Psychological Bulletin
2002, Vol. 128, No. 2, 236–242

0033-2909/02/$5.00 DOI: 10.1037//0033-2909.128.2.236

Methodological Requirements to Test a Possible In-Group Advantage in Judging Emotions Across Cultures: Comment on Elfenbein and Ambady (2002) and Evidence

David Matsumoto
San Francisco State University

H. A. Elfenbein and N. Ambady's (2002) conclusions concerning a possible in-group advantage in judging emotions across cultures are unwarranted. The author discusses 2 methodological requirements for studies to test adequately the in-group advantage hypothesis and an additional requirement in reviewing multiple judgment studies and examining variance in judgment effects across those studies. The few studies that Elfenbein and Ambady reported that support the in-group advantage hypothesis need to be examined for whether they meet the criteria discussed; if they do not, their data cannot be used to support any contention of cultural differences in judgments, let alone the in-group advantage hypothesis. Furthermore, the role of signal clarity needs to be explored in possibly moderating effects across studies; however, this was not done.

In this issue, Elfenbein and Ambady (2002) presented a meta-analysis examining the universality and cultural specificity of emotion recognition. It is timely because of the large number of studies conducted on this issue to date, and it is needed because the basic questions that a meta-analysis can address are important ones that have been pondered for years. One issue that Elfenbein and Ambady raised concerns a possible in-group advantage in judging emotions across cultures. This hypothesis suggests that recognition accuracy is higher when emotions are both expressed and perceived by members of the same cultural group. Elfenbein and Ambady examined this issue across all studies, as well as separately according to communication channel, specific emotions, and different emotions across channels. They also examined how several variables moderate an in-group advantage and drew wide-sweeping conclusions concerning it.

Unfortunately, there are severe limitations in the designs of the studies that Elfenbein and Ambady (2002) used as evidence for the hypothesis and in the logic underlying their conclusions about it. I describe those problems here, raising fundamental design issues that need to be considered when conducting and reviewing judgment research. In doing so, I provide the readership with an alternative way of thinking about the level and type of scholarship that would be necessary to test the in-group advantage hypothesis adequately and to reach conclusions about it.

Below, I describe three methodological requirements that are necessary to test the in-group advantage hypothesis correctly, the first two concerning the conduct of individual studies (the use of balanced designs and stimulus equivalence across expressor cultures) and the third concerned with reviewing multiple studies (signal clarity). I present analyses from previously published studies from my laboratory that meet the methodological requirements described and reanalyze the in-group advantage effects for only balanced studies reported by Elfenbein and Ambady. Finally, I discuss how signal clarity may relate to a possible in-group effect and present a rival hypothesis for the data on the basis of this issue. My position is that (a) most studies that Elfenbein and Ambady reviewed probably did not meet the methodological requirements for adequately testing the in-group advantage hypothesis; (b) even if one grants those studies the methodological issues I raise, a reexamination of their data suggests that the effect is negligible; (c) the studies that do meet the criteria indicate that the effect is nonexistent; and (d) if the in-group advantage effect does exist, it may exist under certain conditions of signal clarity; however, this was not addressed in their review.

Correspondence concerning this article should be addressed to David Matsumoto, Department of Psychology, San Francisco State University, 1600 Holloway Avenue, San Francisco, California 94132. E-mail: dm@sfsu.edu

Three Methodological Requirements to Examine the In-Group Advantage Hypothesis

Only Balanced Designs Can Test the In-Group Advantage Hypothesis

Elfenbein and Ambady (2002) reviewed three types of studies in their meta analysis:

1. Studies in which one set of stimuli was viewed by multiple cultural groups (66 of 97 studies),
2. Studies in which multiple sets of stimuli were viewed by one cultural group (7 of the 97 studies),
3. Studies in which members of each culture viewed stimuli from members of their own and each other group (balanced designs; 21 of 97 studies).

The data that are produced in these studies can be summarized in Table 1, assuming a two-culture comparison. Studies of Type 1 produce data in cells W and X. Studies of Type 2 produce data in cells W and Y. Studies of Type 3 produce all four cells of data.

I strongly contend that studies of Types 1 and 2 cannot not be used for testing a possible in-group advantage. For Type 1, data indicating $W > X$ is used by Elfenbein and Ambady (2002) to support the contention for an in-group advantage of Culture A over Culture B. However, if $Y > Z$ as well, then the difference is not so much an in-group advantage as it is a decoding difference

Table 1
Examination of the Data Obtained in the Studies Reviewed by Elfenbein and Ambady (2002)

	Judges	
Stimuli	Culture A	Culture B
Depicting people of Culture A	W	X
Depicting people of Culture B	Y	Z

between Cultures A and B because the same difference is found regardless of the stimuli being judged. If an in-group advantage were occurring in the manner described by Elfenbein and Ambady, it should follow that Z > Y. Without data in cells Y and Z in studies of Type 1, however, it is impossible to make this determination, and in fact this interpretation is unjustified on the basis of the design of those studies.

The same argument holds true for studies of Type 2, which provide data in cells W and Y. Again, without cells X and Z, there is no way to know whether the difference being observed is an in-group advantage or a decoding issue.

The only type of study that can adequately test an in-group bias in judgment is the balanced design of Type 3 because it provides the data necessary to distinguish whether differences are due to an in-group advantage or to a decoding effect. That is, balanced designs are the only designs that allow for an elimination of a rival hypothesis.

Stimuli Need to Be Equivalent in Emotion-Signaling Properties Across Encoder Cultures

The second methodological requirement to test adequately the in-group advantage hypothesis, or any cultural difference in judgment of emotion stimuli, concerns the characteristics of the stimuli used. If stimuli portraying emotion expressed by people of two different cultural groups are to be judged by members of both of those groups, then the characteristics of the stimuli specific to the emotion message must be exactly equivalent between the two expressor cultures, whereas only the characteristics related to cultural identification should vary.

For example, if faces portraying emotion expressed by people of Cultures A and B are shown to judges of both cultures, then the characteristics of the face related to the emotion must be exactly the same between both cultures' expressors. This means that the same facial muscles must be innervated, with no extraneous muscle movements, and they must be at the same intensity levels. In addition, other aspects of the face related to cultural identification must be the only characteristics of the stimuli that vary systematically (facial physiognomy and morphology).

If the signals specifically related to emotion expressed by Culture A are different than those expressed by Culture B, judgments of these stimuli by observers of Cultures A and B are inherently confounded by differences in the emotion signals. If, for instance, facial expressions from Culture A involve different muscle movements than those of Culture B, or if the muscles innervated are at different intensity levels, then judgment differences between Cultures A and B may be due to differences in the stimuli, not decoding processes.

The only way to address this issue adequately is to measure the actual physical properties of the stimuli related to emotion signaling to ensure that they do not vary across expressor cultures. When facial stimuli are used, such measurement can be achieved by Ekman and Friesen's Facial Action Coding System (FACS; see Ekman & Friesen, 1978). If other stimulus channels are used (e.g., voice), investigators need to demonstrate that the stimuli do not vary on the physical-signal properties specific to emotion in those channels across expressor culture. If the emotion-signal properties are not equivalent among the expressor cultures, the comparison of judge cultures is inextricably confounded by stimulus differences.

Certainly, the original investigators in Elfenbein and Ambady's (2002) review needed to have established such signal equivalence in the first place if their purpose was to test for cultural differences in judgments; and to tell the truth, I am not sure of how many of them did so. I would guess only a few, if any, but that is not the point. The point here is that it was also incumbent on Elfenbein and Ambady to examine whether stimulus equivalence was met in the studies prior to reviewing them with regard to the in-group advantage hypothesis. They did not do so. What they did do was code whether the recognizability of the stimuli used in the studies was validated by a separate consensus sample of raters. However, this does not address the issue adequately because judgments cannot be used to validate equivalence in the physical properties of the signals to be judged, especially if the consensus sample is from only one, not all, of the cultures being studied, which is the typical way in which external consensus samples are obtained.

When Evaluating Judgment Effects Across Studies, Differences May Be Related to Variance Across the Studies in Signal Clarity

Stimuli vary greatly on signal clarity (Ekman, Friesen, & Ellsworth, 1972; O'Sullivan, 1982). Some stimuli are very clear in terms of their emotional content; Ekman and Friesen's Pictures of Facial Affect (PFA; see Ekman & Friesen, 1976) and the Japanese and Caucasian Facial Expressions of Emotion (JACFEE; Matsumoto & Ekman, 1988) are two examples of widely used stimuli that are high in signal clarity. Some stimuli are more ambiguous; the Profile of Nonverbal Sensitivity (PONS; see Rosenthal, Hall, DiMatteo, Rogers, & Archer, 1979), for instance, an audiovisual test of decoding nonverbal cues, was explicitly designed to have less signal clarity overall so as to produce greater individual differences.

Signal clarity is important because emotion-judgment processes differ depending on it. Consider the PFA and JACFEE as examples: When signal clarity is high, there is strong consensus within and across cultures concerning the emotions that are portrayed in the stimuli (e.g., see Biehl et al., 1997; and Matsumoto, 2001, for a review); when signal clarity is lowered, however, by increasing the speed of presentation (only one of the ways in which this may be achieved), accuracy rates are lower, individual differences that are correlated with personality traits exist, and sex differences emerge (Matsumoto, Hall, & Weissman, 2001; Matsumoto et al., 2000).

Because signal clarity has differential effects on the judgment process, it is necessary that reviews of multiple-judgment studies incorporate and measure the level of signal clarity used in the studies being reviewed—judgment effects of any sort, including possible in-group advantage effects, may vary across studies de-

pending on the level of signal clarity in the stimuli used in each study. An effect may occur in some studies that have a certain level of signal clarity, whereas it may not occur in others because of a different level of signal clarity. For this precise reason, the meaning of grouping all studies together regardless of signal clarity in producing overall effect sizes is questionable. Elfenbein and Ambady's (2002) analysis simulating a balanced design, in which they combined unbalanced studies involving either one Caucasian or Asian group, is exactly one of these questionable analyses. Although I can appreciate their attempt to address the issues I am raising here through the available data, their simulation of a balanced design sidesteps the reality that the studies are not able to address those issues adequately in the first place because judge–culture accuracy is inherently confounded with stimuli and thus signal clarity (not to mention the fact that the studies do not meet the requirements for stimulus equivalence across expressor cultures). Differences may exist, but they may exist because of the methodology of the studies, not because of an in-group advantage, and this analysis cannot be used as evidence for it because it does not unequivocally rule out any reasonable rival hypothesis of the data.

Elfenbein and Ambady's coding of the validation of the recognizability of the stimuli by a separate consensus sample of raters is a step in the right direction, but it cannot deal adequately with the signal clarity issue either, because signal clarity may have differential effects on judgments even though the stimuli used met an external criterion for inclusion. To use an example from Biehl et al.'s (1997) research paradigm, statistical significance is achieved if the percentage of judges selecting an emotion category is substantially greater than chance. If there are nine categories provided, chance is 1/9. Even under a more conservative estimate of chance at 1/2 or when correction for category usage is used (Wagner, 1993), agreement rates associated with statistical significance can be substantially lower than the high-consensus agreements one typically obtains with PFA or JACFEE. That means that even if a separate group of judges provide external consensus for recognizability, stimuli that meet these criteria can still vary considerably in signal clarity. If the stimuli that meet criteria are used within a study, then comparisons are not confounded by signal clarity because stimuli at the same level of signal clarity are used. Across studies, however, the problem of effects confounded by variance in signal clarity remains. That Elfenbein and Ambady reported that the size of the in-group advantage did not differ according to whether the studies used consensus-validated stimuli does not necessarily argue against this notion, because signal clarity may have still differed enough within those studies to produce differential effects on judgments. In short, the fact that the issue of signal clarity was not adequately handled leaves the door open for rival explanations of the data across studies that should be considered prior to accepting conclusions about an in-group advantage.

The Available Evidence

Data From My Previous Research

In the past, I have conducted a number of studies (e.g., Biehl et al., 1997; Matsumoto, 1992) meeting the two methodological requirements described above: balanced designs (American and Japanese observers judging Caucasian and Japanese faces) with stimuli that are equivalent in emotion-signaling properties across encoder cultures. The stimuli used in these studies are from the JACFEE (Matsumoto & Ekman, 1988). It consists of 56 expressions—8 examples of 7 emotions—each portrayed by a different individual. Exactly half are portrayed by Caucasians, the other half by Japanese; also, exactly half are portrayed by men, the other half by women. All faces were reliably FACS-coded to ensure that the muscles innervated in the expressions corresponded to the universal signals of emotion (as depicted by Ekman & Friesen, 1975). Especially relevant to this discussion is the fact that, across all examples of the same emotion, the expressions included exactly the same facial muscles innervated, at exactly the same intensity levels, according to FACS-coding, not a group of observers. If one lined up all eight expressions of any of the seven emotions, which Ekman and I did in the creation of the JACFEE, one would readily see that the expressions—that is, the facial muscles innervated and their intensity levels—do not differ across the eight examples of each emotion.

These unique characteristics of the JACFEE were not produced easily. Approximately 75 individuals were photographed while they portrayed various emotions, either through directed action or spontaneous elicitation. For each expressor, a roll of film with 36 shots was taken per session, and each expressor contributed an average of two sessions. Some expressors provided up to four sessions of shooting. This resulted in the production of approximately 5,000 photographs, each of which was evaluated for possible inclusion. All potential expressions were then FACS-coded, and those with equivalent FACS codes were then included in the final pool of expressions. The final JACFEE was produced with the condition that there be an equal number of expressions posed by Caucasian and Japanese men and women, and that each expressor contribute only one expression to the final set. This process took well over a year to complete, prior to obtaining reliability data through judgments.

These characteristics ensure that differences in judgments across expressors of the JACFEE cannot be attributed to differences in the expressions being judged. If judgment differences across expressors are found, they must be due to something other than the characteristics of the stimuli because the stimuli are equivalent in terms of their emotion-signaling properties. The JACFEE has been used in multiple studies with judges from multiple cultural groups (e.g., see Biehl et al., 1997; Hess, Blairy, & Kleck, 1997), providing ample external evidence for the reliability of the emotions judged in the faces, which further bolsters its psychometric properties (but does not substitute for the validation of the equivalence in physical-signal properties of the stimuli).

Data from several studies using the JACFEE are especially relevant to testing the in-group advantage hypothesis. In one study, 41 American and 44 Japanese judges saw the JACFEE and selected a single emotion-category term from a prescribed list that they thought best portrayed the emotion in the face (Matsumoto, 1992). To compute an accuracy score, a response was coded as 1 when the category selected was that intended; all other responses were coded as 0. A five-way analysis of variance (ANOVA) was computed on these accuracy scores, using judge culture, judge sex, emotion, poser race, and poser sex as the independent variables. The interactions of Judge Culture $\times$ Poser Race, Judge Culture $\times$ Poser Race $\times$ Poser Sex, and Judge Culture $\times$ Poser Race $\times$ Poser Sex $\times$ Emotion were nonsignificant, $F(1, 82) = 0.70$, *ns*; $F(5, 410) = 2.53$, *ns*; and $F(5, 410) = 0.33$, *ns*; respectively. The Judge

Table 2
Average Percentage of Judges Correctly Identifying the Intended Emotion in the JACFEE Expressions, as Reported in the JACFEE Brochure

		Judges (%)	
Emotion	Poser race	United States	Japan
Anger	Caucasian	89.41	74.98
	Japanese	85.40	61.20
Contempt	Caucasian	49.51	70.08
	Japanese	45.43	62.99
Disgust	Caucasian	80.69	71.15
	Japanese	81.04	69.00
Fear	Caucasian	72.45	31.52
	Japanese	82.94	43.68
Happiness	Caucasian	97.56	96.23
	Japanese	98.40	99.32
Sadness	Caucasian	91.38	77.72
	Japanese	94.50	73.54
Surprise	Caucasian	92.65	89.97
	Japanese	93.98	88.38

Note. JACFEE = Japanese and Caucasian Facial Expressions of Emotion.

Culture × Poser Race × Emotion interaction was significant, $F(5, 410) = 4.70$, $p < .001$, but the data indicated that Americans had higher accuracy scores than Japanese regardless of poser race for all emotions except happiness, for which there was no difference between judge cultures (thus contributing to the interaction).

These findings cumulatively argue against an in-group advantage hypothesis. A difference in degree may reflect an in-group advantage effect, if the means are in the direction predicted by the effect. That was not the case here. Furthermore, if such an effect did exist, it would mean that it moderated an already existing general-decoding difference between the judge cultures. This is contrary to Elfenbein and Ambady's (2002) position and essentially leaves unchanged the message that judge–culture differences are primarily decoding differences. In addition, three other studies using essentially the same methodology with American and Japanese judges (total *ns* across these studies: 310 Americans, 261 Japanese) also produced results failing to support this hypothesis (Biehl et al., 1997; Matsumoto et al., in press; Matsumoto, Kasri, & Kooken, 1999).

In another study (Matsumoto, 1993), four American ethnic groups (36 European, 46 Asian, 21 African, and 21 Hispanic) viewed all 56 photos of the JACFEE and made both an emotion category judgment and an intensity rating (using a nine-point scale with scores ranging from 0 to 8). Chi-square analyses of the emotion category judgments using ethnicity as the independent variable indicated no European American–Asian American differences in judgments across the photos judged. Recoding of the emotion category data into accuracy scores and computation of ANOVAs also indicated no significant effects between these judge groups. These nonfindings again argue against a possible in-group advantage effect.

Examination of the intensity ratings is also revealing. Emotion scales receiving the highest intensity rating can be interpreted as the most salient emotion judged, thereby providing a measure of accuracy of judgment of the emotion intended. Analyses of the intensity ratings indicated that there were no differences between European American and Asian American judges for either Caucasian or Japanese expressors. This argues against an in-group advantage.

The data reported in the brochure accompanying the JACFEE expressions are also relevant. They include the data reported in Matsumoto (1992) and described above, categorical judgment data from an additional 70 Americans, and intensity ratings by another 124 American and 110 Japanese judges (Matsumoto & Ekman, 1989). The intensity ratings were scalar ratings of seven emotion categories using a nine-point scale. The emotion scale receiving the highest intensity rating was considered the most salient emotion portrayed, and the percentage of observers giving each photo the intended emotion term the highest rating was computed. I averaged the percentage agreement data across the Caucasian and Japanese posers within each emotion and separately for American and Japanese judges (see Table 2). For anger, disgust, fear, sadness, and surprise, American judges had higher agreement levels on Caucasian expressions than did the Japanese judges. The same differences, however, occurred for the Japanese expressions. For contempt, the cultural difference was in the opposite direction, but nevertheless the difference was the same for both Caucasian and Japanese expressors. For happiness, there was no discernible cultural difference for either Caucasian or Japanese expressors.

These data argue unequivocally against an in-group advantage hypothesis and for a decoding effect in judgment that is applied regardless of the expressor race being judged. These studies meet the two methodological criteria I described above, and their findings paint a considerably different picture than that claimed by Elfenbein and Ambady.[1,2]

[1] The in-group advantage may actually exist for cross-cultural comparisons other than American versus Japanese. However, I know of no other studies comparing other cultural groups that meet the stimulus requirements I describe here, probably because no such stimuli other than the JACFEE exist. Even then, Elfenbein and Ambady (2002) reported balanced studies that used stimuli that did not meet the criteria, and I believe an analysis of them indicates that the in-group effect is negligible for the vast majority of them (see below).

[2] It may also be possible that people in different cultures do not typically express emotions exactly the same way, using the same muscles in different ways or different muscles altogether, so that differences across groups in emotional expression in the stimuli that might lead to an in-group advantage in emotion recognition are eliminated by artificially homogenizing the stimuli. Addressing this notion properly would require researchers in the field to first test for cultural differences in actual behaviors in spontaneous emotion-eliciting situations and then to systematically vary the expressions used in a subsequent judgment study, all the while ensuring that the expressions used as stimuli are equivalent in terms of emotion-signaling properties across expressor cultures. This has not been done. These points are related to issues of signal clarity I raise below and, in any case, do not detract from my position that the evidence available to date that meet the methodological criteria I describe does not support the in-group advantage hypothesis.

Examining the Balanced Designs in Elfenbein and Ambady's (2002) Analyses

An examination of only those studies with balanced designs from Elfenbein and Ambady's own data set also produces a considerably different outlook on the in-group advantage hypothesis, even if one were to set aside for a moment the issue of stimulus equivalence across expressor cultures. In their article, they considered three sources of evidence, including studies for which percentage effect sizes could be computed, studies for which an effect size (r) could be computed, and studies for which an interaction (F) between encoder and decoder culture could be computed. In this section I consider the data from balanced designs for each source.

First, I eliminated all studies from their Table 1 listing percentage accuracy effect sizes that were not balanced. The mean in-group advantage effect size for the remaining balanced studies is .03 ($k = 37$ data samples, $SD = .17$; $SE = .03$, 95% confidence interval (CI) $= -.02 < M < .09$). This means that there is an average difference of 3.0% in emotion recognition accuracy in the direction of the supposed in-group advantage. This statistic is substantially different than the 9.3% they reported (Elfenbein & Ambady, 2002, Tables 4 and 6), and in no case would it produce enough of a difference to obtain statistical significance of recognition beyond chance levels in any study. Additionally, a recognition accuracy difference of 3.0% may not be meaningful.

Reexamination of the studies in which an effect size (r) could be calculated (Elfenbein & Ambady, 2002, Table 2) by only including the 14 studies that involved balanced designs produces the same finding. The mean Z_r for these 14 studies was .242. This statistic, however, was almost entirely carried by two studies that had extremely large Z_rs—Albas, McCluskey, and Albas (1976) and Ricci Bitti, Brighetti, Garotti, and Boggi-Cavallo (1989). In fact, when those two studies were not considered, the mean Z_r was .09 ($SD = .19$, $SE = .06$, 95% CI $= -.02 < M < .22$). This is substantially different than the unweighted average effect size of .25 reported by Elfenbein and Ambady (2002, Tables 4 and 8). Although a Z_r of .09 can represent a reliable, but small, effect, given large sample sizes and relatively small variances, the point I make here is that the size of the effect is dramatically different than that reported by Elfenbein and Ambady.

A closer look at the Encoder × Decoder interaction F data (Elfenbein & Ambady, 2002, Table 5) is also revealing. Four studies—Albas et al. (1976); Bond, Omar, Mahmoud, and Bonser (1990); Kretsch (1968); and Shimoda, Argyle, and Ricci Bitti (1978)—really carry the effects they present. For the remaining 12 studies, the mean p was .085, and the mean r^2_{contrast} was .02 (I chose to square the contrast r). Again, this is substantially different than what is reported by Elfenbein and Ambady.

I recognize that I have disregarded six studies here, which is not unsubstantial. My point, however, is that when one examines the effects produced by these six studies in relation to the 101 samples across the three analyses presented by Elfenbein and Ambady, one gets a very different sense of the available data. These studies need to be evaluated for whether they established equivalence in emotion-signaling properties; if they did not, the differences are confounded by stimulus differences across expressors. Moreover, five of the six studies involved the use of voice stimuli, which is related to signal clarity, as I discuss in more detail below. Thus, when one examines only the balanced studies reported by Elfenbein and Ambady, granting those studies amnesty from the requirement that stimuli need to be equivalent in emotion-signaling properties across encoder cultures, the effect is nearly negligible and, in any case, substantially different than that reported by Elfenbein and Ambady.

Conditions Under Which In-Group Advantage Effects May Occur

Diversity Among the Balanced Designs

Despite this strong evidence against the in-group advantage hypothesis, some studies reported by Elfenbein and Ambady (2002) do appear to support it. The first way to examine the validity of these findings would be to investigate the nature of the emotion stimuli used and to determine whether the physical-signaling properties of the stimuli related to emotion were exactly equivalent across the expressor cultures. Again, judgments by external samples cannot be used to make this determination; it would have to be made by actual measurement of the physical properties (muscle innervation, voice characteristics, etc.) of the stimuli involved. If this level of stimulus equivalence cannot be established, physical properties cannot be used to justify interpretations concerning an in-group advantage.

If one grants stimulus equivalence across those studies (which is a big assumption), then the question to raise is, Why does the effect exist for those studies but not for all the others? Inevitably, one comes to the possibility that some study characteristics may produce an in-group advantage. It is therefore imperative that researchers examine the diversity of study characteristics—among the balanced designs—to investigate their possible associations with the in-group effect.

Every study is inherently different; therefore, studies naturally differ according to a number of characteristics. In fact, Elfenbein and Ambady (2002) did a fairly good job of identifying major study characteristics and coding them for their various analyses. For instance, across the 16 studies listed in their Table 5, stimuli were presented via a number of different channels, which break down as follows: voice, 4 studies; full video with sound, 1 study; voice and video, 2 studies; facial photographs, 8 studies; and silent video, 1 study. Elfenbein and Ambady did not examine how the in-group advantage varies according to these characteristics, or others, within the balanced studies. If they did, such analyses might provide some insight into a possible in-group advantage. For instance, of the four studies in their Table 5 in which the in-group advantage was strong, it is interesting to note that all of them included voice as a channel of encoding. Eight of the remaining 12 studies used judgments of facial photographs. In fact, the mean p value for studies using facial photographs was .08, and the mean r^2_{contrast} was .02. These data suggest that if an in-group advantage exists, it may exist when voice is judged. Voice was also the stimulus channel in one of the two studies carrying the in-group advantage effect among the studies in which an effect size (r) could be computed. Thus, it appears to carry the effect in five of the six studies in which an effect apparently occurs.

My earlier comments concerning the equilibration of the emotion-signaling properties of the stimuli across expressor cultures in balanced studies hold true here for studies using voice as well. A number of person characteristics may be recognized from voice, including race and ethnicity, age, gender, emotion, and

others (see Ekman, 1979, for a more extended discussion). Researchers need to measure the actual physical-signal properties of the voice that are related to each of these messages and ensure that stimuli that vary across expressor cultures do not vary on the physical properties related to emotion signaling yet systematically vary according to race and ethnicity. Elfenbein and Ambady (2002) needed to determine whether this level of equilibration was achieved in the studies prior to evaluating their contribution to the in-group advantage hypothesis. If there was no verification of the equivalence of the physical emotion-signaling properties of the voice in those studies, the effects observed in them may have been due to the nonequivalence in the stimuli, not to any in-group advantage.

Similarly, other stimulus channels may be examined for their possible contribution to an in-group advantage. These, too, would need to meet the methodological requirements concerning stimuli described above. Even if one grants stimulus equivalence, however, a question arises concerning whether there currently is adequate representation of various study characteristics (and thus sample size) in those analyses to draw any serious conclusions about the possible in-group effect. For example, in the data presented by Elfenbein and Ambady (2002), there is only one balanced study with full video with sound, and another one with silent video. There are only two studies with voice and video. The four studies with voice and eight studies with facial photographs only begin to represent numbers of studies that can be used to adequately and reliably draw conclusions about an effect. If there were adequate numbers of studies that used each of the various types of stimuli channels among the balanced studies, if they met the stimulus requirements I describe, and if their associations with the in-group advantage were tested and documented, then and only then could a conclusion be drawn on the basis of evidence that meets a level for scholarly acceptance. This evidence does not exist.

A Possible Explanation of In-Group Advantage Effects if They Existed

If an in-group advantage exists, it may be associated with stimulus ambiguity (i.e., signal clarity), and an inverted U may characterize their relationship. Possible in-group advantage effects may be minimized when stimuli are very clear or not clear at all. When stimuli are very clear, they may be easily recognized by all groups of people, regardless of any biases in decoding. This may certainly explain why such effects have often not been obtained in our previous studies using the PFA or JACFEE. When stimuli are not clear at all, the in-group advantage may also be neutralized because they are too ambiguous to be recognized by anyone. Thus, the in-group advantage may occur when stimuli have some midrange value of signal clarity. That is, if judges rate stimuli with a moderately high signal–noise ratio, they may rely on cues or processes that are idiosyncratic to their cultural group when judging emotions.

Elfenbein and Ambady (2002) well recognized that signal clarity affects accuracy. For instance, they reported that cross-cultural accuracy was lower for studies that used tone of voice than it was for studies that used other channels and that dynamic channels overall were less accurately recognized across cultures than the static channels. Moreover, the relationship between signal clarity and a possible in-group advantage is suggested by their findings. For example, they noted that cross-cultural accuracy on the PONS test varied across channels from 19.6% for content-filtered sound to 73.0% for facial video with random-spliced sound. Effect sizes related to in-group advantage differed according to channel, and follow-up analyses led Elfenbein and Ambady to conclude that "providing information from additional channels of communication can reduce such cross-cultural differences" (p. 221; i.e., reduce the in-group advantage).

Similarly, in their results concerning specific emotions, Elfenbein and Ambady (2002) reported that the "in-group advantage was lowest with happiness and anger, whereas it was greatest with fear and disgust" (p. 222). Happiness is often the most easily recognized emotion, thus having the greatest signal clarity, whereas accuracy levels for fear are generally much lower. These data are also suggestive of the possible role of signal clarity moderating the in-group advantage effect sizes.

In reporting findings related to differences in emotions across channels, Elfenbein and Ambady (2002) reported that "happiness was most accurately recognized from the face, but the least recognized from the voice" (p. 222). The in-group advantage was correspondingly smallest for happiness in the face but highest for happiness in the voice. These findings again support contentions for the moderating effects of signal clarity on the possible in-group advantage.

To be sure, some findings argue against the moderating role of signal clarity. For instance, contempt was associated with a low in-group advantage, but it was also the most poorly recognized emotion, having a 43.2% accuracy level. The fact that there were no differences in In-Group Advantage $\times$ Research Team also argues against this notion, as there generally are differences in the signal clarity of stimuli across teams. Still, there are sufficient findings reported here to suggest an alternative view of the data Elfenbein and Ambady presented, which unfortunately remains untested formally in the studies they reviewed.

Conclusion

In this article, I have discussed two methodological requirements for studies to test adequately the in-group advantage hypothesis and an additional requirement in reviewing multiple-judgment studies and examining variance in judgment effects across those studies. I showed that when previously published studies from my laboratory that met the two criteria are examined, the effect is nonexistent. I also showed that when only balanced studies in Elfenbein and Ambady's (2002) review are examined, the in-group advantage effect is negligible. The few balanced studies that Elfenbein and Ambady reported that support the in-group advantage hypothesis need to be examined for whether the stimuli that were used met the requirements for stimulus equivalence across expressor cultures; if they did not, their data cannot be used to support any contention of cultural differences in judgments, let alone the in-group advantage hypothesis. If stimulus equivalence can be granted in these studies, and that is a big if, then the role of signal clarity needs to be explored in possibly moderating such effects.

Elfenbein and Ambady's (2002) conclusions concerning this issue are problematic because they attempted to shut the door on other rival, alternative explanations of cultural differences in judgments even though the evidence does not allow that door to be closed. Not only did they throw out the possible effect of display

rules on decoding, but they also discarded the possible influence of language differences in emotion taxonomies on the judgment process (although they do contradict themselves later in the discussion by discussing the potential importance of linguistic and conceptual reasons to contribute to the in-group advantage). I do not believe it reasonable to discard such possibilities when first, the evidence does not support such a position and second, there are so few studies that have been designed to specifically address them in the first place. In fact, I know of no cross-cultural study to date that has linked display rules with judgments of emotion.

There are other problems with their conclusions. Researchers and theorists may easily, and mistakenly, swallow the notion of in-group advantage in decoding. Indeed, it is an attractive view for those who wish to propound cultural differences, and it has a catchy title. Such a view of the judgment process that is clearly not supported by data can easily lead to theories that advocate "fundamental" intergroup differences which, in turn, can easily polarize cultures against each other and contribute to the academic construction of walls and barriers among people that may be unnecessary and in any case unjustified by the literature. Clearly, such developments have serious consequences for intergroup and interpersonal relations as well.

At the same time, I am not saying that the in-group advantage absolutely does not exist, nor do I intend to suggest that the studies I have conducted, although meeting the methodological criteria I describe, should close the door on this issue. As I described above, there may be conditions in which such an effect exists, and there are certainly many other types of studies that can and should be done to test the variety of cultural differences in judgment processes, such as those raised in Elfenbein and Ambady's (2002) Footnote 3.

The evidence available to date simply does not support Elfenbein and Ambady's (2002) conclusions concerning the in-group advantage hypothesis and psychologists need not be so hasty in accepting their claims about it. The strength of a meta-analysis is determined in large part not only by the methodological quality of the studies included in the analysis, both within each study and across studies, but also by the methodological appropriateness of the studies to address the pertinent questions raised by the meta-analysis.

References

Albas, D. C., McCluskey, K. W., & Albas, C. A. (1976). Perception of the emotional content of speech: A comparison of two Canadian groups. *Journal of Cross-Cultural Psychology, 7,* 481–490.

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.

Bond, C. F., Omar, A., Mahmoud, A., & Bonser, R. N. (1990). Lie detection across cultures. *Journal of Nonverbal Behavior, 14,* 189–204.

Ekman, P. (1979). Facial signs: Facts, fantasies, and possibilities. In T. Sebeok (Ed.), *Sight, sound, and sense* (pp. 124–156). Bloomington: Indiana University Press.

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

Ekman, P., & Friesen, W. V. (1976). *Pictures of facial affect.* Palo Alto, CA: Consulting Psychologists Press.

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

Ekman, P., Friesen, W. V., & Ellsworth, P. (1972). *Emotion in the human face: Guide-lines for research and an integration of findings.* New York: Pergamon Press.

Elfenbein, H. A., & Ambady, N. (2002). On the universality and cultural specificity of emotion recognition: A meta-analysis. *Psychological Bulletin, 128,* 203–235.

Hess, U., Blairy, S., & Kleck, R. (1997). The relationship between the intensity of emotional facial expressions and observers' decoding. *Journal of Nonverbal Behavior, 21,* 241–257.

Kretsch, R. A. (1968). *Communication of emotional meaning across national groups.* Unpublished doctoral dissertation, Columbia University, New York.

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

Matsumoto, D. (1993). Ethnic differences in affect intensity, emotion judgments, display rule attitudes, and self-reported emotional expression in an American sample. *Motivation & Emotion, 17,* 107–123.

Matsumoto, D. (2001). Culture and emotion. In D. Matsumoto (Ed.), *The handbook of culture and psychology* (pp. 171–194). New York: Oxford University Press.

Matsumoto, D., Consolacion, T., Yamada, H., Suzuki, R., Franklin, B., Paul, S., et al. (in press). American–Japanese cultural differences in judgments of emotional expressions of different intensities. *Cognition & Emotion.*

Matsumoto, D., & Ekman, P. (1988). *Japanese and Caucasian Facial Expressions of Emotion and Neutral Faces (JACFEE and JACNeuF)* [Slides]. (Available from 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 & Emotion, 13,* 143–157.

Matsumoto, D., Hall, J. A., & Weissman, M. (2001). *Sex differences in judgments of multiple emotions from facial expressions.* Manuscript submitted for publication.

Matsumoto, D., Kasri, F., & Kooken, K. (1999). American–Japanese cultural differences in judgments of expression intensity and subjective experience. *Cognition & Emotion, 13,* 201–218.

Matsumoto, D., LeRoux, J. A., 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.

O'Sullivan, M. (1982). Measuring the ability to recognize facial expressions of emotion. In P. Ekman (Ed.), *Emotion in the human face* (pp. 281–314). New York: Cambridge University Press.

Ricci Bitti, P. E., Brighetti, G., Garotti, P. L., & Boggi-Cavallo, P. (1989). Is contempt expressed by pancultural facial movements? In J. P. Forgas & J. M. Innes (Eds.), *Recent advances in social psychology: An international perspective* (pp. 329–339). Amsterdam: Elsevier.

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.

Shimoda, K., Argyle, M., & Ricci Bitti, P. E. (1978). The intercultural recognition of emotional expressions by three national racial groups: English, Italian and Japanese. *European Journal of Social Psychology, 8,* 169–179.

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

Received September 5, 2001
Revision received October 15, 2001
Accepted October 15, 2001 ■