MATSUMOTO AND EKMAN'S JAPANESE AND CAUCASIAN FACIAL EXPRESSIONS OF EMOTION (JACFEE): RELIABILITY DATA AND CROSS-NATIONAL DIFFERENCES

Michael Biehl, David Matsumoto, Paul Ekman, Valerie Hearn, Karl Heider, Tsutomu Kudoh, and Veronica Ton

ABSTRACT: Substantial research has documented the universality of several emotional expressions. However, recent findings have demonstrated cultural differences in level of recognition and ratings of intensity. When testing cultural differences, stimulus sets must meet certain requirements. Matsumoto and Ekman's Japanese and Caucasian Facial Expressions of Emotion (JACFEE) is the only set that meets these requirements. The purpose of this study was to obtain judgment reliability data on the JACFEE, and to test for possible cross-national differences in judgments as well. Subjects from Hungary, Japan, Poland, Sumatra, United States, and Vietnam viewed the complete JACFEE photo set and judged which emotions were portrayed in the photos and rated the intensity of those expressions. Results revealed high agreement across countries in identifying the emotions portrayed in the photos, demonstrating the reliability of the JACFEE. Despite high agreement, cross-national differences were found in the exact level of agreement for photos of anger, contempt, disgust, fear, sadness, and surprise. Cross-national differences were also found in the level of intensity attributed to the photos. No systematic variation due to either preceding emotion or presentation order of the JACFEE was found. Also, we found that grouping the countries into a Western/Non-Western dichotomy was not justified according to the data. Instead, the cross-national differences are discussed in terms of possible sociopsychological variables that influence emotion judgments.

Cross-cultural research has documented high agreement in judgments of facial expressions of emotion in over 30 different cultures (Ekman,

1994), including preliterate cultures (Ekman, Sorensen, & Friesen, 1969; Ekman & Friesen, 1971). Recent research, however, has reported cultural differences in judgment as well. Matsumoto (1989, 1992a), for example, found that American and Japanese subjects differed in their rates of recognition. Differences have also been found in ratings of intensity (Ekman et al., 1987).

Examining cultural differences requires a different methodology than studying similarities. Matsumoto (1992a) outlined such requirements: (1) cultures must view the same expressions; (2) the facial expressions must meet criteria for validly and reliably portraying the universal emotions; (3) each poser must appear only once; (4) expressions must include posers of more than one race.

Matsumoto and Ekman's (1988) Japanese and Caucasian Facial Expressions of Emotion (JACFEE) was designed to meet these requirements. JACFEE was developed by photographing over one hundred posers who voluntarily moved muscles that correspond to the universal expressions (Ekman & Friesen, 1975, 1986). From the thousands of photographs taken, a small pool of photos was coded using Ekman and Friesen's (1978) Facial Action Coding System (FACS). A final pool of photos was then selected to ensure that each poser only contributed one photo in the final set, which is comprised of 56 photos, including eight photos each of anger, contempt, disgust, fear, happiness, sadness, and surprise. Four photos of each emotion depict posers of either Japanese or Caucasian descent (2 males, 2 females).

Two published studies have reported judgment data on the JACFEE, but only with American and Japanese subjects. Matsumoto and Ekman (1989), for example, asked their subjects to make scalar ratings (0–8) on seven emotion dimensions for each photo. The judgments of the Americans and Japanese were similar in relation to strongest emotion depicted in the photos, and the relative intensity among the photographs. Americans, however, gave higher absolute intensity ratings on photos of happiness, anger, sadness, and surprise. In the second study (Matsumoto, 1992a), high agreement was found in the recognition judgments, but the level of recognition differed for anger, disgust, fear, and sadness.

While data from these and other studies seem to indicate the dual existence of universal and culture-specific aspects of emotion judgment, the methodology used in many previous studies has recently been questioned on several grounds, including the previewing of slides, judgment context, presentation order, preselection of slides, the use of posed expressions, and type of response format (Russell, 1994; see Ekman, 1994, and Izard, 1994, for reply). Two of these, judgment context and presentation order, are especially germane to the present study and are addressed here.

Within-subject designs may inflate recognition levels because of "a more direct comparison of various facial expressions" (Russell, 1994, p. 112). This claim, however, has been disputed (Ekman, 1994; Izard, 1994). Ekman (1994) argued that we make such comparisons all the time when interacting with others in everyday situations. Izard (1994) cited evidence (Izard, 1971; Izard, Haynes, Fantauzzo, Slomine, & Castle, 1993) where preceding emotion varied with little or no change in recognition scores and concluded that "decisions about the message in a facial expression in daily life typically take full advantage of all of the information in the visual field, and such decisions are likely to be context free only in the laboratory" (p. 293).

Presentation order in within-subject designs may also inflate agreement because subjects are responding to the context of the same presentation order (Russell, 1994). Ekman (1994), however, pointed out that order did vary across studies reporting similar findings.

Russell (1994) compared judgment accuracy percentages between Western and Non-Western countries to highlight these possible concerns. In his analysis, the percentage judgment data from 31 countries were used as dependent variables. The results of these analyses showed that Western countries tended to have higher percentage data. From these findings, Russell (1994) inferred that the methods used in these studies, including within-subject designs and presentation orders, may bias the results toward showing high agreement.

The purpose of this study was to demonstrate the reliability of the JACFEE to elicit pancultural agreement in emotion judgments, while at the same time testing for cross-national differences in agreement levels and intensity ratings. Another purpose of this study was to test empirically some of Russell's claims concerning the influence of methodology on judgments. Judges from Hungary, Japan, Poland, Sumatra, United States, and Vietnam were presented all 56 photographs in the JACFEE set twice. Subjects first judged the emotion portrayed in the expressions by selecting a single emotion term from a list of alternatives. On the second viewing, they rated the overall intensity of the expressions. We tested the following hypotheses:

1. Observers within and across countries will demonstrate high levels of agreement on the emotion portrayed in the JACFEE photos, consistent with previous findings concerning universality in judgments (Ekman, 1994).
2. There will be cross-national differences in the exact level of agreement, consistent with findings reported in Matsumoto (1992a) and Russell (1994).

3. There will be cross-national differences in the absolute intensity levels attributed to the expressions, consistent with findings from Ekman et al. (1987) and Matsumoto and Ekman (1989).
4. Emotion judgments will vary systematically as a function of the type of emotion that precedes it; consistent with findings reported in Russell & Fehr (1987) and argued by Russell (1991; 1994).
5. Emotion judgments will vary systematically as a function of presentation order, consistent with Russell's (1994) arguments.

Method

Judges

Judges were 271 Americans, 75 Poles, 45 Hungarians, 44 Japanese, 34 Vietnamese, and 32 Sumatrans. The single choice (but not intensity) data for 41 of the Americans and for the 44 Japanese were reported earlier in Matsumoto (1992a). Also, the single choice (but not intensity) data for the contempt expressions for Japanese, Polish, Hungarian, and Vietnamese subjects were reported in Matsumoto (1992b). The intensity data for 24 photos used in an American-Japanese comparison of display rules were also reported earlier (Matsumoto, 1990). As the focus in this study was the reliability of the JACFEE set, we merged all available data for a complete analysis of all JACFEE photos.

The American sample included 128 males (mean age 19.88) and 143 females (mean age 19.78). The 75 Poles included 40 males (mean age 19.41) and 35 females (mean age 20.31). The Hungarian sample included 12 males (mean age 19.83) and 33 females (mean age 19.87). The 44 Japanese included 22 males (mean age 21.09) and 22 females (mean age 20.18). The Vietnamese sample included 19 males and 15 females ranging in age from 18–40. The 32 Sumatrans included 17 males (mean age 21.24) and 15 females (mean age 20.73). All subjects were born and raised in their respective countries and, except for the Vietnamese, were college undergraduates participating in partial fulfillment of class requirements. Subjects were recruited from universities in metropolitan areas including the San Francisco Bay Area (United States), Warsaw (Poland), Szeged (Hungary), Osaka (Japan), and Padang (Sumatra). The Vietnamese subjects were recruited from English as a Second Language (ESL) courses in the San Francisco Bay Area, and were in the U.S. for less than a year. The fact that, with the exception of the Vietnamese sample, all participants were students at universities in major urban areas provided some basis for inferring some degree of equivalence in socioeconomic class within country. No other

demographic data, however, were obtained, and the results reported below should be interpreted with this caveat.

Facial Stimuli, Judgment Tasks, and Procedures

The procedures for collecting the judgment data were the same for all countries, with a slight modification in Sumatra. The stimuli included all 56 photos of Matsumoto and Ekman's (1988) JACFEE, and additional photos taken from the original pool of photos from which the JACFEE was selected and from stimulus sets used in previous judgment studies. The additional photos also portrayed emotion either in full face, partially, or blended with other emotions. Subjects were tested in groups, and were shown the stimuli twice. On each pass, the stimuli were presented one at a time, for 10 seconds each, in a random order. Hungarian, Japanese, Vietnamese, and one group of American subjects received one random order. Polish, Sumatran, and two groups of American subjects each received different random orders. (Different presentation orders occurred because of differences in the total stimulus sets used across countries, depending on the number of additional photos included in the stimulus set.) During the first viewing, subjects selected a single term from seven choices (anger, contempt, disgust, fear, happiness, sadness, and surprise) that best described the emotion portrayed. On the second viewing, subjects judged the intensity of each expression using a 9-point scale (0–8) labeled NOT AT ALL (0), A LITTLE (1), A MODERATE AMOUNT (4), and A LOT (8). There was no mention of any emotion terms in this second viewing, to eliminate the influence of cultural differences in the exact meaning of the words on intensity ratings. The procedures for the Sumatran subjects differed from other countries in that the two judgments were separated by a week.

All protocols were originally written in English and then translated into the target languages. The accuracy of the translations was verified by back-translation procedures. Consistent with the purposes of this study, only the data from the JACFEE expressions were used in the analyses.

Results

Hypothesis 1: Universal Agreement in Emotion Judgments and Reliability of the JACFEE

Main analyses. The single-choice data were first analyzed by calculating the percent of subjects selecting the intended emotion on each of the JACFEE photos for all six countries (Table 1). The percentages were gener-

TABLE 1

Percent of Subjects Within Each Culture Who Chose Predicted Emotion

Photo ID	Emotion	USA	JPN	SUM	VIE	POL	HUN
ES1-2C17	ANCAMA	88.3	77.3	96.9	87.9	90.5	88.9
RH-1C24	ANCAMA	87.4	63.6	78.1	91.2	81.1	93.3
KG-1C21	ANCAFE	71.9	70.5	71.0	73.5	85.1	82.2
LR-1C24	ANCAFE	93.4	88.6	100.0	90.9	91.9	88.9
AF-1C30	ANJAMA	86.1	43.2	64.5	75.8	86.5	86.0
BM-1C22	ANJAMA	86.0	54.5	43.8	69.7	77.0	84.4
NM-2C14	ANJAFE	85.9	65.9	83.9	81.8	82.4	82.2
AL-1C21	ANJAFE	75.6	50.0	90.6	76.5	83.8	88.9
Anger Average		*84.3*	*64.2*	*78.6*	*80.9*	*84.8*	*86.9*
JH-1C10	COCAMA	64.3	61.4	71.9	84.8	85.1	88.6
ER-2C11	COCAMA	66.1	75.0	87.1	94.1	87.8	93.3
KN-1C09	COCAFE	63.0	84.1	79.3	88.2	70.3	86.7
WW-1C09	COCAFE	61.6	70.5	87.1	91.2	83.3	84.4
PM-1C11	COJAMA	63.5	84.1	83.3	82.4	83.8	88.9
SC-1C08	COJAMA	68.3	81.8	90.0	76.5	77.0	84.4
YW-2C04	COJAFE	46.3	70.5	62.5	90.9	85.1	80.0
AK2-1C10	COJAFE	67.2	86.4	75.0	88.2	81.1	84.4
Contempt Average		*62.5*	*76.7*	*79.5*	*87.0*	*81.7*	*86.3*
JB1-1C33	DICAMA	87.0	77.3	56.3	36.4	79.2	84.4
BC-1C15	DICAMA	85.5	75.0	87.5	58.8	86.3	93.3
GM-1C14	DICAFE	69.8	70.5	71.0	67.6	86.5	77.8
EG-1C21	DICAFE	75.0	84.1	84.4	73.5	85.1	80.0

YY1-1C21	DIJAMA	83.9	84.1	90.3	79.4	87.8	84.1
EK-1C22	DIJAMA	77.0	61.4	50.0	30.3	76.7	79.5
YY2-1C21	DIJAFE	91.4	70.5	78.1	70.6	86.5	97.8
ES2-1C22	DIJAFE	78.1	75.0	90.3	44.1	78.4	77.8
DISGUST AVERAGE		*81.0*	*74.7*	*76.0*	*57.6*	*83.3*	*84.3*
SG-4C19	FECAMA	75.8	43.2	53.1	76.5	73.0	71.1
JL-1C34	FECAMA	72.7	47.7	76.7	61.8	63.5	60.0
KB-1C36	FECAFE	75.8	45.5	25.0	64.7	62.2	51.2
SB-1C36	FECAFE	85.9	50.0	46.7	55.9	71.6	77.8
DM-1C25	FEJAMA	74.1	70.5	71.9	58.8	67.1	82.2
HU-1C35	FEJAMA	88.9	54.5	70.0	52.9	64.9	84.4
AM-2C25	FEJAFE	86.3	56.8	65.6	85.3	81.1	91.1
DT-2C31	FEJAFE	74.1	68.2	45.2	79.4	68.9	73.3
FEAR AVERAGE		*79.2*	*54.6*	*56.8*	*66.9*	*69.0*	*73.9*
DG-1C05	HACAMA	93.7	95.5	100.0	100.0	100.0	93.3
SA-1C35	HACAMA	98.1	90.9	100.0	94.1	90.5	95.6
EA-1C06	HACAFE	99.3	100.0	100.0	100.0	100.0	97.8
TA-1C04	HACAFE	98.9	100.0	100.0	97.1	98.6	97.8
DL-1C05	HAJAMA	98.5	100.0	100.0	100.0	97.3	97.8
HF-1C04	HAJAMA	97.0	100.0	96.8	97.1	100.0	97.8
JL-1C03	HAJAFE	97.4	100.0	96.8	100.0	98.6	100.0
LK-1C04	HAJAFE	97.8	100.0	96.9	100.0	100.0	100.0
HAPPINESS AVERAGE		*97.6*	*98.3*	*98.8*	*98.5*	*98.1*	*97.5*
JC-1C28	SACAMA	87.6	31.8	73.3	30.3	48.6	64.4
SW1-1C20	SACAMA	94.1	93.2	93.5	87.9	100.0	95.6
NH1-1C31	SACAFE	94.4	95.5	100.0	94.1	93.2	88.9

Table 1 (Continued)

Photo ID	Emotion	USA	JPN	SUM	VIE	POL	HUN
DC-1C22	SACAFE	91.1	56.8	68.8	88.2	95.9	84.4
CF-4C07	SAJAMA	91.9	81.8	60.0	71.9	84.7	77.8
GO-1C31	SAJAMA	88.4	52.3	81.3	88.2	89.0	91.1
RK-1C31	SAJAFE	93.3	84.1	93.8	91.2	94.6	73.3
EN-1C28	SAJAFE	91.1	79.5	78.1	91.2	98.6	91.1
Sadness Average		*91.5*	*71.9*	*81.1*	*80.4*	*88.1*	*83.3*
JG-1C17	SUCAMA	96.7	90.9	90.3	94.1	97.3	97.8
AG-1C13	SUCAMA	90.7	90.9	90.3	97.1	95.9	88.9
SS1-1C14	SUCAFE	86.6	95.5	68.8	85.3	78.4	95.6
MM-1C17	SUCAFE	96.3	97.7	93.8	91.2	94.6	97.8
AK1-1C14	SUJAMA	76.8	77.3	84.4	84.8	61.6	86.7
ST-1C15	SUJAMA	95.1	95.5	96.8	94.1	95.9	97.8
KK1-1C33	SUJAFE	96.6	97.7	96.9	91.2	94.6	95.6
YF-1C04	SUJAFE	95.2	93.0	93.5	94.1	94.5	93.3
Surprise Average		*91.8*	*92.3*	*89.4*	*91.5*	*89.1*	*94.2*
Total Average		*84.0*	*76.1*	*80.0*	*80.4*	*84.9*	*86.6*

Note. Key for Emotion Category: First two letters indicate emotion: AN = anger, CO = contempt, DI = disgust, FE = fear, HA = happiness, SA = sadness, SU = surprise; second two letters indicate poser race: CA = Caucasian, JA = Japanese; last two letters indicate poser sex: MA = male, FE = female.

ally well above chance (1/7), and were comparable with previous data supporting universality in emotion recognition. Across all photos and countries, the intended emotion category was the modal response 321 out of 336 opportunities (56 photos × 6 countries). Using a conservative estimate of 50% agreement in modal response across countries as chance, a binomial test of the ratio of agreement to opportunities was highly significant, $p < .0001$, as predicted. (We acknowledge here, however, the potential unreliability of this test given non-independence in the data because of repeated measurement from the same sample within each country.) These results supported Hypothesis 1.

Additional analyses. The analysis presented above, which is common to judgment studies in the past, defined modal response of a country without criteria concerning the level of agreement within that country. For example, if an emotion term was selected by 35% of the judges in a country, and no other emotion term was selected by a greater percentage of judges, that emotion term was considered the modal response for that country. Previous judgment studies have typically reported recognition rates above 70% across countries for the universal expressions (which is well above what would be considered chance—1/7—given the judgment task). No study, however, has examined the threshold at which agreement across countries diminishes. To extend the findings reported above, we tallied the number of times the percent of judges within a country selected the intended emotion term, at the same time specifying a minimum percentage level that needed to be met before a country was tallied, starting with 70% and increasing in 5% increments. At each level, we computed a binomial test of the ratio of the number of countries meeting criteria for recognition within the country to the total number of opportunities, with the assumption of chance at 50% agreement across countries at that level of recognition (we again acknowledge the possible unreliability of the statistic). At the 70%, 75%, and 80% recognition levels, the binomial tests were still significant at the $p < .0001$ level, but failed to reach significance at the 85% level. We then computed these analyses separately for each emotion, reckoning that different emotions may have different recognition threshold levels of agreement. Happiness had the highest agreement level (95%, $p < .0001$), followed by surprise (90%, $p < .01$), and sadness (80%, $p < .01$). Disgust ($p < .01$), contempt ($p < .01$), and anger ($p < .001$) were significant at the 75% agreement level. Photos of fear were significant at the 60% agreement level ($p < .001$). Thus, there appears to be quite a range of recognition levels across the emotions that produce agreement across countries. These data are consistent with previous published data indicat-

ing that happiness and surprise often elicit high, and fear lower, agreement levels (Ekman, 1994).

One of the problems with "traditional" methods of data analysis that utilize percentage data such as those above is that recognition accuracy is confounded with frequency of usage of each response category. Wagner (1993), in fact, has suggested that these methods inflate accuracy rates and has developed a new scoring procedure that corrects for this problem. His procedure involves computation of an unbiased hit rate which is the "joint probability both that a stimulus is correctly identified (given that it is presented) and that a response is correctly used (given that it is used)" (p. 16). We computed these new accuracy scores along with corresponding chance proportions, and conducted a four-factor analysis of variance (ANOVA) using judge country (6) and judge gender (2) as between-subject factors, and emotion (7) and accuracy vs. chance (2) as within-subject factors. Results revealed a significant interaction for Judge Country X Emotion X Accuracy vs. Chance, $F(30, 2862) = 11.48$, $p < .001$.[1] One-way ANOVAs comparing accuracy vs. chance scores were computed for each emotion, separately for all countries. All 42 analyses (7 emotions X 6 countries) were significant at the $p < .001$ level, indicating that judges in all countries selected the intended emotion term at a rate far greater than chance even given computation of the unbiased hit rates. These results further contributed considerable support for Hypothesis 1.

Hypothesis 2: Cross-National Differences in Emotion Recognition

One consideration in testing cross-national differences in judgment concerns the treatment of the countries. Grouping countries into a priori categories has the advantage of aiding in the construction of classifications that may be useful in interpreting differences should they be found. For example, Russell (1994) suggested the utility of comparing Western versus Non-Western (W-NW) countries, and presented data to suggest that differences in recognition levels could reliably be interpreted according to differences along this distinction. To test this notion, we classified the U.S., Poland, and Hungary as "Western" cultures, and Japan, Vietnam, and Sumatra as "Non-Western" cultures. We converted the nominal judgment data into continuous scores by calculating the number of times each judge selected the predicted emotion term for each of the four poser types (Caucasian males and females, Japanese males and females); thus, scores ranged from 0 (did not select predicted emotion term for either of the two photos) to 2. We then computed a five-way ANOVA on these scores, using judge culture (2) and judge gender (2) as between-subject factors, and emotion (7), poser

race (2), and poser gender (2) as within-subject factors. A significant main effect for judge culture indicated that Western cultures did indeed have higher accuracy scores than the Non-Western cultures, $F(1, 413) = 22.51$, $p < .001$, $R^2 = .05$ (means [SD] = 1.70 [.22] and 1.57 [.23], respectively). Also, the Judge Culture X Emotion X Poser Race X Poser Gender interaction was significant, $F(6, 2478) = 5.19$, $p < .001$, $R^2 = .01$. We thus tested judge culture differences separately for the four poser types across all seven emotions. The analyses indicated no W-NW differences for happy or surprise expressions; Non-Western cultures had significantly higher accuracy scores than Western cultures on all four poser types of contempt; and Western cultures had significantly higher accuracy scores than Non-Western cultures on three of four comparisons of anger expressions, three of four comparisons of disgust expressions, and all four comparisons for both fear and sad expressions.[2]

The W-NW distinction, however, runs the risk of glossing over important national differences within classification. Keeping countries separate has the advantage of eliminating this risk, but has the disadvantage of making more problematic the search for common themes across the countries that can be used to interpret differences should they be found. The risks involved in grouping countries can be tested empirically, and we did so by computing two five-factor ANOVAs on the recognition scores, using judge country (3) and judge gender (2) as between-subject factors, and poser race (2), poser gender (2) and emotion (7) as within-subject factors. In one analysis, the U.S., Poland, and Hungary comprised the three levels of the judge country factor; in the other analysis, Japan, Sumatra, and Vietnam comprised the three levels of the judge country factor. Results revealed a significant Judge Country X Emotion X Poser Race X Poser Gender interaction for both Western, $F(12, 1908) = 4.26$, $p < .001$, and Non-Western countries $F(12, 522) = 2.57$, $p < .001$. These findings indicated the existence of cross-national differences in recognition levels within both classifications, and suggested that the previous W-NW differences be interpreted with this caveat.

Because of possible cross-national differences within the W-NW classification, we computed a five-factor ANOVA on the recognition scores keeping all countries separate. Results revealed a significant Judge Country X Emotion X Poser Race X Poser Gender interaction, $F(30, 2430) = 3.88$, $p < .001$.[3] One-way ANOVAs comparing judge countries were computed separately for each poser type within the seven emotions, with Newman-Keuls follow-up comparisons. Several consistent significant differences were found. Japanese were worse at identifying anger than Americans, Hungarians, Poles, and Vietnamese; worse at identifying fear than Ameri-

cans, Hungarians, and Poles; and worse at identifying sadness than Americans and Poles. Americans were worse at identifying contempt than Hungarians, Japanese, Poles, and Vietnamese. All other countries were better at identifying disgust than Vietnamese. There was no consistent pattern of differences for surprise, nor any significant differences for happiness.[4] Findings from all analyses presented in this section supported Hypothesis 2.

Hypothesis 3: Cross-National Differences in Judgments of Intensity

As in Hypothesis 2 above, we tested Hypothesis 3 two ways, once using the W-NW distinction, the second time treating the six countries independently. In the first analysis, we computed one-way ANOVAs on the intensity ratings separately for each of the 56 expressions, using only those judges who selected the emotion term in the single-choice judgment task intended by the expression. This selection eliminated the confounding effects of subjects who judged a face as a non-intended emotion. Consequently, these analyses did not collapse across the two photos for each poser type, and had to be computed separately for each photo, unlike the analyses above for Hypothesis 2. Across the 56 comparisons, the findings for two emotions were clear, with Non-Western cultures giving significantly higher intensity ratings than Western cultures on seven of eight fear expressions, but with Western cultures giving higher intensity ratings on all eight happy expressions. The findings for the other emotions, however, were less clear. The Western cultures gave higher intensity ratings on four of eight expressions of anger, four of contempt, three of disgust, three of sadness, and four of surprise. The Non-Western cultures, however, gave significantly higher ratings on two expressions of anger and one of contempt. All other comparisons were not significant.[5]

As above in Hypothesis 2, we also examined whether the W-NW distinction was empirically justified by computing two five-factor ANOVAs on the intensity ratings for both Western (Hungary, Poland, U.S.) and Non-Western (Japan, Sumatra, Vietnam) categories, using the same factors as above. Results again revealed significant Judge Country X Emotion X Poser Race X Poser Gender interactions for both Western, $F(12, 2136) = 5.05$, $p < .001$, and Non-Western countries $F(12, 534) = 6.14$, $p < .001$. Thus, the findings presented immediately above should be interpreted with this caveat.

In the second test of Hypothesis 3, we computed one-way ANOVAs comparing judge countries separately for each expression, with Newman-Keuls follow-up comparisons. There were consistent significant differences between the countries in intensity ratings for photos of contempt, disgust,

and happiness. Sumatrans perceived contempt less intensely than Hungarians, disgust less intensely than Japanese, and happiness less intensely than all other countries. No consistent pattern of significant differences was found for photos of anger, fear, sadness, and surprise.[6] Both sets of analyses presented in this section offer considerable support for Hypothesis 3.

Hypotheses 4 and 5: The Influence of Methodology on Judgments

The single-choice percentage for each photo was calculated separately by the emotion preceding it for each country and presentation order. We averaged these percentages across the eight photos per emotion (these percentages available from authors upon request), and tested differences in the mean percentages using ANOVA (type of preceding emotion as the independent variable), separately for each emotion. Of the seven analyses, only one (happiness) was significant, $F(5, 37) = 3.15, p < .05$. Inspection of the mean percentages for happiness, however, indicated that this significant result may have been a function of restrictions in variance in the percentages (range 96.36 to 100.00). Thus, Hypothesis 4 was rejected.

In order to test Hypothesis 5, the presentation orders were divided into quarters and accuracy percentages were calculated for each country, order, and emotion. If Hypothesis 5 were true, the data should indicate an increasing trend in the percentages across the four quarters for each presentation order, and this trend should be present in well over a majority of the data reviewed. Instead, an increasing trend was found only eight out of a possible 63 (7 emotions X 9 country/orders) times. In addition, we computed correlations between the position in sequence of a photo and the percent of subjects recognizing the predicted emotion. Of the 63 correlations computed, only six positive correlations were statistically significant. Finally, order effects were also tested by computing accuracy scores (0–1) for each photo, and then averaging the accuracy scores within each of the four quarters of presentation. An increasing trend effect across the four presentation quarters was found in three of 42 chances (6 countries X 7 emotions; different presentation orders for U.S. subjects were averaged). Thus, Hypothesis 5 was rejected.

Discussion

The results indicated high agreement within and across countries in the judgments of emotion in each of the JACFEE expressions. The percent

agreement within each country was well above chance (1/7), and the number of times the countries agreed was also well above chance. The same findings were obtained when accuracy rates were recalculated using Wagner's (1993) method of producing unbiased hit rates. Altogether, these data are entirely comparable with data from previous studies that establish the universality in judgments of these facial expressions (Ekman, 1994), and demonstrate the pancultural agreement in emotion judgments of the JACFEE.

The findings, however, also indicated that there were considerable differences between the countries in the exact level of agreement. The Japanese, for example, seemed to fare worse than the other countries in agreeing on expressions of anger, fear, and sadness. Americans had less agreement than other countries on contempt, while Vietnamese had less agreement than other countries on disgust. These findings are not due to artifacts of stimuli quality, because the findings would not be consistent across photographs, and the percentages of agreement are similar to previous research (and one would have to argue that the same ambiguities existed in previous research stimuli). We do not interpret the existence of cross-national differences in judgments as evidence against universality, either. Rather, they highlight the importance of learned rules of emotion judgments that differ from culture to culture. Earlier, Ekman and Friesen's (1969) neuro-cultural theory explained how emotional expressions can be universal while at the same time culture-specific according to culturally learned display rules. We suggest that a similar mechanism exists for judging emotions as well. That is, judgments of which emotion is portrayed in the faces are most likely based on the universality of those expressions, which some would argue are biologically innate. This feature contributes to the levels of agreement in judgments within and across countries that are substantially higher than chance. Still, people of different countries and cultures learn culture-specific rules of decoding (cf. Buck, 1984) that introduce tendencies to alter agreement levels in judgments. These tendencies may exist not necessarily in the general emotion category used to classify an expression, but rather in the exact semantics and general affective meanings associated with that category label. Thus, while cultures will all tend to view each emotion as that emotion intended, differences in agreement levels among cultures will occur because of differences in the semantic or affective meanings and associations of the emotion terms used as response alternatives. These latter differences comprise what may be considered culture-specific rules of decoding emotions, and would explain why agreement levels are often substantially higher than chance yet variable across countries at those high levels, as obtained in this and other studies.

Agreement levels for happy and surprise expressions were highest, and for fear lowest, and these data are comparable with previous judgment data as well. Differences in the agreement levels across emotions may arise because of differences in the social or semantic meanings of the emotions, or because of differences in their expressions. For example, judgments may be affected by the frequency of their occurrence, and thus judgment, of these expressions in real life. Happy and surprise expressions may occur most frequently, for example, and fear the most infrequently, and the degree of practice in making such judgments may result in differences in judgment agreement levels. Second, judgments may be affected by the degree of complexity of the facial components involved in the expressions. Happy and surprise expressions are relatively "simple" expressions, given the number of facial components involved and the degree of voluntary control over those muscles. Fear, on the other hand, is one of the most complex expressions, given the number of muscles innervated and the relative lack of control over the naturally antagonistic action of some of those muscles. Finally, there may be overlap among the emotions in the semantic or affective meanings of the emotion categories used as response alternatives. Such degrees of overlap may contribute to differences in agreement levels when single-choice responses are obtained and analyzed. An element of fear, for example, may be a bit of surprise. If the percentage data for surprise as a response category were added to that of fear for judgments of fear photos, then the agreement levels would be considerably higher and comparable to those of the other emotions. Future research needs to test whether any, or none, of these ideas has merit.

The cross-national comparisons in intensity ratings using raw scores revealed differences in contempt, disgust, and happiness. The existence of cross-national differences in intensity ratings replicates the findings of previous studies (Ekman et al., 1987; Matsumoto & Ekman, 1989). These differences could not have been influenced by translation terms, since there was no mention of these terms in this session. Cultural differences in intensity ratings may be interpreted in several ways. In our previous studies (Ekman et al., 1987; Matsumoto & Ekman, 1989), we have interpreted these differences as indicative of learned cultural norms of judgment and expression. For example, we interpreted previous American-Japanese differences as a function of the notion that, when the Japanese express emotions, they express them less intensely than do Americans, which leads to lower intensity ratings when judging expressions. There are, however, other alternative explanations, and which of these are correct, if any, need to be addressed in future work. For example, it may be the case that the exact opposite is true. That is, emotions may be expressed more intensely

in the Japanese culture; thus, when judging the same stimulus, they interpreted the expression as less intense than did the Americans. Yet another interpretation would suggest that intensity ratings are confounded by confidence in the ratings, and that low confidence produces low ratings. We would speculate that the differences are due to cultural norms in decoding emotions, rather than to differences in the average intensity of actually-occurring emotions or confidence. But, our interpretation remains speculation until future research can address this issue directly.

As future research struggles with these alternative explanations of the findings, one major issue concerns the treatment of countries and considerations of the cultural factors that contribute to differences in judgments of emotion or intensity ratings. In his article, Russell (1994) utilized a W-NW dichotomy to analyze data and support his claims about possible bias in methodology. Indeed, this classification produced differences in this study in both recognition data and intensity ratings. We, however, believe this classification to be unjustified, for several reasons. First, our analyses indicated that the countries within both of these classifications differed on both the recognition and intensity judgments. We believe this to be a more conservative approach, testing the validity of the assumption of homogeneity among the countries being classified prior to any such arbitrary classification. Second, there is a major difference in the unit of analysis in our study compared to Russell's (1994). Russell used each country's percentage agreement levels as data; thus, country were the cases in his analysis. In our study, judges were cases. Third, it is not clear what the criterion for inclusion in either the "Western" or "Non-Western" groups should be. In both this study and Russell's (1994), a distinction between Western and Non-Western countries was assumed, and countries were arbitrarily placed into those categories.

Instead of using such notions of culture as W-NW, which are quickly becoming antiquated, we would argue for the use of a theoretical framework about culture that explains differences in display and perception across all countries without lumping them together. We have proposed such a theoretical framework (Matsumoto, 1989, 1990) that goes beyond the traditional definition of equating culture and country, and instead accounts for differences in emotional display rules across cultures based on sociopsychological dimensions, such as those Hofstede (1980, 1983) postulated. These dimensions include Individualism (vs. Collectivism), Power Distance (the degree to which cultures emphasize or minimize power differentials), Uncertainty Avoidance (the degree to which cultures adopt rules and rituals to minimize anxiety concerning the unknown), and Masculinity (the degree to which cultures foster gender differences). We be-

lieve these types of dimensions are more powerful in explaining cross-cultural differences in judgments as well as expressions. For example, the fact that the Japanese were worse at identifying anger, fear, and sadness than most countries is consistent with the collectivistic nature of Japan. Negative emotions would be more likely to disrupt group harmony and social interactions and yield lower accuracy scores due to lack of experience with these emotions in a variety of contexts. In individualistic cultures, however, these emotions are more and may even be emphasized, yielding higher accuracy scores. Similar hypotheses concerning the nature of other cultural differences using other dimensions may also be postulated. Future research generating scores for each country on these dimensions can then assess the degree to which these tendencies exist along a continuum of cultural difference, instead of lumping countries into arbitrary dichotomies.

Finally, there were no systematic differences due either presentation order or preceding emotion (Ekman, 1994; Izard, 1994). Russell's studies (Russell, 1991; Russell & Fehr, 1987), however, did find differences in recognition rates due to preceding emotion. It is possible that his methodology, employing a much smaller number of emotions and photos, increases the chances of subjects considering and comparing previous stimuli and subsequently producing differences. These demands may disappear when tested in a larger set of stimuli, where it is much more difficult for subjects to remember previous stimuli.

Notes

1. Other effects were also significant. The significant main effect of judge culture indicated that differences in accuracy scores existed regardless of judge gender and type of emotion rated, $F(5, 477) = 7.08$, $p < .001$, $R^2 = .07$ (means ranged from a low of .66 for the Japanese judges to a high of .80 for the Hungarian judges). A significant main effect for judge gender indicated that females were generally more accurate in judging the emotions than males, $F(1, 477) = 31.51$, $p < .001$, $R^2 = .06$. A significant main effect for emotion indicated that the emotions were judged at differential rates of accuracy, from a low of .61 for contempt to a high of .95 for happiness, $F(6, 2862) = 106.72$, $p < .0001$, $R^2 = .18$. A full report on the entire analysis can be obtained from David Matsumoto.
2. A detailed report of these findings can be obtained from David Matsumoto.
3. Other effects were also significant. A significant judge culture main effect indicated that the countries differed in overall accuracy across all expressions, $F(5, 405) = 5.88$, $p < .001$ (means ranged from a low of 1.51 for Japanese judges to a high of 1.74 for Hungarian judges). A significant judge gender main effect indicated that female judges generally had higher accuracy scores than did male judges, $F(1, 405) = 16.90$, $p < .001$. A significant main effect for emotion indicated that the emotions were judged at differential rates of accuracy, from a low of 1.43 for contempt to a high of 1.96 for happiness, $F(6, 2430) = 74.45$, $p < .0001$. A significant main effect for poser gender indicated that expressions posed by females were more accurately judged than those posed by males, $F(1, 405) = $

20.89, $p < .0001$. Other effects were also significant, and a full report on the entire analysis can be obtained from David Matsumoto.
4. A detailed report of these findings can be obtained from David Matsumoto.
5. A detailed report of these findings can be obtained from David Matsumoto.
6. These analyses were also computed on ratings standardized within country. Analysis of standardized scores eliminate the possible confound of cultural response sets in the data (see Matsumoto, 1994) and offer researchers another glimpse of the data. A different pattern of country differences emerged for the standardized scores. Sumatrans perceived happiness less intensely than all other countries, while Americans perceived disgust less intensely than Japanese. Additionally, Hungarians perceived Caucasian anger less intensely than Vietnamese; and Poles perceived female surprise more intensely than Americans. A detailed report of both raw score and standardized score findings can be obtained from David Matsumoto.

References

Buck, R. (1984). *The communication of emotion.* New York: Guilford.

Ekman, P. (1994). Strong evidence for universals in facial expressions: A reply to Russell's mistaken critique. *Psychological Bulletin, 115,* 268–287.

Ekman, P., & Friesen, W. (1971). Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology, 17,* 124–129.

Ekman, P., & Friesen, W.(1975). *Unmasking the face: A guide to recognizing emotions from facial clues.* Englewood Cliffs, NJ: Prentice Hall.

Ekman, P., & Friesen, W. (1978). *Facial action coding system.* Palo Alto, CA: Consulting Psychologists Press.

Ekman P., & Friesen, W. (1986). A new pancultural facial expression of emotion. *Motivation and Emotion, 10,* 159–168.

Ekman, P., Friesen, W., O'Sullivan, M., Diacoyanni-Tarlatzis, I., Krause, R., Pitcairn, T., Scherer, K., Chan, A., Heider, K., LeCompte, W. A., Ricci-Bitti, P. E., Tomita, M., & Tzavaras, A. (1987). Universals and cultural differences in the judgments of facial expressions of emotion. *Journal of Personality and Social Psychology, 53,* 712–717.

Ekman, P., Sorenson, E. R., & Friesen, W. (1969). Pancultural elements in facial displays of emotions. *Science, 164,* 86–88.

Hofstede, G. (1980). *Cultures Consequences.* Sage Publications, Beverly Hills, CA.

Hofstede, G. (1983). Dimensions of natural cultures in fifty countries and three regions. In J. Deregowski, S. Dziurawiec, & R. A. Anais (Eds.), *Expiscations in cross-cultural psychology.* Lisse: Swets and Zeitlinger.

Izard, C. E. (1971). *The face of emotion.* New York: Appleton-Century-Crofts.

Izard, C. (1994). Innate and universal facial expressions: Evidence from developmental and cross-country research. *Psychological Bulletin, 115,* 288–299.

Izard, C. E., Haynes, O. M., Fantauzzo, C. A., Slomine, B. S. & Castle, J. M., (1993). *The morphological stability and social validity of infants' facial expressions in the first nine months of life.* Manuscript submitted for publication.

Matsumoto, D. (1989). Cultural influences on the perception of emotion. *Journal of Cross-Cultural Psychology, 20,* 92–105.

Matsumoto, D. (1990). Cultural similarities and differences in display rules. *Motivation and Emotion, 14,* 195–214.

Matsumoto, D. (1992a). American-Japanese cultural differences in the recognition of universal facial expressions. *Journal of Cross-Cultural Psychology, 23,* 72–84.

Matsumoto, D. (1992b). More evidence for the universality of a contempt expression. *Motivation and Emotion, 16,* 363–368.

Matsumoto, D. (1994). *Cultural influences on research methods and statistics.* Pacific Grove, CA: Brooks Cole.

Matsumoto D. & Ekman, P. (1988). *Japanese and Caucasian facial expressions of emotion (JACFEE)* [Slides]. San Francisco, CA: Intercultural and Emotion Research Laboratory, Department of Psychology, San Francisco State University.

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

Russell, J. A. (1994). Is there universal recognition of emotion from facial expression? A review of cross-cultural studies. *Psychological Bulletin, 115,* 102–141.

Russell, J. A. (1991). Negative results on a reported facial expression of contempt. *Motivation and Emotion, 15,* 281–291.

Russell, J. A., & Fehr, B. (1987). Relativity in the perception of emotion in facial expressions. *Journal of Experimental Psychology: General, 116,* 223–237.

Wagner, H. L. (1993). On measuring performance in category judgments studies of nonverbal behavior. *Journal of Nonverbal Behavior,* 17(1), Spring, 3–28.