Statistically significant differences in culture means may or may not reflect practically important differences between people of different cultures. To determine whether differences between culture means represent meaningful differences between individuals, further data analyses involving measures of cultural effect sizes are necessary. In this article the authors recommend four such measures and demonstrate their efficacy on two data sets from previously published studies. They argue for their use in future cross-cultural research as a complement to traditional tests of mean differences.

# DO BETWEEN-CULTURE DIFFERENCES REALLY MEAN THAT PEOPLE ARE DIFFERENT?
## A Look at Some Measures of Cultural Effect Size

DAVID MATSUMOTO
ROBERT J. GRISSOM
*San Francisco State University*
DALE L. DINNEL
*Western Washington University*

Euro-Americans have higher independent self-construal scores than Hong Kong Chinese, while Hong Kong Chinese have higher interdependent self-construal scores than Euro-Americans.

Singelis, Bond, Sharkey, and Lai, 1999

Estonians exhibit greater tendencies for legitimization, relativism, absoluteness, and universality in their moral reasoning than Finns.

Keltikangas-Jarvinen, Terav, and Pakaslahti, 1999

Chinese in Singapore are more compliant than Chinese in Taiwan in response to requests from friends; but, Singaporean Chinese are also more direct in refusing compared to Taiwanese Chinese.

Bresnahan, Ohashi, Liu, Nebashi, and Liao, 1999

Korean children are more interested in numbers than Americans; American children are more interested in words and ideas than Japanese children. Japanese children are less interested than Koreans and Americans in being alone; but Americans are more interested in people than the two Asian groups.

Henderson, Marx, and Kim, 1999

AUTHORS' NOTE: Address correspondence to David Matsumoto, Department of Psychology, San Francisco State University, 1600 Holloway Avenue, San Francisco, CA, 94132; telephone (415) 338-1114, fax (510) 217-9608, e-mail dm@sfsu.edu. We would like to thank Chu Kim, Laila Ansari, Brenda Franklin, and Sunita Paul for their assistance in our general research program.

JOURNAL OF CROSS-CULTURAL PSYCHOLOGY, Vol. 32 No. 4, July 2001 478-490

> In dealing with conflict, Japanese use more avoidance tactics than Americans, while Americans use more assertion. Further, Japanese use avoidance relatively more strongly, but assertion and third party intervention less strongly than do Americans.
>
> Ohbuchi, Fukushima, and Tedeschi, 1999

> Hungarians engage in more self-disclosing to partners, friends, and parents than Russians and Georgians, but less so to acquaintances.
>
> Goodwin, Nizharadze, Luu, Kosa, and Emelyanova, 1999

Cross-cultural comparison is a cornerstone of cross-cultural research, as witnessed above in a summary of findings from selected articles published in two recent issues of this journal. The importance of such comparisons is not debated; they are important in challenging findings from monocultural studies, in the construction of culturally relevant theories, and in the design of intercultural training and education programs. A standard methodology in conducting these comparisons begins with the selection of measures of psychological constructs that produce quantitative data from two or more cultures or countries. Differences are tested by comparing variance between the culture means relative to the variance within the cultures, typically using *t* or *F* tests. When the chance probability of obtaining *t* or *F* values is sufficiently low (≤ 5%), the result is considered statistically significant.

But does statistical significance reflect differences between people of the different cultures? Not necessarily. The sole computation of *t*s or *F*s precludes our ability to interpret meaningful differences between people, because *p* values merely indicate the strength of the evidence against the null hypothesis of no difference between population means. Statistical significance, assuming no Type I error, only reflects some unknown, nonzero difference between the population means. And the larger the sample sizes, the easier it is for smaller differences to become statistically significant. Therefore, a statistically significant difference may actually reflect a trivially small difference between population means. Interpretations of cultural differences between people based on statistically significant findings may be based on practically insignificant differences between means.

"Practically insignificant" means that the nonzero difference between culture means is so small that it is of little or no practical significance. A synonymous phrase would be "substantively insignificant." In the phrase "statistically significant," "significant" means "signifying," signifying that there is sufficient evidence against a null hypothesis of no difference. In the phrase "practically significant," "significant" means "large" or "important." An example would be a statistically significant difference between boys and girls that is so small that it does not warrant separate instructional practice for teaching.

A significant *t* or *F* in a cross-cultural study, therefore, does not necessarily mean that most people of one culture have an appreciably greater score than most people of another, nor that the average person from one culture will have a substantially higher score than the average person from the other, nor that a randomly selected individual from one culture will very likely have a higher score than a randomly selected individual from the other. In short, statistical significance may not have any meaningful implications for predicting differences on the level of individuals, and those who consider it as indicating meaningful differences between people across cultures may, in fact, create or perpetuate stereotypes about those people and cultures. Although most researchers are aware of these limitations of significance testing, most reports rely solely on them, largely ignoring the practical significance of the results.

For example, suppose a study comparing Americans and Japanese on individualism versus collectivism found that Americans had significantly higher means on individualism (in actuality, recent reviews indicate that exactly the opposite is true; see Matsumoto, 1999; Takano & Osaka, in press). Let us also say, however, that the significant mean differences were not associated with large differences in the samples. If we concluded that Americans are generally individualistic (I) and Japanese are generally collectivistic (C), we would be making a mistake; it may very well be that there are only minor differences between them on IC despite the statistically significant differences in the means.

These possible mistakes have several implications. Theoretically, they may lead to the construction of knowledge based on stereotypes. Research is then created to test this bias, which is perpetuated because of the continued use of limited data analytic techniques. Practically, programs for intercultural sensitivity, training, competence, adjustment, and the like are based on cultural stereotypes, providing consumers with incorrect guidelines that may be more harmful than helpful. The notion of American individualism relative to Japanese collectivism, which in fact is not supported by research, for instance, is an example of a construct widely used in theories, research design, and applied program creation and implementation.

Discussions concerning the limitations of *t*, *F*, and of null hypothesis significance tests (NHST) are not new. One of the first concerns about NHST is its dependence on sample size, which is also relevant for cross-cultural research. In addition, Cohen (1962) pointed out the utility of power analysis in psychological research and highlighted the high error rates that are typically associated with NHST. Mistakes commonly cited include accepting the null hypothesis when it fails to be rejected, automatically interpreting rejected null hypotheses as theoretically or practically meaningful, and failing to consider the likelihood of Type II errors (Loftus, 1996; Shrout, 1997; see also Wilcox, 1998). Some writers (e.g., Hunter, 1997) have recommended outright bans against NHST, arguing that error rates are as high as 60%, not the 5% traditionally thought.[1] Others have argued for the continued use of NHST in addition to the incorporation of effect size statistics and confidence intervals (e.g., Abelson, 1997; Harris, 1997).

To examine the degree to which cross-cultural data are indicative of meaningful differences between individuals based on culture-group membership, further analyses are needed to estimate the otherwise unknown degree of difference between two cultures' populations. (We limit ourselves in the remainder of this article to a discussion of two-sample data.[2]) Collectively, these techniques are known as measures of effect size, and when used in cross-cultural research, we refer to them as cultural effect size. Although their importance has long been recognized, their use in cross-cultural research is still quite limited. We argue here, therefore, for their incorporation in cross-cultural work, a position that is consistent with the recommendation of the APA's Task Force on Statistical Inference (Wilkinson, The Task Force on Statistical Inference, & American Psychological Association Board of Scientific Affairs, 1999).

There are many measures of effect size available (Cohen, 1988; Feingold, 1992, 1995; Feingold & Mazzella, 1998; Hedges & Olkin, 1985; Rosenthal, 1991; Wilcox, 1997) as well as informative overviews and summaries of them (e.g., Rosenthal, Rosnow, & Rubin, 2000). Some provide information that is highly relevant in cross-cultural studies, wherein researchers are concerned with the representation of group-level cultural differences on the individual level. By "individual level," we allude to the fact that significance tests only tell us what proportions of *t* or *F* distributions are beyond the value attained by our *t* or *F* result. The measures of effect size recommended here do not address proportions of such distributions; rather, they inform us about outcomes in terms of proportions of members of a culture or in terms of the performance of the average member of a culture relative to proportions of members of another culture.

In the remainder of this article, we recommend four measures of effect size that we believe are most appropriate for cross-cultural research. We selected them for their ease of computation and interpretation and their relevance. Using these measures, we reanalyze data from two previously published studies to demonstrate their efficacy. Finally, we argue for the incorporation of effect size statistics to complement traditional NHST.

## FOUR MEASURES OF CULTURAL EFFECT SIZE

### THE STANDARDIZED DIFFERENCE BETWEEN TWO SAMPLE MEANS

The first measure is the well-known standardized difference between two population means, estimated by $g = (\overline{X}_A - \overline{X}_B)/s_p$ in the case of comparing two means from a two-group or multigroup study. This estimate assumes homogeneity of variance as it pools variances to obtain $s_p$ or $\sqrt{MS_w}$ (Hedges & Olkin, 1985) and normality when interpreted in terms of the percentile attained by the average-performing member of one culture with respect to the distribution of scores of the other. It allows researchers to estimate what percentage of people in a culture has higher scores than the average member of a statistically significantly lower scoring culture.

For example, if the mean of culture A is statistically significantly higher than the mean of culture B and $g = +1.00$, an average member of culture A scores 1.00 standard deviation unit higher than an average member of culture B. A standard score of +1.00 lies at the 84th percentile. Therefore, we can conclude that the average member of culture A is outscoring 84% of the members of culture B. A value of $g$, however, that is only slightly above 0 indicates that average-scoring members of the statistically significantly higher scoring culture may be outscoring only slightly more than 50% of the lower scoring culture. Thus, the statistically significant difference may be associated with little or practically no appreciable difference on the level of the individuals. Cohen (1988) suggests that values around .20, .50, and .80 reflect small, moderate, and large differences, respectively. (Cohen's $d$ uses the assumed common population standard deviation, [sigma], in the denominator instead of $s_p$.)

### PROBABILISTIC SUPERIORITY EFFECT SIZE MEASURE

The second measure is the probability that a randomly sampled member of population a will have a score that is higher than a randomly sampled member of population b, $\Pr(X_a > X_b)$. If there is no difference between the two distributions, $\Pr(X_a > X_b) = .50$. The more superior that distribution a is compared to distribution b, the more $\Pr(X_a > X_b)$ moves away from .5 toward 1. (Computation of these statistics may result in values less than .50 if culture a is the lower scoring culture.)

When raw data are not available, $\Pr(X_a > X_b)$ can be estimated by the common language (CL) effect size statistic (McGraw & Wong, 1992) that assumes normality and homogeneity of variance. It is based on a $z$ score, $z_{cl} = (\overline{X}_a - \overline{X}_b)/\sqrt{S_a^2 + S_a^2}$, and is the proportion of area under the standardized normal curve that is below the obtained value of $z_{cl}$. For example, if $z_{cl} = -1.00$ or $+1.00$, $\Pr(X_a > X_b)$ is estimated to be approximately .16 or .84, respectively. The latter would indicate that 84% of the time a randomly sampled member of culture A will outscore a randomly sampled member of culture B. The closer $z_{cl}$ is to 0, however, the less different are the two distributions; that is, the chance that a randomly sampled member of culture A will score higher than a randomly sampled member of culture B approaches .50.

When raw data are available, the recommended unbiased estimator of $\Pr(X_a > X_b)$ is the "probability of superiority" estimator (Grissom, 1994), PS = $U/(mn)$, where $U$ is the Mann-Whitney statistic and m and n are sample sizes. The $U$ is a count of the number of members of one sample whose scores outrank those of the other (assuming no ties or equal allocation of ties to each sample). For example, suppose that $m = n = 10$ (although sample sizes need not be equal) and that the score obtained by each member of sample $m$ is compared to the score obtained by each member of sample $n$, resulting in $mn = 10 \times 10 = 100$ such comparisons. Suppose further that 70 comparisons resulted in the sample $m$ member having a score superior to that of the sample $n$ member. Then $U/(mn) = 70/100 = .70$, the proportion of times that a member of sample $m$ has a score that is superior to that of a member of sample $n$. A proportion in a sample estimates a probability in a population. In the present case the estimate of $\Pr(X_a > X_b) = .70$. The PS has the advantage over CL of not assuming normality or homogeneity of variance. (For more information on the PS, see Grissom, 1994; for constructing confidence intervals for $\Pr(X_a > X_b)$, see Wilcox, 1997, 1998.) Also, $\Pr(X_a > X_b)$ can be estimated from values of $g$ using a table in Grissom (1994; $g$ is denoted "delta" in Grissom), assuming homogeneity of variance.

## COHEN'S *U1*

The third measure is Cohen's (1988) *U1*, the percentage *non*overlap of two distributions. Assuming normality, homogeneity of variance, and populations of equal size, *U1* can be estimated from values of $d$ (or Hedge's $g$) using a table in Cohen (1988). For example, if $d = 0$, the estimate is $U1 = 0\%$; there is 0% nonoverlap, 100% overlap, between the two cultures' scores. If $d = +1.00$, there is an estimated 55.4% nonoverlap, 44.6% overlap. Although Cohen's *U1* assumes equal population sizes, one can use these values to compare theoretical equal-size versions of the two cultures' populations.

## THE POINT BISERIAL CORRELATION

The final measure is the point biserial correlation between culture groups, coded dichotomously into any two values for the two cultures (the sign of the correlation will depend on which culture has higher scores; the value, however, will be unaffected), and the dependent variable $Y$. This is simply the Pearson $r$ when $Y$ is continuous and $X$ is a dichotomy. This measure is easy to interpret because it uses the familiar scale of $r$, 0 to 1. Values closer to 1 indicate substantial differences between cultures; values closer to 0 indicate minimal or even negligible differences, regardless of statistical significance.

Although the value of $r_{pb}$ does not depend on homogeneity of variance, the result of a $t$ test testing whether $r_{pb}$ is significantly different from 0 can be affected by heterogeneity of variance. When SS values are based on two equal-sized groups, $r_{pb} = \sqrt{SS_B / SS_T}$. Unequal sample sizes attenuate $r_{pb}$. The attenuation-corrected $r_{pb}$, denoted $r_c$, is given by $r_c = ar_{pb} / \sqrt{(a^2 - 1)r_{pb}^2 + 1}$, where $a = \sqrt{.25 / pq}$, and $p$ and $q$ are the proportions of the total participants that are in each group (Hunter & Schmidt, 1990).

The point biserial correlation is preferable to other correlational measures of effect size that estimate the proportion of variance explained, such as eta squared, because the squaring nature of the latter results in a directionless measure and creates an impression that effect sizes are smaller than they are (Rosenthal & Rubin, 1982). For instance, suppose that 100 members of one culture and 100 members of another culture were categorized as either individualistic or collectivistic (IC). Suppose further that the coefficient of determination, $r_{pb}^2$, were found to be .10. One might conclude that culture is not an important predictor of IC because culture "only explains" 10% of the variance in the dependent variable. In this example, however, $r_{pb} = (.10)^{1/2} = .32$ which, under a simplifying assumption, is roughly equivalent to $g = .68$, a moderately large culture effect size by Cohen's (1988) criteria.

Moreover, using Rosenthal and Rubin's (1982) binomial effect size display (BESD), a technique that allows for the conversion of an effect size into prediction of group membership, the 10% of the variance explained by culture in these data translates to 66% of the members of one sampled culture being found to be collectivistic, whereas 34% of the other culture was so categorized. Thus, when we examine these data in terms of individual members of cultures, we find that the relatively small $r_{pb}^2 = .10$ translates to nearly twice as many members of one culture falling into a category than members of another culture.

## SUMMARY

Several sources (e.g., Cohen, 1988; Dunlap, 1999; Grissom, 1994) describe the relationships among the measures recommended here, and we summarize their major characteristics in Table 1. The measures $g$ and $r_{pb}$ show the direction of cultural difference by their sign and by which culture's mean is coded 1 or produces the first mean in the calculation. The PS and CL reflect direction of difference by whether they are below or above .50 and by which culture is designated culture a. The *U1* measure shows the direction of difference according to the value of *U1* and which culture has the higher mean.

Although the transition from group to individual level interpretations is important in cross-cultural work, there is no universally accepted, objective standard of how large an effect has to be in order to be considered meaningful. Cohen's (1988) criteria for small, medium, and large effects in terms of values of $d$ provide some rough benchmarks. Because the measures of effect size presented here can readily be translated to $d$, $r_{pb}$, CL, or PS, any author or reader of a cross-cultural study can translate any of our recommended measures into one of the categories of small, medium, or large effect, or estimates involving proportions of individuals, if desired.

Our position, however, is that how authors characterize their effects as small, medium, or large (or meaningful or not meaningful) is irrelevant provided that they report the appropriate effect size statistics that allow readers to make their own interpretations of meaningfulness. $F$, $t$, or $p$ do not provide readers the information for determining if their own standard of a meaningful difference between cultures has been attained; reporting measures of effect size permits readers to evaluate the effect in terms of one's own sense of meaningfulness. In the example given previously, $r_{pb} = .32$ can translate to a binomial effect size display showing that 66% of one culture are collectivistic whereas 34% of another culture are so. One reader may interpret this result as meaningful, another as not. We prefer not to provide such guidelines for interpretation; instead, our thrust is to suggest solely that measures such as those recommended here be estimated and reported, so that readers can make their own evaluations. In a sense, reporting effect sizes "democratizes" the evaluation of meaningfulness.[3]

**TABLE 1**

**Summary of the Major Characteristics of the Four Effect Size Estimators Presented**

| *Effect Size Measure* | *Estimator* | *Assumption(s), Limitation(s)* | *Use* |
|---|---|---|---|
| Standardized difference between two cultures' populations' means | $g = (\overline{X}_A - \overline{X}_B)/s_p$ | Normality & homogeneity of variance | Tells how many *SD* units the mean of culture population A is above or below the mean of culture population B. Under normality, also tells what percentage of one population's members are outscored by the mean-scoring members of another population. |
| $Pr(X_a > X_b)$ | Common language (CL) method: The proportion of area under the standardized normal curve that is below the obtained value of $z_{cl} = (\overline{X}_a - \overline{X}_b)/\sqrt{S_a^2 + S_b^2}$<br>Probability of superiority (PS) method: PS = *U*/*mn* | CL assumes normality and homogeneity of variance; PS makes no assumptions | Tells the probability that a randomly sampled member of culture population A will outscore a randomly sampled member of culture population B. |
| Cohen's *U1* | Estimated from table in Cohen (1988) | Normality, homogeneity of variance, populations of equal size | If the two cultures are (theoretically) of equal size, tells the percentage of nonoverlap of their distributions. |
| Point-biserial correlation | Pearson *r* when X is the dichotomy of membership in culture A or B and Y is a continuous dependent variable, or by $r_{pb} = \sqrt{SS_b / SS_T}$<br>See text for correction when sample sizes are unequal. | A *t* test for the significance of $r_{pb}$ assumes homogeneity of variance | Measures degree and direction of correlation between culture group membership and the dependent variable. |

TABLE 2
**Reanalysis of Matsumoto and Ekman, 1989**

| | *United States* (n = *124)* | | *Japan* (n = *110)* | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| *Emotion & Scale*[a] | M | SD | M | SD | Hedge's g | *PS*[b] | *CL*[b] | U1(%) | $r_{pb}$ |
| Anger | 6.47 | 0.97 | 4.58 | 1.53 | 1.47 | .84 | .85 | 68.1 | .35 |
| Contempt | 2.33 | 1.74 | 4.32 | 1.71 | –1.16 | .21 | .21 | 58.9 | .25 |
| Disgust | 5.81 | 1.47 | 5.43 | 1.80 | 0.23 | .56 | .57 | 14.7 | .01 |
| Fear | 5.79 | 1.37 | 3.60 | 1.51 | 1.52 | .86 | .85 | 70.7 | .37 |
| Happiness | 7.26 | 0.72 | 6.62 | 1.11 | 0.68 | .67 | .68 | 43.0 | .10 |
| Sadness | 6.21 | 1.25 | 4.80 | 1.54 | 1.00 | .76 | .76 | 55.4 | .20 |
| Surprise | 6.41 | 1.19 | 4.86 | 1.40 | 1.19 | .80 | .80 | 62.2 | .26 |

NOTE: PS = probability of superiority method; CL = common language method.
a. Emotion refers to the emotion displayed in the facial expression. Scale refers to the emotion category rating scale used by judges in the original study. This distinction is kept for those who refer to the original study, where multiple scales are rated on each emotion portrayed in the face.
b. As mentioned in the text, the calculation of PS and CL may result in values < .50 if the first culture entered is the lower scoring culture.

## TWO EXAMPLES FROM PREVIOUSLY PUBLISHED STUDIES

### MATSUMOTO AND EKMAN, 1989

In a study examining cultural differences in judgments of facial expressions of emotion, Matsumoto and Ekman (1989) showed American and Japanese judges facial expressions of emotion posed by Asians and Caucasians. The judges rated each on multiple, 9-point scales of intensity. With the exception of one emotional expression, contempt, Americans had significantly higher mean ratings than the Japanese on the target scales. Table 2 provides the descriptive statistics for the two cultures separately for each expression on the target scale (taken from Yrizarry, Matsumoto, & Wilson Cohn, 1998, who reported a more extensive analysis of the original data). All culture mean differences are statistically significant. (We ignore here poser race and poser gender effects for ease of presentation.)

Computation of the four effect size measures provides interesting information above and beyond that provided by the mean differences. Hedge's *g* indicated that the average American scored 1.47 standard deviation (*SD*) units above the average Japanese when judging anger and .68 *SD* units when judging happiness. Across all emotions except contempt, the mean effect size showed that the average American scored .85 *SD* units above the average Japanese. This indicates that the average American scored higher than about 80% of Japanese persons.

Estimation of the $\Pr(X_a > X_b)$ paints a similar picture. The $z_{cl}$ indicated that the probability that a randomly sampled American would score higher than a randomly sampled Japanese when judging anger was .85; for surprise, it was .80. The same estimates using the PS statistics were .84, and .80, respectively.

The *U1* estimates ranged from a low of 14.7% for disgust to a high of 68.1% for anger. The former indicated that there is only a small nonoverlap, and thus a considerable overlap, between the two distributions for disgust; the latter indicated a considerable nonoverlap for anger. The point biserial correlation suggested that cultural differences were moderate on

**TABLE 3**

**Reanalysis of Kleinknecht, Dinnel, Kleinknecht, Hiruma, and Harada, 1997**

| | *United States* (n = *182)* | | *Japan* (n = *161)* | | | | | |
|---|---|---|---|---|---|---|---|---|
| *Variable* | M | SD | M | SD | *Hedge's* g | *PS* | U1(%) | $r_{pb}$ |
| Independent self-construal | 4.73 | 0.65 | 4.75 | 0.62 | –0.027 | .50 | 0.00 | .00 |
| Interdependent self-construal | 4.63 | 0.67 | 4.43 | 0.72 | 0.29 | .42 | 21.30 | .02 |
| Difference between independent and interdependent self-construals | 0.10 | 0.96 | 0.32 | 0.91 | –0.23 | .45 | 14.70 | .11 |
| Embarassability | 108.80 | 23.99 | 112.27 | 18.69 | –0.16 | .46 | 14.70 | .02 |
| Taijin kyofusho | 83.65 | 28.06 | 93.50 | 30.16 | –0.34 | .41 | 21.30 | .05 |
| Social interaction anxiety | 26.35 | 12.22 | 31.50 | 12.89 | –0.41 | .38 | 27.40 | .08 |
| Social phobia | 18.37 | 10.69 | 20.65 | 13.78 | –0.187 | .47 | 14.70 | .05 |

NOTE: PS = probability of superiority method.

judgments of anger and fear (.35 and .37, respectively), low to moderate on sadness, surprise, and happiness (.20, .26, .10), and negligible on disgust (.01).

The findings on disgust are interesting. Hedge's *g* for disgust was only .23, indicating that the average American scored only .23 *SD* units above the average Japanese. The PS was estimated to be only about .56, which is not very much higher than chance (.50). There was only a 14.7% nonoverlap between the American and Japanese distributions on disgust, and $r_{pb}$ was only .01. This is in contrast to the ANOVA findings indicating that the culture means are "highly" significantly different from each other, $F(1, 5184) = 10.42$, $p < .001$, which would erroneously but often be considered a substantial difference.

## KLEINKNECHT, DINNEL, KLEINKNECHT, HIRUMA, AND HARADA, 1997

In this study, American and Japanese participants completed three measures of social anxiety: the Social Phobia Scale (SPS), the Social Interaction Anxiety Scale (SIAS), and the Taijin Kyofusho Scale (TKS), which was developed to measure social anxiety in Japan. The respondents also completed an Embarrassability Scale (ES) and a Self-Construal Scale, which produces an independent and interdependent self-construal score. The Japanese had significantly higher means than the Americans on the three social anxiety scales and embarrassability, $t(340) = 4.332$, $p < .001$; $t(340) = 2.955$, $p < .01$; $t(340)$ 3.713, $p < .001$; and $t(340) = 1.644$, $p < .10$, for ES, TKS, SIAS, and SPS, respectively. A typical conclusion would be that Japanese experience significantly greater social anxiety and embarrassment than Americans.

The four cultural effect size statistics, however, indicate that these differences may be practically unimportant (see Table 3). Hedge's *g*, for example, suggests that the average Japanese person scores only between 0.16 and 0.41 *SD*s above the average American. The PS values estimate that the probability of a randomly selected Japanese individual having a greater score than a randomly selected American on the TKS is only .59 (i.e., 1 – .41), which is not substantially greater than chance (.50). For ES, which produced the largest significant difference in the means, the PS was only 1 – .46 = .54. The *U1* statistics indicated that there was only a 14.70% nonoverlap in the distributions between Americans and Japanese on ES and SP; for TKS and SIAS, it was 21.30% and 27.40%, respectively. Finally, the point biserial correlations indicated that each of the four social anxiety scales had only a weak relationship with cultural group membership, ranging from .02 to.08.

Further analyses examined whether the observed cultural differences in the social anxiety measures can be accounted for by the self-construals. But although Americans were hypothesized to have higher scores on independent self-construals, this difference was not significant, $t(340) = .265$, *ns*. Each of the four effect size measures for this scale indicated that there was essentially no difference between the two cultures as well (see Table 3). There was a statistically significant difference on interdependent self-construals; it was, however, in the opposite direction predicted, with Americans having significantly higher scores, $t(340) = 2.844$, $p < .005$. Effect size analyses indicated that this difference was also relatively small (Hedge's $g = .29$, PS = .58, $U1$ = 21.30%, and $r_{pb} = .02$).

We also computed the difference between independent and interdependent self-construals in each respondent and tested cultural differences on these differences. The comparison was statistically significant, $t(340) = 2.348$, $p < .05$; it was the Japanese, however, that had relatively higher independent self-construal scores. Once again, the four cultural effect size estimates suggested that this difference was relatively small (Hedge's $g$ = .23, PS = .55, $U1$ = 14.70%, and $r_{pb} = .11$).

## CONCLUSION

Statistics such as *t*s and *F*s have dominated data analysis in cross-cultural research, and other areas of psychology as well. To be sure, they have their place and have led to many important findings. Research using them to document cultural differences is a cornerstone of cross-cultural psychology. Problems occur, however, when we interpret statistical significance to reflect meaningful differences between individuals. As we have argued, *t*s and *F*s cannot tell us about meaningful differences on the level of people, and their sole, continued use in this fashion will only foster stereotypes in research, theory, and practice because group differences are used to infer differences among people.

Fortunately, alternative methods for analyzing data exist, some of which we have discussed here. They provide us with valuable information about the magnitude of cultural differences that are unavailable from *t*s or *F*s. They allow us to make finer estimations of the degree to which observed group differences are represented on the level of individuals. They allow theorists to think more constructively and realistically about conceptual issues, forcing them to go beyond mere global, stereotypic notions that are assumed to be true for all members of a culture. And they provide important guidelines concerning applications of cultural differences.

Just pause to consider the wealth of knowledge concerning cultural differences in any area of cross-cultural comparison that some may assume to be important or large on the level of individuals because previous research has documented statistically significant differences between culture means. How many of these are actually reflective of meaningful differences on the level of individuals? Unfortunately, the answer is unknown, unless tests of group differences in means are accompanied by measures of cultural effect size such as those presented here. If theories, research, and practical work that are supposedly applicable to individuals are based on such limited group difference comparisons, theories, research, and applied programs based on these cultural differences may be based on a house of cards. Future theories in cross-cultural psychology, and all areas of psychology, should be built on a better foundation.

Of course, this foundation starts with better research methodology, and the points we raise are not intended to suggest that statistical methods can compensate for limitations of design in cross-cultural research, particularly when preexisting cultural groups are used as the independent variable (see also Wilkinson et al., 1999). Clearly, one of the biggest challenges facing cross-cultural research concerns the need to replace culture with specific, measurable psychological variables that are hypothesized to account for cultural differences. Incorporation of such context variables is an important improvement in method that transcends methods of data analysis in any cross-cultural study (Bond & Tedeschi, in press; Poortinga, Van de Vijver, Joe, & van de Koppel, 1987; Van de Vijver & Leung, 1997).

Still, data analysis and statistical inference are important parts of methods, and it is these issues that this article addresses. Although we have illustrated four useful measures of cultural effect size when comparing two cultures, it is important to recognize that numerous other indices of effect size are available, as mentioned earlier. For example, Rosnow and Rosenthal (1996) presented a "counternull statistic" to estimate the nonnull value of effect size that is as well supported by the result as is the null hypothesized value of effect size (usually $g = 0$ or $r_{pb} = 0$). Rosnow and Rosenthal also present a measure of effect size that can be used to make focused contrasts of two cultures at a time in a multiple-culture ANOVA design.

The ability to explain, understand, and predict behavior on the individual level is one of the founding pillars of psychology. Although it is important for researchers to examine the influence of many social categories on human behavior—gender, culture, socioeconomic status, and the like—ultimately our goal is to understand individual differences on psychological phenomena and the influence of social structures on those individual differences. Thus, we urge researchers to consider using these and other measures in their future empirical work, and journal editors to require their authors to report such statistics. Only the continued development and refinement of methods in cross-cultural research can help it to further enhance its contributions to psychology throughout the world.

## NOTES

1. This argument is based on the fact that if the null hypothesis is false, Type I error is impossible; the only type of error that could possibly occur would be Type II error. In these cases, the maximum potential error rate for the significance test is thus 97.5% for two-tailed tests and 95% for one-tailed tests. Hunter (1997) cites studies that have computed the error rate for the statistical significance test in leading psychological journals and that conclude that the error rate was about 60% at the time of the study. For more information, see Hunter (1997).

2. A number of other resources discuss the issue of computing and interpreting effect sizes from multisample data, including Rosenthal, Rosnow, and Rubin (2000), McGraw and Wong (1992), and Cortina and Nouri (2000).

3. Authors may be inclined more than readers to interpret their findings as meaningful. Differences in interpretations of meaningfulness based on effect sizes, as we argue here, may be a consequence of the democratization of these evaluations, which generally does not occur with null hypothesis significance tests, because of the accepted standard of statistical significance among researchers.

## REFERENCES

Abelson, R. P. (1997). On the surprising longevity of flogged horses: Why there is a case for the significance test. *Psychological Science*, *8*, 12-15.

Bond, M. H., & Tedeschi, J. T. (in press). Polishing the jade: A modest proposal for improving the study of social psychology across cultures. In D. Matsumoto (Ed.), *Handbook of culture and psychology.* New York: Oxford University Press.

Bresnahan, M. J., Ohashi, R., Liu, W. Y., Nebashi, R., & Liao, C. (1999). A comparison of response styles in Singapore and Taiwan. *Journal of Cross-Cultural Psychology*, *30*, 342-358.

Cohen, J. (1962). The statistical power of abnormal-social psychological research. *Journal of Abnormal and Social Psychology*, *65*, 145-153.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New York: Academic Press.

Cortina, J. M., & Nouri, H. (2000). *Effect sizes for ANOVA designs*. Thousand Oaks, CA: Sage.

Dunlap, W. P. (1999). A program to compute McGraw and Wong's common language effect size indicator. *Behavior Research Methods, Instruments, and Computers*, *31*, 706-709.

Feingold, A. (1992). Sex differences in variability in intellectual abilities: A new look at an old controversy. *Review of Educational Research*, *62*, 61-84.

Feingold, A. (1995). The additive effects of differences in central tendency and variability are important in comparisons between groups. *American Psychologist*, *50*, 5-13.

Feingold, A., & Mazzella, R. (1998). Gender differences in body image are increasing. *Psychological Science*, *9*, 190-195.

Goodwin, R., Nizharadze, G., Luu, L.A.N., Kosa, E., & Emelyanova, T. (1999). Glasnost and the art of conversation: A multilevel analysis of intimate disclosure across three former communist cultures. *Journal of Cross-Cultural Psychology*, *30*, 72-90.

Grissom, R. J. (1994). Probability of the superior outcome of one treatment over another. *Journal of Applied Psychology*, *79*, 314-316.

Harris, R. J. (1997). Significance tests have their place. *Psychological Science*, *8*, 8-11.

Hedges, L. V., & Olkin, L. (1985). *Statistical methods for meta-analysis*. San Diego, CA: Academic Press.

Henderson, B. B., Marx, M. H., & Kim, Y. C. (1999). Academic interests and perceived competence in American, Japanese, and Korean children. *Journal of Cross-Cultural Psychology*, *30*, 32-50.

Hunter, J. E. (1997). Needed: A ban on the significance test. *Psychological Science*, *8*, 3-7.

Hunter, J. E., & Schmidt, F. L. (1990). *Methods and meta-analysis*. Newbury Park, CA: Sage.

Keltikangas-Jarvinen, L., Terav, T., & Pakaslahti, L. (1999). Moral reasoning among Estonian and Finnish adolescents: A comparison of collectivist and individual settings. *Journal of Cross-Cultural Psychology*, *30*, 267-290.

Kleinknecht, R. A., Dinnel, D. L., Kleinknecht, E. E., Hiruma, N., & Harada, N. (1997). Cultural factors in social anxiety: A comparison of social phobia symptoms and *taijin kyofusho*. *Journal of Anxiety Disorders*, *11*, 157-177.

Loftus, G. R. (1996). Psychology will be a much better science when we change the way we analyze data. *Current Directions in Psychological Science*, *5*, 161-170.

Matsumoto, D. (1999). Culture and self: An empirical assessment of Markus and Kitayama's theory of independent and interdependent self-construals. *Asian Journal of Social Psychology*, *2*, 289-310.

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

McGraw, K. O., & Wong, S. P. (1992). A common language effect size statistic. *Psychological Bulletin*, *111*, 361-365.

Ohbuchi, K., Fukushima, O., & Tedeschi, J. T. (1999). Cultural values in conflict management: Goal orientation, goal attainment, and tactical decision. *Journal of Cross-Cultural Psychology*, *30*, 51-71.

Poortinga, Y. H., Van de Vijver, F.J.R., Joe, R. C., & van de Koppel, J.M.H. (1987). Peeling the onion called culture: A synopsis. In C. Kagitcibasi (Ed.), *Growth and progress in cross-cultural psychology* (pp. 22-34). Berwyn, PA: Swets North America.

Rosenthal, R. (1991). *Meta-analytic procedures for social research*. Newbury Park, CA: Sage.

Rosenthal, R., Rosnow, R. L., & Rubin, D. B. (2000). *Contrasts and effect sizes in behavioral research: A correlational approach*. Cambridge, UK: Cambridge University Press.

Rosenthal, R., & Rubin, D. B. (1982). A simple, general purpose display of magnitude of experimental effect. *Journal of Educational Psychology*, *74*, 166-169.

Rosnow, R. L., & Rosenthal, R. (1996). Computing contrasts, effect sizes, and counternulls on other people's published data: General procedures for research consumers. *Psychological Methods*, *1*, 331-340.

Shrout, P. E. (1997). Should significance tests be banned? Introduction to a special section exploring the pros and cons. *Psychological Science*, *8*, 1-2.

Singelis, T. M., Bond, M. H., Sharkey, W. F., & Lai, C.S.Y. (1999). Unpackaging culture's influence on self-esteem and embarassability: The role of self-construals. *Journal of Cross-Cultural Psychology*, *30*, 315-341.

Takano, Y. & Osaka, M. (in press). An unsupported common view: Comparing Japan and the U.S. on individual/collectivism. *Asian Journal of Social Psychology.*

Van de Vijver, F.J.R., & Leung, K. (1997). *Methods and data analysis for cross-cultural research*. Thousand Oaks, CA: Sage.

Wilcox, R. R. (1997). *Introduction to robust estimation and hypothesis testing*. San Diego, CA: Academic Press.

Wilcox, R. R. (1998). How many discoveries have been lost by ignoring modern statistical methods? *American Psychologist*, *53*, 300-314.

Wilkinson, L., The Task Force on Statistical Inference, American Psychological Association Board of Scientific Affairs. (1999). Statistical methods in psychology journals: Guidelines and explanations**.** *American Psychologist*, *54*, 594-604.

Yrizarry, N., Matsumoto, D., & Wilson Cohn, C. (1998). American-Japanese differences in multiscalar intensity ratings of universal facial expressions of emotion. *Motivation and Emotion*, *22*, 315-328.

*David Matsumoto is professor of psychology and director of the Culture and Emotion Research Laboratory at San Francisco State University. He earned his B.A. from the University of Michigan and his M.A. and Ph.D. from the University of California, Berkeley. He has studied emotion, communication, and culture for over 15 years and is the author of more than 250 works on culture and emotion, including original research articles, paper presentations, books, book chapters, videos, and assessment instruments. He is the author of* Culture and Psychology: People Around the World *(Wadsworth) and the editor of the* Handbook of Culture and Psychology *(Oxford University Press).*

*Robert J. Grissom earned his A.B. degree in psychology from Brown University and his Ph.D. in experimental psychology from Princeton University. He is now professor emeritus at San Francisco State University, where he had been coordinator of the graduate program in psychological research and now conducts research on statistical methods. He has recently published several articles in the areas of animal and human memory, therapeutic efficacy, hearing, and social psychology.*

*Dale L. Dinnel is an associate professor of psychology at Western Washington University. He recently coauthored an undergraduate statistics book,* Basic Statistics for the Behavioral Sciences*, with Robert M. Thorndike. In addition, he has been involved in research on cross-cultural variations in social phobia symptoms and cross-cultural variations in conceptualizations of self, including self-worth protection models of achievement motivation.*