ORIGINAL ARTICLE

Facial attractiveness: Ranking of end-of-treatment facial photographs by pairs of Chinese and US orthodontists

Tian-Min Xu,[a] Edward L. Korn,[b] Yan Liu,[c] Hee Soo Oh,[d] Ki Heon Lee,[e] Robert L. Boyd,[f] and Sheldon Baumrind[g]

Beijing, China, Rockville, Md, Gwangju, Korea, and San Francisco, Calif

Introduction: In this study, we assessed agreement and disagreement among pairs of Chinese and US orthodontists in the ranking for "facial attractiveness" of end-of-treatment photographs of growing Chinese and white orthodontic patients. **Methods:** Two groups of orthodontist-judges participated: from the University of the Pacific, School of Dentistry, in California and from Peking University School and Hospital of Stomatology in China. Each judge independently ranked standard clinical sets of profile, frontal, and frontal-smiling photographs of 43 white patients and 48 Chinese patients. Pearson correlations were generated for a total of 1980 rankings by pairs of judges. **Results:** The resulting correlations ranged from +0.004 to +0.96 with a median of +0.54. Of these, 18.7% were lower than 0.4; 41.0% were lower than 0.5; 68.8% were lower than 0.6; 91.6% were lower than 0.7; and only 8.4% were greater than 0.7. As had been anticipated, correlations between judges were higher when they ranked patients of their own ethnicity than when they ranked patients of different ethnicity, but the differences were smaller than had been expected. The rankings of no pair of judges correlated negatively. This is to say that no pair of judges, whether of the same or different ethnicity, ranked the patients so that those 1 judge tended to find attractive were consistently found unattractive by the other. **Conclusions:** The distribution of levels of agreement between pairs of orthodontists did not differ substantially whether the pairs included 2 US orthodontists, 2 Chinese orthodontists, or 1 US and 1 Chinese orthodontist. As might be expected, the pairs of Chinese orthodontists agreed with each other slightly better on average when ranking Chinese patients, and the pairs of US orthodontists agreed with each other slightly better on average when ranking white American patients, but the overall differences were small. These findings appear consistent with the inference that, on average, judgments of "facial attractiveness" by orthodontists at the 2 venues are more similar than had been expected for patients of Chinese and white ethnicity. (Am J Orthod Dentofacial Orthop 2008;134:74-84)

[a]Professor and chair, Department of Orthodontics, Peking University School and Hospital of Stomatology, Beijing, China.
[b]Mathematical statistician, Biometric Research Branch, National Cancer Institute, NIH, Rockville, Md.
[c]Assistant professor, Department of Orthodontics, Peking University School and Hospital of Stomatology, Beijing, China.
[d]Assistant professor, Department of Orthodontics, School of Dentistry, University of the Pacific, San Francisco, Calif.
[e]Associate professor, Department of Orthodontics, Chonnam National University, Gwangju, Korea.
[f]Professor and chair, Department of Orthodontics, School of Dentistry, University of the Pacific, San Francisco, Calif.
[g]Professor, Department of Orthodontics, and director, Craniofacial Research Instrumentation Laboratory, School of Dentistry, University of the Pacific, San Francisco, Calif.
Supported in part by NIH-NIDR grants DE07332 and DE08713 and the American Association of Orthodontists Foundation.
Reprint requests to: Sheldon Baumrind, Craniofacial Research Instrumentation Laboratory, Room 617, Arthur A. Dugoni School of Dentistry, University of the Pacific, 2115 Webster St, San Francisco, CA 94115; e-mail, sbaumrind@pacific.edu.
Submitted, April 2006; revised and accepted, August 2006.
0889-5406/$34.00

doi:10.1016/j.ajodo.2006.08.023

During the past 2 decades, orthodontists have responded to the concerns of their patients by becoming increasingly concerned with facial esthetics. Given this increased focus on appearance, we orthodontists must understand as well as possible precisely how we perceive the "facial attractiveness" of our patients. Past orthodontic research in the area of facial attractiveness has focused on the evaluation of various manufactured models: eg, the study of profile silhouettes,[1-4] computer-modified ("morphed") images of the face,[4-10] and commentaries on the faces of movie stars and beauty contest winners conventionally considered attractive.[11,12] Most studies focused almost exclusively on the profile views of the face. The judges in these studies were drawn from a number of social groups such as artists, parents, and patient peers, as well as from orthodontists.[13-20] Recent publications in the orthodontic literature have also included reports on differences in perceptions of facial attractiveness among various ethnic groups.[4,10,20-24]

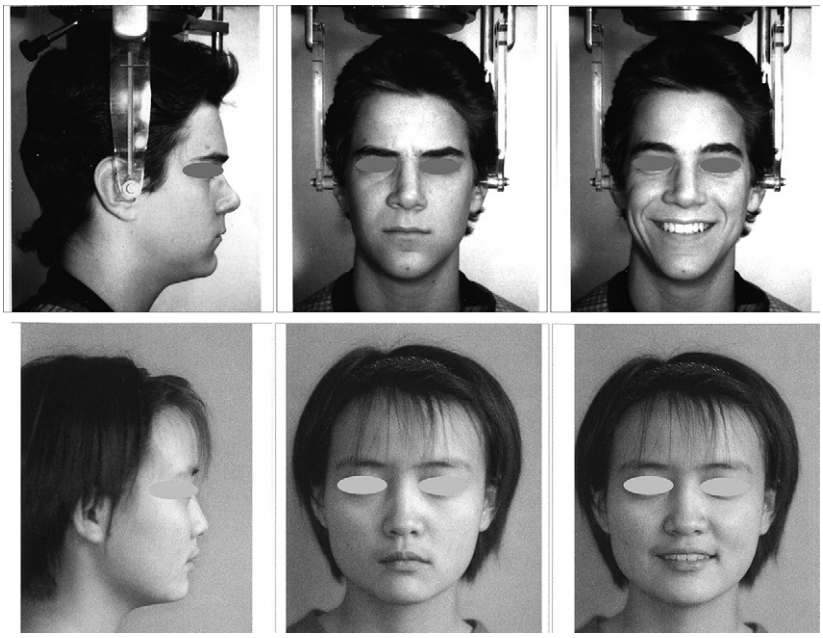

Fig 1. Representative triplets of 1 US and 1 Chinese subject. The images that the judges evaluated were without eye shields. The eyes were shielded for publication.

The design and materials of this study were somewhat closer to clinical experience. Our focus was on the evaluation by orthodontists of sets of semi-standardized photographs of the kind typically used as part of the standard clinical orthodontic record. The set of images for each patient consisted of profile, frontal, and frontal-smiling views such as those shown in Figure 1. For convenience, we refer to the set of 3 images of each patient as a "triplet."

Sets of photographs of this kind have been used for many years for the global characterization of patients' appearance, as well as for the specific evaluation of the proportions of the nose, lips, and chin.[25-35] This study focuses on the agreement and disagreement between pairs of orthodontists ranking the "facial attractiveness" of treated orthodontic patients from triplets of end-of treatment photographs. In addition, we began to investigate how orthodontists from different cultural backgrounds compare with each other in their evaluations of patients from other ethnic backgrounds. In this article, we were more concerned with the agreement and disagreement among orthodontists than with the rankings of the individual triplets. In a subsequent article, we will report on the rankings of the triplets and how they correlate with representative cephalometric measurements generally considered to reflect "facial attractiveness."

The specific issues we addressed were (1) to what extent do pairs of orthodontists of the same ethnicity and cultural background rank photographs of treated patients of their own ethnicity in the same way? (2) to what extent do pairs of orthodontists of the same ethnicity and cultural background rank photographs of treated patients of different ethnicity in the same way? (3) to what extent do pairs of orthodontists of different ethnicity and cultural background rank photographs of treated patients in the same way? and(4) how do the rankings of experienced orthodontists compare with those of residents?

MATERIAL AND METHODS

In this study, 2 cohorts of orthodontists ranked facial photographs of 2 cohorts of patients. One cohort of orthodontists was educated and practices in the United States. The other cohort of orthodontists was educated and practices in China. All US judges had received their primary orthodontic education in accredited university graduate programs in the United States. All the Chinese judges had received their primary orthodontic education in Chinese orthodontic residencies. One cohort of patients was from the United States, and the other was from China. Each judge evaluated the photographs from both cohorts of patients.

The judges were chosen from among the academic and clinical faculties and residents of the Department of Orthodontics at Peking University School of Stomatology, and the Department of Orthodontics at the University of the Pacific School of Dentistry. Demographic

Table I. Judge demographics

	n	*Sex ratio (M:F)*	*Age (y)*	*Specialty experience (y)*
Chinese judges				
Judges				
Senior faculty	5	2:3	36.8 ± 1.5	13.0 ± 1.6
Junior faculty	5	2:3	31.4 ± 1.7	8.0 ± 1.6
Total faculty	10	4:6	34.1 ± 3.2	10.5 ± 3.0
Third-year residents	5	1:4	28.0 ± 2.0	3.2 ± 1.0
Second-year residents	5	3:2	27.6 ± 3.3	2.7 ± 2.7
First-year residents	5	1:4	24.6 ± 2.6	0.5 ± 0.0
Total residents	15	5:10	26.7 ± 2.9	2.1 ± 1.9
Total Chinese judges	25	9:16	29.7 ± 4.7	5.5 ± 4.8
US judges				
Senior faculty	5	5:0	66.0 ± 5.6	37.2 ± 9.3
Junior faculty	5	2:3	42.4 ± 4.8	10.6 ± 7.7
Total faculty	10	7:3	54.2 ± 13.4	23.9 ± 16.2
Second-year residents	5	3:2	32.2 ± 1.9	1.2
First-year residents	5	3:2	27.4 ± 2.3	0.2
Total residents	10	6:4	29.8 ± 3.2	0.7 ± 0.5
Total US judges	20	13:7	42.0 ± 15.7	12.3 ± 16.3

Values for some residents count experience before residency.
M, Male; *F,* female.

information for the 25 Chinese judges and the 20 US judges is included in Table I. Faculty judges are classified according to their academic role; residents are classified by years of study.

The same strategy was used for sampling facial photographs in both the United States and China. The US patients were sampled from end-of-treatment records from the private practice of Dr Helmer Pearson, director of the Graduate Orthodontic Clinic at the New Jersey Dental School Department of Orthodontics. The Chinese patients were sampled at the Graduate Orthodontic Clinic, Peking University, School of Stomatology. At each venue, the first step was to randomly order the charts of all patients during a specified 3-year interval. Then, after the identification of a sufficiently large sample with full records, a stratified subsample consisting of 48 patient records was created, divided into 4 groups of 12 records each. Each group contained triplets for 3 Class I nonextraction, 3 Class I extraction, 3 Class II nonextraction, and 3 Class II extraction patients. In 1 group of US patients (group C), it was necessary to drop 3 subjects before data acquisition because the quality of the photographic images was unsatisfactory. The ratio of female to male subjects in each sample was 3:1, approximating the ratio of the sexes in the patient populations in both Chinese and US practices. Only the end-of-treatment triplet photographs of each patient were used in this study. See Figure 2 for further details of the sampling process.

Standardized profile, frontal, and frontal-smiling photographs taken at the end of treatment were acquired from each patient's treatment file. The original US images were 4 × 5-in grayscale photographic prints; the Chinese images were 2 × 2-in color slides. The Chinese images were scanned at 1200 dpi in Beijing by using an AGFA T1200 scanner (AGFA-Gevaert Corp, Mortsel, Belgium). The US images were scanned at 300 dpi at the Department of Orthodontics, University of Medicine and Dentistry of New Jersey. The scans of both sets of images were printed at a common final scale at the University of the Pacific on an Epson Stylus Color 800 inkjet printer (Epson America, Long Beach, Calif). The profile, frontal, and frontal-smiling images for each patient were printed side by side as grayscale triplets on 8.5 × 11-in high-quality paper (Fig 1).

The 8 groups of triplets were viewed under constant conditions for all judges at both universities. To blind the judges, the triplets in each group were rerandomized before presentation on a laboratory table as shown in Figure 3. The order of the presentation of the triplets in each group was the same for all judges.

The instructions to each judge were the same at both institutions. Although all the Chinese judges read English, their instructions were translated into Chinese. The images were examined by 1 judge at a time. For each group of patients, the judge picked up the triplet she or he considered "most attractive," then the triplet considered next most attractive, and so on until all 12 images had been removed from the table. An observer recorded the order of ranks, assigning a value of 12 to the triplet considered "most attractive," descending to a value of 1 for the triplet

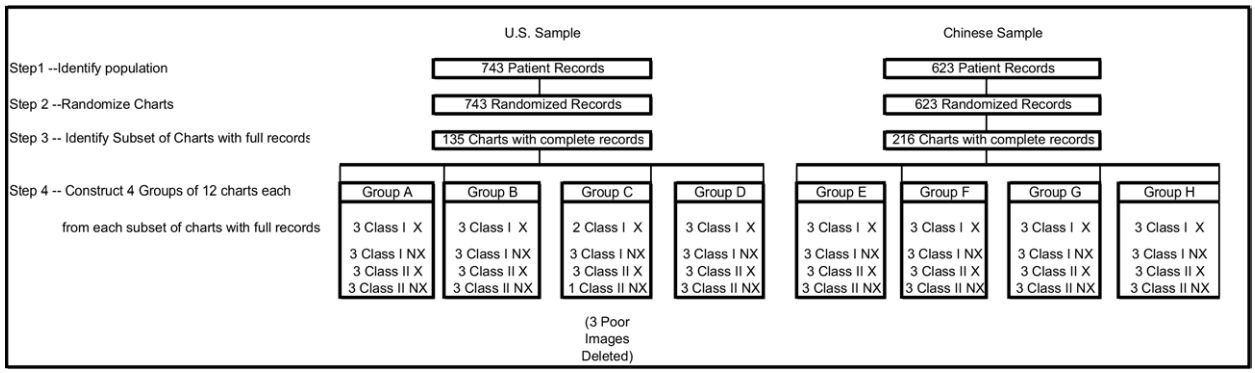

Fig 2. Schematic showing the procedures for drawing the equivalent Chinese and US samples. Step 1: at each venue, patients who received treatment during a specified time period were identified. Step 2: each patient was assigned a random number. All subsequent procedures were conducted with the charts sorted in random order. This ensured that the sample was representative of the population (ie, practice) from which it was drawn. Step 3: proceeding in random order, all charts with complete records were identified and duplicated for further studies. (For the purposes of the general project of which this study is a part, a "complete record" was considered to be one in which a lateral cephalogram, study casts, full-mouth intraoral or panoramic x-rays, and a facial photographic triplet were available at the beginning and end of full-bonded orthodontic treatment.) Step 4: parallel stratified subsets of 48 Chinese subjects and 45 US subjects with complete records were selected for the study. The subset from each institution comprised the first 48 randomly ordered charts from that venue that satisfied the Angle Class and extraction or nonextraction criteria for this study, except that 3 subjects were lost from US group C because their images were technically unsatisfactory.

Fig 3. A judge examining a group of images. Each judge examined all triplets, 1 group of images at a time, ordering patients by attractiveness with the most attractive triplet first and the least attractive last.

considered "least attractive." (For group C, the 9 triplets were ranked from 9 to 1.) Each judge ranked all 4 groups of Chinese patients at 1 session and all 4 groups of US patients at another session.

The values for the remaining patients in group C, originally ranked 9 to 1, were each multiplied by 1.3 to yield transformed values from 11.7 to 1.3. Later, after data for all remaining patients had been gathered, it was considered appropriate to delete the data for 2 additional US patients (each from a different group that originally had 12 patients) because they were the only nonwhites in the US sample. For each of these 2 groups, the rankings of the 11 remaining judges were first computationally modified to run from 11 to 1 and then multiplied by 1.08333 to yield transformed values from 11.92 to 1.08. These adjustments were made so that the rankings of each group would have the same mean (6.5) and approximately the same standard deviation, thus facilitating the merging of the data from all 4 US groups. These operations resulted in a final US sample of 43 patients.

Statistical analysis

Pearson correlations for all judge pairs were computed by using the SAS statistical package (version 9.1, SAS, Cary, NC). Because the original data were ordered by rank, the calculated values for the Chinese patients were identical to those that would have been computed by using the Spearman method. Among the US patients, the Pearson values differed slightly from their corresponding Spearman values because of the

transformations in the groups with fewer than 12 subjects. However, these differences were not consequential. Furthermore, it would have been inappropriate to treat all the pairwise correlations as if they were independent because many of the correlations involved the same judge. Therefore, for purposes of computing the statistical significance of differences between different groupings of judges, standard errors of the means of correlation coefficients were computed by jackknifing the judges.[36] *P* values comparing means of correlation coefficients were calculated with a z-test using these jackknifed standard errors. These standard errors and *P* values account for the variability of the judges but not for the variability of the cases. Therefore, the *P* values should be interpreted as reflecting how different the true mean correlations for the given set of subjects are. All *P* values are 2-sided.

RESULTS

Overall, there were 990 separate pairings between Chinese and US judges in the examination of the Chinese patients and an additional 990 separate pairings in the examination of the US patients. Among the most interesting findings was that the rankings of no pair of judges correlated negatively. This is to say that there was no situation in which 2 judges, whether of the same or different ethnicity, had concepts of attractiveness so that the faces that 1 judge tended to find attractive were consistently found unattractive by the other. It was also true that the level of agreement among pairs of judges was highly variable, ranging from Pearson r scores as low as +0.004 to as high as +0.96. Among the total of 1980 judge-pair correlations (r), 41% were lower than +0.5; 68.8% were lower than +0.6.; 91.6% were lower than +0.7; and only 8.4% were greater than +0.7; and 18.5% of all judge-pair correlations failed to reach statistical significance at the 0.05 level. (Statistical significance at the 0.05 level required an r value greater than 0.288 for the 48 Chinese patients and greater than 0.304 for the 43 US patients.)

In Figure 4 and Table II, we present a more detailed examination of the findings for all patients by all pairs of judges.

Figure 4 and Table II compare the correlations in the rankings of all judge-pairs for all Chinese and US patients. In ranking the Chinese patients, agreement of the Chinese judges with each other was statistically significantly higher than agreement of the US judges with each other (A > C, $P = 0.02$). The mean difference in correlation (r) between Chinese judge-pairs and US judge-pairs was 0.11. Half the Chinese judge-pairs (A) had correlations greater than 0.62, whereas only a quarter of the US judge-pairs (C) were greater than 0.60.

Conversely, in the ranking of the US patients, agreement among the US judge-pairs tended to be higher than agreement among the Chinese judge-pairs, as might have been expected. However, this difference did not reach statistical significance (F > D, $P = 0.096$).

We had also expected that the correlations for pairings between judges of different ethnicity would be lower than those in which both judges were of the same ethnicity. The findings supported this expectation. For the ranking of the Chinese patients, pairings of mixed judges had lower correlations than did pairings of Chinese judges (A > B, $P < 0.001$), whereas, for the ranking of the US patients, pairings of mixed judges had lower correlations than US judges (F > E, $P = 0.005$). The data further indicated that pairings of judges of different ethnicity correlated with each other about as well as parings in which 2 judges of the same ethnicity evaluated patients of the other ethnicity (B ~ C and D ~ E).

We were also interested in examining patterns of agreement between faculty and residents at each institution. The tables and figures that follow are analogous to those reported for the full sample in Table II and Figure 4, but the numbers of judges are considerably smaller, and the investigations reported in these tables and figures should be considered exploratory rather than definitive.

Figure 5 and Table III compare the correlations in the rankings of the Chinese faculty with each other and with those of the Chinese residents. When Chinese orthodontists evaluated Chinese patients, the residents appeared to agree with each other (C) more strongly than they agreed with the faculty (B), or than the faculty agreed with each other (A). Only a quarter of the faculty appeared to agree with each other more strongly than a correlation of 0.63, whereas half the residents appeared to agree with each other more strongly than 0.65. When Chinese orthodontists evaluated US patients, there was no evidence of consequential differences between faculty pairs (D), resident pairs (F), and faculty-resident pairs (E). None of these relationships was strong enough to be statistically significant, and these findings serve only as preliminary estimates.

Figure 6 and Table IV compare the correlations in the rankings of the US faculty with each other and with those of the US residents. The sample sizes in these comparisons are smaller than those in the preceding

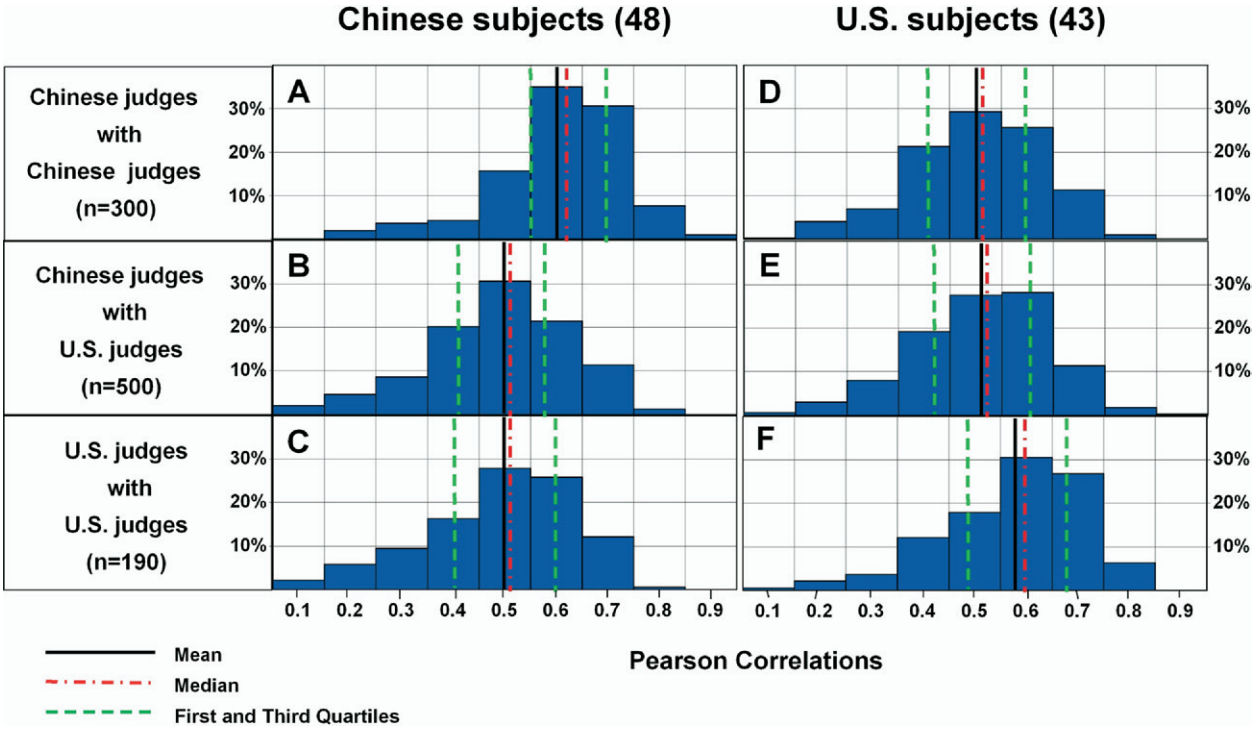

Fig 4. Graphic representation of the statistical values of Table II.

Table II. Descriptive statistics for the pairings of all judges for all subjects

Judge pairings	*Chinese patients (48)*	*US patients (43)*
Chinese judges with Chinese judges (n = 300)	(A) Mean* = 0.60 ± 0.13 SE† = 0.033 Median‡ = 0.62 q1, q3§ = 0.55, 0.69	(D) Mean = 0.50 ± 0.13 SE = 0.024 Median = 0.51 q1, q3 = 0.42, 0.60
Chinese judges with US judges (n = 500)	(B) Mean = 0.49 ± 0.14 SE = 0.026 Median = 0.50 q1, q3 = 0.41, 0.57	(E) Mean = 0.51 ± 0.13 SE = 0.021 Median = 0.52 q1, q3 = 0.43, 0.61
US judges with US judges (n = 190)	(C) Mean = 0.49 ± 0.14 SE = 0.037 Median = 0.50 q1, q3 = 0.40, 0.60	(F) Mean = 0.57 ± 0.14 SE = 0.032 Median = 0.59 q1, q3 = 0.48, 0.67

*Mean ± standard deviation.
†Standard error. Standard errors in this table have been adjusted for the fact that the correlations of each judge are all other judges are not independent.
‡Half of all judge-pairs had correlations lower than this value; the other half had correlations higher than this value.
§*q1,* First quartile: one quarter of the judge-pairs had correlations lower than this value; *q3,* third quartile: three quarters of the judge-pairs had correlations lower than this value.

table and figure and are sufficient only for preliminary impressions. There are some indications that US residents might differ with US faculty more when they evaluate white patients (D, E, and F) than they do when they rank Chinese patients (A, B, and C). However, these indications are small and reach statistical significance barely if at all. Hence, as was true of the comparison between Chinese faculty and residents, the findings of Table IV and Figure 6 can only be considered preliminary estimates.

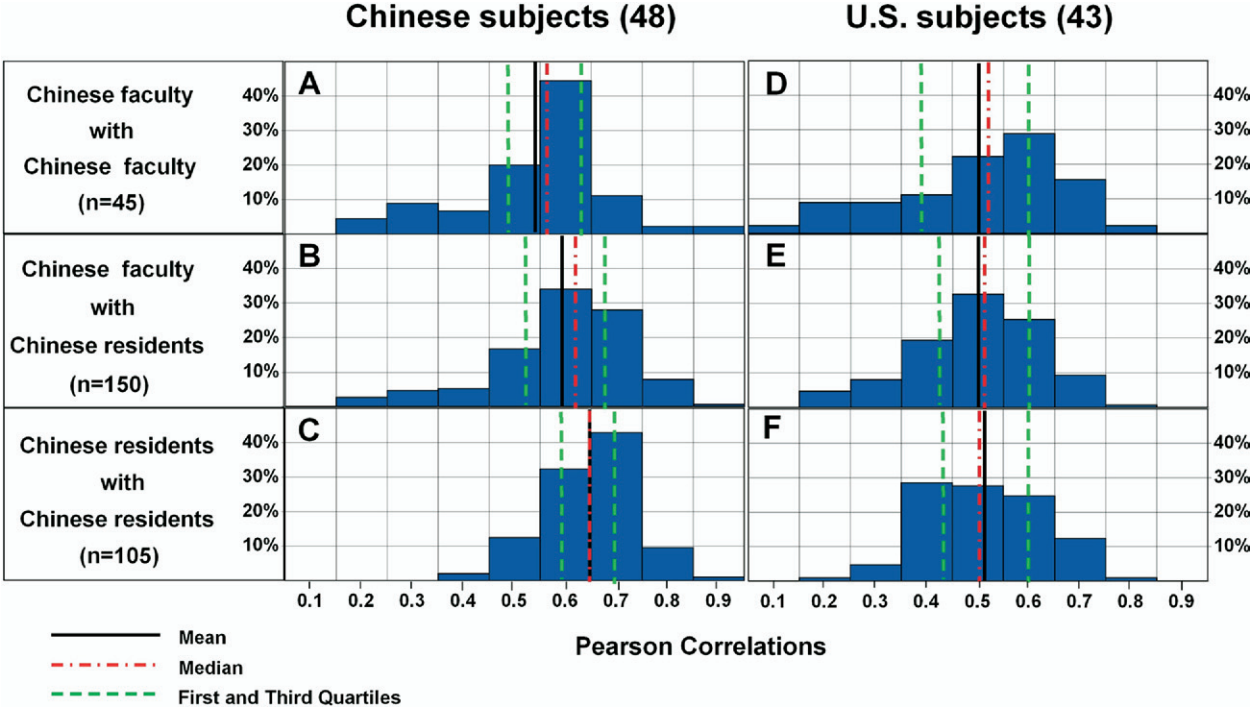

Fig 5. Graphic representation of the statistical values of Table III.

Table III. Descriptive statistics for the Chinese judges for all subjects

Judge pairings	*Chinese patients (48)*	*US patients (43)*
Chinese faculty with Chinese faculty (n = 45)	(A) Mean* = 0.54 ± 0.15 Median† = 0.56 q1, q3‡ = 0.48, 0.63	(D) Mean = 0.50 ± 0.16 Median = 0.52 q1, q3 = 0.38, 0.60
Chinese faculty with Chinese residents (n = 150)	(B) Mean = 0.59 ± 0.14 Median = 0.62 q1, q3 = 0.53, 0.68	(E) Mean = 0.50 ± 0.13 Median = 0.51 q1, q3 = 0.42, 0.60
Chinese residents with Chinese residents (n = 105)	(C) Mean = 0.65 ± 0.09 Median = 0.65 q1, q3 = 0.59, 0.69	(F) Mean = 0.51 ± 0.12 Median = 0.50 q1, q3 = 0.43, 0.60

*Mean ± standard deviation.
†Half of all judge-pairs had correlations lower than this value; the other half had correlations higher than this value.
‡*q1,* First quartile: one quarter of the judge-pairs had correlations lower than this value; *q3,* third quartile: three quarters of the judge-pairs had correlations lower than this value.

DISCUSSION

From a clinical perspective, the most salient observation from this study is the similarity in mean level of agreement when pairs of clinicians of different ethnicity ranked the "facial attractiveness" of treated patients of different ethnicities. In terms of the full range of variability that clearly exists in human facial attractiveness, the variability in each of our 2 samples was relatively small. Within each sample, there was at the end of orthodontic treatment considerable similarity in age, dress, and photographic pose, with only small residual deviations from normal occlusion. Yet orthodontists who were products of different cultures and educated under different conditions agreed reasonably strongly on average about the relative ranking of different patients. This finding implies a high degree of similarity in standards of judgment among orthodontic clinicians at these 2 venues. It also implies that the

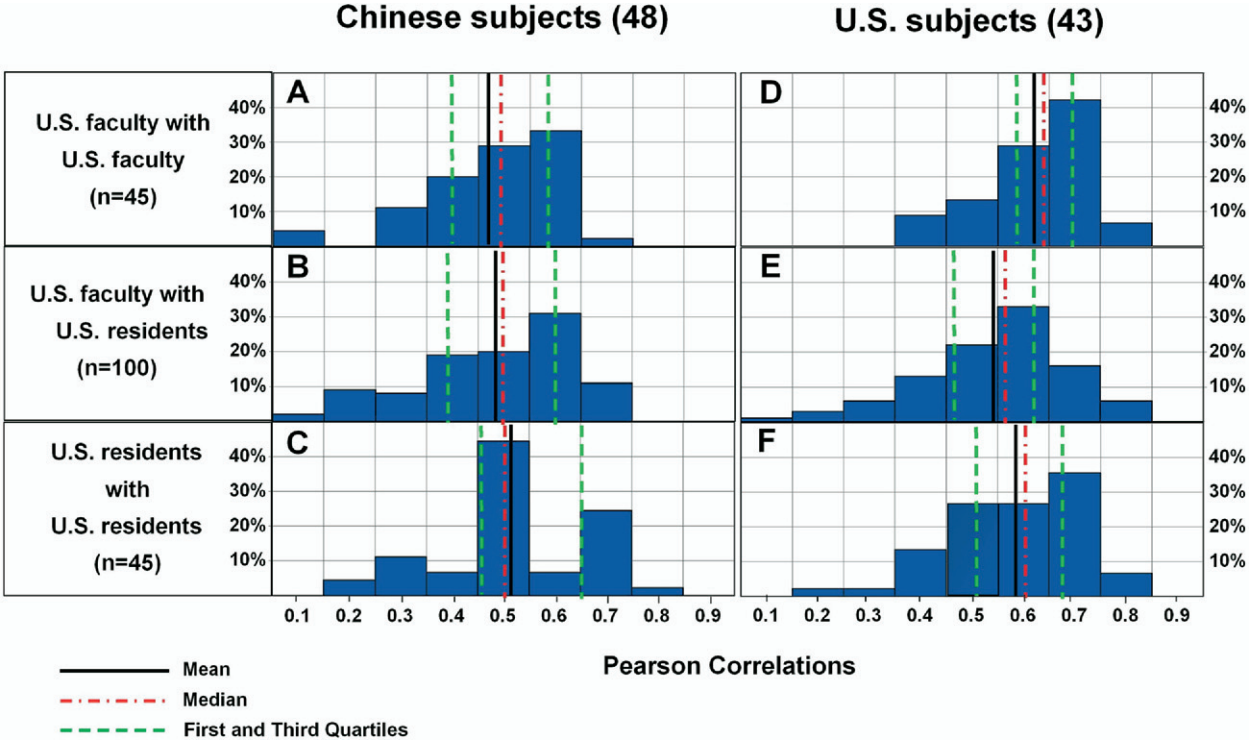

Fig 6. Graphic representation of the statistical values of Table IV.

Table IV. Descriptive statistics for the US judges for all subjects

Judge pairings	*Chinese patients (48)*	*US patients (43)*
US faculty with US faculty (n = 45)	(A) Mean* = 0.48 ± 0.13 Median† = 0.49 q1, q3‡ = 0.40, 0.58	(D) Mean = 0.62 ± 0.11 Median = 0.64 q1, q3 = 0.58, 0.69
US faculty with US residents (n = 100)	(B) Mean = 0.48 ± 0.15 Median = 0.49 q1, q3 = 0.39, 0.60	(E) Mean = 0.54 ± 0.14 Median = 0.56 q1, q3 = 0.46, 0.62
US residents with US residents (n = 45)	(C) Mean = 0.51 ± 0.14 Median = 0.50 q1, q3 = 0.46, 0.65	(F) Mean = 0.58 ± 0.13 Median = 0.60 q1, q3 = 0.51, 0.67

*Mean ± standard deviation.
†Half of all judge-pairs had correlations lower than this value; the other half had correlations higher than this value.
‡*q1,* First quartile: one quarter of the judge-pairs had correlations lower than this value; *q3,* third quartile: three quarters of the judge-pairs had correlations lower than this value.

sources of the differences in judgment that do exist among them will be difficult to identify and characterize.

This similarity was observed despite considerable differences in the demographic composition of the orthodontic faculties at Beijing and San Francisco (Table I). In general, the Chinese faculty was more homogeneous with respect to age and years of experience than its US counterpart, and it contained a higher proportion of women. On average, the US judges were older and had more years of clinical experience. If there had been large differences between the way the Chinese and US judges ranked the patients, these demographic dissimilarities between the 2 groups of judges might have been thought to be responsible for the differences. But the mean differences actually observed between the rankings of the Chinese and US judges were small, thus reinforcing our impression that the similarities in judgment between the 2 groups in the ranking of "facial

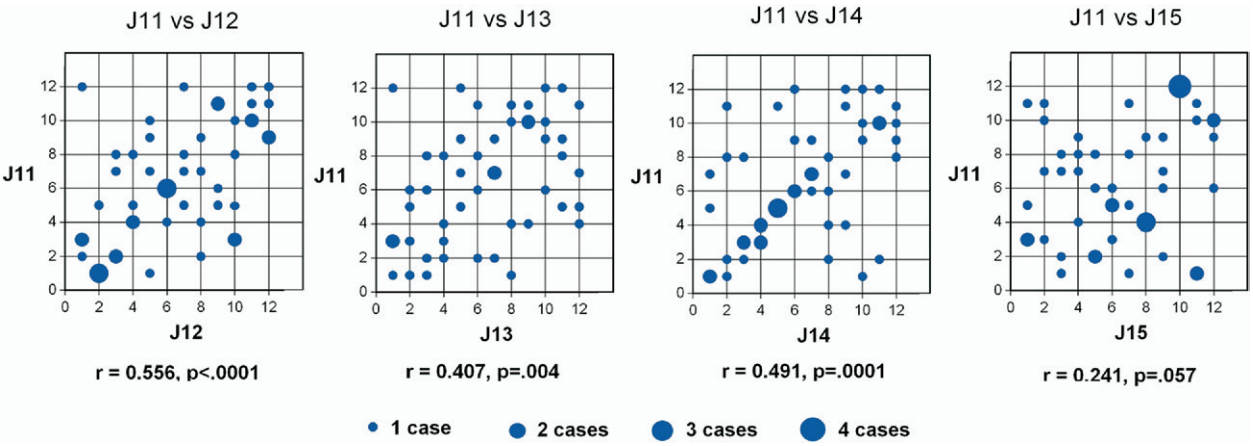

Fig 7. Representative scatterplots and correlations: scatterplots of the 48 Chinese subjects showing the correlations (r) and *P* values. The rankings from 1 to 12 by 1 judge (J11) were plotted separately against those of 4 other judges (J12-J15). In each plot, the rankings of judge J11 were plotted on the y-axis, and those of 1 of the other judges were plotted on the x-axis with a point representing a patient at the intersection. If more than 1 patient received the same combination of scores from both judges, the point is enlarged proportionally. The higher the correlation, the more closely grouped are the values of the individual patients and the lower is the probability that the results can be accounted for merely by chance. Note the differences in the distributions for different pairs of judges. Note also that, even with high statistical significance, the values for individual patients in the same plot can be widely scattered.

attractiveness" are robust across differences in age and sex.

We noted further that, although agreement between pairs of judges at the 2 venues tended to be strongly significant statistically, the actual values of the Pearson correlations reported in this study are distinctly lower than those in many other studies of facial attractiveness. In some studies, correlations on the order of 0.8 and 0.9 have been reported.[13,15,37-39] We believe that the difference between those findings and ours is a property of the fact that somewhat different questions were asked. In most previous orthodontic studies of facial attractiveness, the responses of a group of judges were pooled in an attempt to discern overall judge preferences for particular subjects.[4-10,13-20,40] In this study, we focused on the smallest unit of comparison between judges—that of 1 judge with 1 other judge. (We did so because the comparison between 2 judges seems to model well the clinical situation in which 2 clinicians exchange views concerning a patient of common interest.) However, averaging the correlations of a large number of judges leads to higher absolute r value. Had we pooled the rankings of all judges in this study to obtain the best overall estimate of the ranking of each patient, the mean correlation of all 45 judges for all 91 patients is r = 0.88. For all 45 judges for the 48 Chinese patients, the corresponding value is r = 0.86, and, for all 45 judges for the 43 US patients, the value is r = 0.91.

Readers also need to be aware to that the larger any sample is, the lower will be the absolute value of r needed to achieve statistical significance. This means that, for samples as large as 43 or 48, statistical significance can be achieved even with considerable interjudge differences in individual cases. For a visual sense of what the paired judgments we used look like in terms of interjudge agreement for individual patients, see Figure 7.

Our findings concerning the questions proposed in the introduction can be summarized as follows. Among 300 Chinese judge-pairs ranking Chinese patients, the average correlation was r = 0.60 ± 0.13. When the same 300 pairs of Chinese judges ranked US white patients, the average correlation was r = 0.50 ± 0.13. Among 190 US judge-pairs ranking US white patients, the average correlation was r = 0.57 ± 0.14. When the same 190 pairs of US judges ranked Chinese patients, the average correlation was r = 0.49 ± 0.14. Among 500 judge-pairs of mixed ethnicity (1 Chinese judge with 1 US judge), the average correlation ranking Chinese patients was r = 0.49 ± 0.14, and the average correlation ranking US white patients was r = 0.51 ± 0.13.

When Chinese patients were evaluated, the rankings of pairs of Chinese orthodontists correlated with each other significantly better than did those of pairs of US orthodontists (P = 0.020). When white patients

were evaluated, the rankings of pairs of US orthodontists appeared to correlate with each other more strongly than those of pairs of Chinese orthodontists, but the strength of the relationship fell short of statistical significance ($P = 0.096$).

The range of correlations for all types of judge-pairs was large. Within each type of comparison, values for individual judge-pairs ranged from $r = 0.2$ to $r = 0.8$.

Separate comparisons between the responses of faculty and residents at each venue involved reduced sample sizes and led to equivocal results. Findings for these comparisons have been tabulated and presented but should be viewed only as preliminary estimates.

CONCLUSIONS

There were real differences between the mean rankings of pairs of Chinese orthodontists and the mean rankings of pairs of US orthodontists, but these differences were small. Within any judge pair (Chinese with Chinese, US with US, or US with Chinese), disagreement between the 2 orthodontists could be quite considerable, but no pair of judges correlated negatively. This implies that in no comparison did 2 judges, whether of the same or different ethnicity, have concepts of facial attractiveness such that the patients whose faces 1 judge tended to find more attractive were consistently found less attractive by another judge.

These findings are consistent with the impression that, on average, the judgments of "facial attractiveness" by orthodontists at these 2 venues are much less different than had been expected for patients of either Chinese or white ethnicity.

REFERENCES

1. Barrer JG, Ghafari J. Silhouette profiles in the assessment of facial esthetics: a comparison of cases treated with various orthodontic appliances. Am J Orthod 1985;87:385-91.
2. Czarnecki ST, Nanda RS, Currier GF. Perceptions of a balanced facial profile. Am J Orthod Dentofacial Orthop 1993;104:180-7.
3. Cox NH, Van der Linden FP. Facial harmony. Am J Orthod 1971;60:175-83.
4. Lew KK, Soh G, Loh E. Ranking of facial profiles among Asians. J Esthet Dent 1992;4:128-30.
5. Hier LA, Evans CA, BeGole EA. Comparison of preferences in lip position using computer animated imaging. Angle Orthod 1999;69:231-8.
6. Giddon DB, Sconzo R, Kinchen JA, Evans CA. Quantitative comparison of computerized discrete and animated profile preferences. Angle Orthod 1996;66:441-8.
7. Maganzini AL, Tseng JY, Epstein JZ. Perception of facial esthetics by native Chinese participants by using manipulated digital imagery techniques. Angle Orthod 2000;70:393-9.
8. Kitay D, BeGole EA, Evans CA. Computer-animated comparison of self-perception with actual profiles of orthodontic and nonorthodontic subjects. Int J Adult Orthod Orthognath Surg 1999;14:125-34.
9. Anderson NK, Evans CA, Giddon DB. Comparison of perceptions of computer- animated left- and right-facing profiles. J Prosthodont 1999;8:72-9.
10. Soh J, Chew MT, Wong HB. A comparative assessment of the perception of Chinese facial profile esthetics Am J Orthod Dentofacial Orthop 2005;127:692-9.
11. Riedel RA. An analysis of dentofacial relationships. Am J Orthod 1957;43:103-19.
12. Peck H, Peck S. A concept of facial esthetics. Angle Orthod 1970;40:284-317.
13. Bell R, Kiyak HA, Joondeph DR, McNeill RW, Wallen TR. Perceptions of facial profile and their influence on the decision to undergo orthognathic surgery. Am J Orthod 1985;88:323-32.
14. Prahl-Andersen B, Boersma H, van der Linden FPGM, Moore AW. Perceptions of dentofacial morphology by laypersons, general dentists and orthodontists. J Am Dent Assoc 1979;98: 209-12.
15. Peerlings RH, Kuijpers-Jagtman AM, Hoeksma JB. A photographic scale to measure facial esthetics. Eur J Orthod 1995;17: 101-9.
16. Kerr WJS, O'Donnell JM. Panel perception of facial attractiveness. Br J Orthod 1990;17:299-304.
17. Foster EJ. Profile preference among diversified groups. Angle Orthod 1973;43:34-40.
18. Cochrane SM, Cunningham SJ, Hunt NP. Perceptions of facial appearance by orthodontists and the general public. J Clin Orthod 1997;31:164-8.
19. Cochrane SM, Cunningham SJ, Hunt NP. A comparison of the perception of facial profile by the general public and 3 groups of clinicians. Int J Adult Orthod Orthognath Surg 1999;14:291-5.
20. Sushner NI. A photographic study of the soft-tissue profile of the Negro population. Am J Orthod 1977;72:373-85.
21. Hwang HS, Kim WS, McNamara JA Jr. Ethnic differences in the soft tissue profile of Korean and European-American adults with normal occlusions and well-balanced faces. Angle Orthod 2002; 72:72-80.
22. Mantzikos T. Esthetic soft tissue profile preferences among the Japanese population. Am J Orthod Dentofacial Orthop 1998;114: 1-7.
23. Türkkahraman H, Gökalp H. Facial profile preferences among various layers of Turkish population. Angle Orthod 2004;74: 640-7.
24. Mejia-Maidl M, Evans CA, Viana G, Anderson NK, Giddon DB. Preferences for facial profiles between Mexican Americans and Caucasians. Angle Orthod 2005;75:953-8.
25. Holdaway RA. A soft tissue cephalometric analysis and its use in orthodontic treatment planning. Part I. Am J Orthod 1983;84:1-28.
26. Ricketts RM. Planning treatment on the basis of facial pattern and an estimate of its growth. Angle Orthod 1957;27:14-37.
27. Burstone CJ. Integumental contour and extension patterns. Angle Orthod 1959;23:146-57.
28. Hsu BS. Comparisons of the five analytic reference lines of the horizontal lip position: their consistency and sensivity. Am J Orthod Dentofacial Orthop 1993;104:355-60.
29. Ackerman JL, Proffit WR, Sarver DM. The emerging soft tissue paradigm in orthodontic diagnosis and treatment planning. Clin Orthod Res 1999;2:49-52.
30. Bergman RT. Cephalometric soft tissue facial analysis. Am J Orthod Dentofacial Orthop 1999;116:373-89.
31. Arnett GW, Jelic JS, Kim J, Cummings DR, Beress A, Worley CM, et al. Soft tissue cephalometric analysis (diagnosis and

treatment planning of dentofacial deformity). Am J Orthod Dentofacial Orthop 1999;116:239-53.
32. Arnett GW, Bergman RT. Facial keys to orthodontic diagnosis and treatment planning (part II). Am J Orthod Dentofacial Orthop 1993;103:395-411.
33. Arnett GW, Bergman RT. Facial keys to orthodontic diagnosis and treatment planning (part I). Am J Orthod Dentofacial Orthop 1993;103:299-312.
34. Holdaway RA. A soft-tissue cephalometric analysis and its use in orthodontic treatment planning (part I). Am J Orthod 1983;84:1-28.
35. Merrifield LL. The profile line as an aid in critically evaluating facial esthetics. Am J Orthod 1966;52:804-22.
36. Korn EL, Graubard BI. Analysis of health surveys. New York: Wiley; 1999. p 29.
37. Tatarunaite E, Playle R, Hood K, Shaw W, Richmond S. Facial attractiveness: a longitudinal study. Am J Orthod Dentofacial Orthop 2005;127:676-82.
38. Langlois JH, Kalakanis LE, Rubenstein AJ, Larson AD, Hallamam MJ, Smoot MT. Maxims or myths of beauty: a meta-analytic and theoretical review. Psychol Bull 2000;126:390-423.
39. Feingold A. Good-looking people are not what we think. Psychol Bull 1992;111:304-41.
40. Phillips C, Tulloch C, Dann C. Rating of facial attractiveness. Community Dent Oral Epidemiol 1992;20:214-20.