gender separately for target and nontarget ratings. We did so and found that women gave higher ratings than men to target emotions on anger, disgust, fear, happiness, sadness, and surprise. Six of the seven comparisons on the nontarget emotions, however, were not significant. These analyses suggest that women saw more of the target emotion than men did, but that there were no differences in ratings of nontarget emotions.

Discussion

When participants viewed prototypical facial expressions of seven emotions for 10 s and made ratings on seven emotion scales, women showed more variability in their ratings for a given stimulus expression. This greater variation among women was associated with a tendency for them to give the target emotion higher ratings than men. These results correspond to previous findings using the JACFEE (Matsumoto & Ekman, 1988) with 10-s exposures and categorical judgments, as women were more successful in picking the correct (target) emotion from a multiple-choice test than men were (Biehl et al., 1997). It is unclear whether there were gender differences on the nontarget emotions, and we have no a priori reason to give more weight to one interpretation of the significant interactions reported above than the other. Regardless, the findings provide evidence that the ratings made by women were not only more extreme than men's but also more appropriate in their patterning.

The second study examined whether the gender differences obtained using the new judgment task would occur even if the stimuli were presented so fast as to be on the edge of conscious awareness. If gender differences occur even under such conditions, that would suggest that the gender differences are not affected by exposure speed and would imply relatively automatic cognitive processing differences between men and women. In addition, we were able to perform an additional analysis of accuracy in which data from perceivers in the first study were treated as establishing a standard. Specifically, their ratings of each item were treated as the "correct" picture of how much each of the seven emotions was conveyed. Because the JACFEE (Matsumoto & Ekman, 1988) faces were selected to be prototypical, ratings were, not surprisingly, highest for the target emotion. However, perceivers did "see" other emotions. To illustrate, the ratings of disgust expressions by the standardization sample, averaged over eight faces on a scale ranging from 0 (*absent*) to 8 (*strong*), were 5.81 for disgust, 2.42 for contempt, 1.84 for anger, 0.42 for surprise, 0.30 for sadness, 0.11 for fear, and 0.06 for happiness (reported in Yrizarry et al., 1998). Because the stimuli were preselected to be prototypical and because the standardization participants were allowed to view

Table 1
Scalar Ratings by Participant Gender and Rating Type (Target vs. Nontarget) in Studies 1 and 2

Expression and rating type	Women	Men	*F*	*p*	*r*
		Study 1[a]			
Anger					
Target	6.79	6.33	2.12	.148	.14
Nontarget	1.28	1.11			
Contempt					
Target	2.48	2.30	0.11	.745	.03
Nontarget	0.98	0.92			
Disgust					
Target	6.45	5.46	10.21	.002	.30
Nontarget	0.91	0.94			
Fear					
Target	6.20	5.59	1.46	.230	.12
Nontarget	1.31	1.06			
Happiness					
Target	7.53	7.09	5.66	.019	.23
Nontarget	0.23	0.20			
Sadness					
Target	6.79	5.89	11.91	.001	.32
Nontarget	0.36	0.33			
Surprise					
Target	6.73	6.20	5.02	.027	.21
Nontarget	0.42	0.41			
		Study 2[b]			
Anger					
Target	3.41	3.05	3.88	.050	.10
Nontarget	0.93	0.95			
Contempt					
Target	2.01	1.48	8.96	.003	.16
Nontarget	0.73	0.71			
Disgust					
Target	4.09	3.41	12.67	.000	.18
Nontarget	0.75	0.80			
Fear					
Target	3.28	2.79	3.90	.049	.10
Nontarget	1.05	1.00			
Happiness					
Target	5.67	5.46	2.11	*ns*	.08
Nontarget	0.30	0.38			
Sadness					
Target	3.39	2.57	17.59	.000	.22
Nontarget	0.57	0.62			
Surprise					
Target	5.70	5.12	10.05	.002	.16
Nontarget	0.38	0.42			

Note. *F* and effect size refer to the Gender × Rating Type interaction. Effect size *r* = eta.
[a] Degrees of freedom for Study 1 = 1, 104. [b]Degrees of freedom for Study 2 = 1, 360.

each expression for a full 10 s and were under no time or cognitive load constraints, it seemed justified to consider their responses to be a gold standard against which to compare responses when the same stimuli were shown for a much briefer period of time in Study 2. *Accuracy* in Study 2 was defined as the correlation, calculated for each participant for each item, between his or her seven ratings of an expression and the corresponding seven averaged ratings from the standardization sample. The higher the correlation, the more

the perceiver "saw" the same relative amounts of each emotion in the stimulus as the standardization group did. A higher correlation is interpreted as indicating greater accuracy.

Study 2

Method

Instrument. The Japanese and Caucasian Brief Affect Recognition Test (JACBART; Matsumoto et al., 2000) was used. JACBART items were created by embedding onto videotape a JACFEE (Matsumoto & Ekman, 1988) expression for a very brief exposure in the middle of a 1-s presentation of that same expressor's neutral face. This format eliminates afterimages of the target expression. Items are presented randomly, and no emotion is presented consecutively. An orienting tone accompanied by a presentation number is shown 1 s prior to each item.

Three versions of the JACBART (Matsumoto et al., 2000) were produced, each differing only in the presentation speed of the target JACFEE (Matsumoto & Ekman, 1988) expression—.07, .13, or .20 s. Participants were randomly assigned to view one version. In this article, data were collapsed across the three versions because judge gender did not interact with version in prior analyses (Matsumoto et al., 2000).

Participants and procedures. Participants were 126 male and 237 female students at San Francisco State University participating in partial fulfillment of class requirements. They viewed the JACBART (Matsumoto et al., 2000) in small groups (ranging from 32% to 41% male) on a 17-in (43-cm) video monitor. For each expression, they completed the same scalar judgments as in Study 1. They were provided with a sheet of definitions of the seven emotion words used in the judgment tasks taken from a standard dictionary and were told not to focus on the neutral face that started and ended each item but on the target expression embedded within. Participants were then given two examples of completed rating sheets to illustrate how to use them. The videotape was stopped after each item so that participants could complete their ratings.

Scoring of scalar accuracy. For each item, scalar accuracy was computed as described above. Twelve accuracy scores were computed by averaging scores over items within the following expression types: anger, contempt, disgust, fear, happiness, sadness, and surprise; Caucasian and Japanese; males and females; and total.

Results

Gender differences in patterning of responses. As in Study 1, we calculated the within-participant standard deviation on each JACBART (Matsumoto et al., 2000) item and then averaged these standard deviations across all 56 items for each participant. The mean standard deviation was 2.14 for women and 1.95 for men, $t(360) = 3.40$, $p < .001$. Women's ratings differentiated among the rating scales more than men's.

We then performed the same analysis of target versus nontarget ratings as in Study 1 (see Table 1). The relevant interactions were significant for every emotion except happiness. As in Study 1, two interpretations are available for these interactions. In one, analyzing interaction residuals leads to the conclusion that women gave relatively higher ratings to target emotions and relatively lower ratings to nontarget emotions than men. In the other, simple effects analyses of raw cell means indicated that women gave significantly higher ratings on target emotions than men did on all emotions except happiness, which was in the same direction but not significant, but that there were no gender differences on six of seven nontarget emotions, although women gave lower ratings to the nontarget emotions for five of the seven emotions. As in Study 1, we have no a priori reason to favor one interpretation over the other.

Gender differences in scalar accuracy. Accuracy scores were compared between men and women using *t* tests. Women had significantly higher accuracy than men on nearly all of the test subscales as well as the total (see Table 2). That is, women had a stronger correlation between their ratings of each expression on the seven emotion scales and the corresponding standardization ratings for that expression. This finding is especially interesting when one considers that another group of participants in Matsumoto et al. (2000; $n = 89$) responded to the JACBART (Matsumoto et al., 2000) using fixed-choice categorical judgments, and there were no gender differences in accuracy using this scoring method (all ps $> .42$).

Relation of response patterning to scalar accuracy. Because the expressions contained in the JACBART (Matsumoto et al., 2000) were created to be prototypical (i.e., to show a clearly predominant emotion), the more extreme response pattern of women (irrespective of which interpretation of the significant interactions one believes) was thus not only different from men's but also represented a more accurate description of the expressions. It follows that this difference may be related to the scalar accuracy

Table 2
Comparison of Men and Women on Scalar Accuracy for Study 2

	Women		Men				
Item type	*M*	*SD*	*M*	*SD*	*t*	*p*	*r*
Caucasian	0.59	0.18	0.53	0.17	3.35	.001	.17
Japanese	0.58	0.18	0.50	0.18	3.98	.000	.20
Male	0.58	0.17	0.52	0.18	3.44	.001	.18
Female	0.59	0.18	0.52	0.18	3.87	.000	.20
Anger	0.50	0.26	0.44	0.27	2.14	.033	.11
Contempt	0.35	0.26	0.26	0.24	3.30	.001	.17
Disgust	0.59	0.24	0.53	0.23	2.60	.010	.14
Fear	0.56	0.28	0.47	0.30	2.73	.007	.14
Happiness	0.84	0.20	0.81	0.21	1.68	.094	.09
Sadness	0.48	0.31	0.35	0.30	3.57	.000	.18
Surprise	0.80	0.21	0.77	0.22	1.29	*ns*	.07
Total	0.59	0.17	0.52	0.17	3.78	.000	.20

Note. All degrees of freedom = 361. Effect size *r* = eta.

scores. To investigate this, we first calculated *extremity scores,* defined as the arithmetic difference between ratings of the target emotion and the averaged ratings of nontarget emotions for each item for each participant. We then correlated each participant's extremity scores across the 56 items with that participant's accuracy scores, producing a correlation for each participant reflecting how well his or her accuracy was related to the degree to which he or she used the extreme scoring pattern. As hypothesized, these correlations were substantially positive for both men (mean $r = .51$) and women (mean $r = .49$), and both were highly significantly greater than zero for men, $t(125) = 18.89$, and for women, $t(236) = 25.79$, both ps $< .0001$.

Discussion

Study 2 demonstrated that women were more accurate than men in judging the emotional meaning from nonverbal cues even when the stimuli were presented so fast as to be at the edge of conscious awareness. This effect, however, occurred only when judges were allowed to rate multiple emotions in the stimuli and by using a correlational method for scoring accuracy that is sensitive to patterning, not just right–wrong categorical judgments; when judges made a single emotion category judgment, there were no gender differences. The overall gender difference on scalar accuracy ($r = .20$) was identical to that established in the literature for categorical measures of accuracy at longer exposure durations ($r = .20$; Hall, 1978, 1984; Hall et al., 2000).

Study 2 also showed that men and women produced a different pattern of rating emotions in facial expressions. Although two interpretations exist for the significant interactions that we obtained, they both agree that women gave higher ratings to the correct (target) emotion; they disagree in whether there is a gender difference in the nontarget ratings. In terms of statistical significance, the interaction results were stronger in Study 2, being significant for six of seven rather than four of seven emotions. However, there is not much difference between mean effect sizes in the two studies (mean effect size for Study 1 interactions = .19; mean effect size for Study 2 interactions = .14).

General Discussion

That women are more accurate than men in judging emotional meaning from nonverbal cues even under situations of minimal stimulus information is an important finding that implies differential cognitive processing capabilities for men and women. This finding, if replicated, could lead to important new research about the origins of these differences. It could be, for example, that women are socialized to decode emotions better than men from such an early age that the ability to do so is more automated for women than for men. Alternatively, it could be that female brains are better equipped to decode emotions than are male brains from birth.

The fact that the somewhat different rating pattern exhibited by men versus women was evident both for 10-s and less than 1-s stimulus exposures indicates that it is a gender difference not uniquely linked to exposures. Questions remain, however, about why men's and women's rating patterns diverged. One

possibility is that women more quickly ascertained that the expression was prototypical (in gestalt fashion) and accordingly used the more extreme ratings that are indeed appropriate in such a case (that is, rating the target emotion much higher than other emotions). In contrast, men may have used a less gestalt approach, trying to analyze individually each of the seven possible emotions for each face. Considering the brevity of the stimulus (even when it was 10-s long), such a strategy may be excessively time-consuming and result in more guesswork and ratings closer to the midpoint of the scale.

A second possibility is that women were more confident in their judgments and therefore more willing to commit to giving higher ratings to the target emotion, whereas the men were not as confident and consequently "hedged their bets" by being more conservative, that is, less extreme, in their judgments. Lower confidence on the part of men could result either from not being sure of what emotions were on the face at the time of viewing or not remembering the stimulus as well as women did at the time they made their ratings (which was after the stimulus was no longer in view).

The positive correlations for both genders between accuracy and the tendency to distinguish more between target and nontarget emotions might lead to the conclusion that accuracy as we calculated it is essentially redundant with the extreme rating pattern. However, the correlation made use of the profile of ratings across all seven rated emotions, whereas the extremity scores combined all nontarget emotions into one average. Therefore, the correlational index of accuracy is likely to be a more sensitive indicator. Indeed, for the JACBART (Matsumoto et al., 2000), stimuli in Study 2 proved to be much more sensitive than a comparison of hits based on categorical judgments, for which there were no gender differences. Thus, for the JACBART, men and women identified the predominant emotion equally well, but women did a better job of discerning the relative mix of perceived emotions in the stimuli.

References

Biehl, M., Matsumoto, D., Ekman, P., Hearn, V., Heider, K., Kudoh, T., & Ton, V. (1997). Matsumoto and Ekman's Japanese and Caucasian Facial Expressions of Emotion (JACFEE): Reliability data and cross-national differences. *Journal of Nonverbal Behavior, 21,* 3–21.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Mahwah, NJ: Erlbaum.

Costanzo, M., & Archer, D. (1989). Interpreting the expressive behavior of others: The Interpersonal Perception Task. *Journal of Nonverbal Behavior, 13,* 225–245.

Ekman, P., & Friesen, W. V. (1978). *The Facial Action Coding System (FACS)*. Palo Alto, CA: Consulting Psychologists Press.

Hall, J. A. (1978). Gender effects in decoding nonverbal cues. *Psychological Bulletin, 85,* 845–857.

Hall, J. A. (1984). *Nonverbal sex differences: Communication accuracy and expressive style*. Baltimore: Johns Hopkins University Press.

Hall, J. A., Carter, J. D., & Horgan, T. G. (2000). Gender differences in the nonverbal communication of emotion. In A. H. Fischer (Ed.), *Gender and emotion: Social psychological perspectives* (pp. 97–117). Paris: Cambridge University Press.

Katsikitis, M., Pilowsky, I., & Innes, J. M. (1997). Encoding and decoding of facial expression. *Journal of General Psychology, 124,* 357–370.

Matsumoto, D. (1986). *Cross-cultural communication of emotion*. Unpublished doctoral dissertation, University of California, Berkeley.

Matsumoto, D., & Ekman, P. (1988). *Japanese and Caucasian Facial Expressions of Emotion and Neutral Faces (JACFEE and JACNeuF)*. (Available from the Human Interaction Laboratory, University of California, San Francisco, 401 Parnassus Avenue, San Francisco, CA 94143)

Matsumoto, D., & Ekman, P. (1989). American-Japanese cultural differences in intensity ratings of facial expressions of emotion. *Motivation and Emotion, 13,* 143–157.

Matsumoto, D., LeRoux, J., Wilson-Cohn, C., Raroque, J., Kooken, K., Ekman, P., et al. (2000). A new test to measure emotion recognition ability: Matsumoto and Ekman's Japanese and Caucasian Brief Affect Recognition Test (JACBART). *Journal of Nonverbal Behavior, 24,* 179–209.

Nowicki, S., Jr., & Duke, M. P. (1994). Individual differences in the nonverbal communication of affect: The Diagnostic Analysis of Nonverbal Accuracy Scale. *Journal of Nonverbal Behavior, 18,* 9–36.

Rosenthal, R., Hall, J. A., DiMatteo, M. R., Rogers, P. L., & Archer, D. (1979). *Sensitivity to nonverbal communication: The PONS test*. Baltimore: Johns Hopkins University Press.

Rosnow, R. L., & Rosenthal, R. (1989). Definition and interpretation of interaction effects. *Psychological Bulletin, 105,* 143–146.

Yrizarry, N., Matsumoto, D., & Wilson-Cohn, C. (1998). American-Japanese differences in multiscalar intensity ratings of universal facial expressions of emotion. *Motivation and Emotion, 22,* 315–327.

Received May 1, 2003
Revision received October 1, 2003
Accepted October 8, 2003 ■