

ORIGINAL ARTICLE

Reliability assessment and correlation analysis of evaluating orthodontic treatment outcome in Chinese patients

Guang-Ying Song¹, Zhi-He Zhao², Yin Ding³, Yu-Xing Bai⁴, Lin Wang⁵, Hong He^{6,7}, Gang Shen⁸, Wei-Ran Li¹, Sheldon Baumrind⁹, Zhi Geng¹⁰ and Tian-Min Xu¹

This study aimed to assess the reliability of experienced Chinese orthodontists in evaluating treatment outcome and to determine the correlations between three diagnostic information sources. Sixty-nine experienced Chinese orthodontic specialists each evaluated the outcome of orthodontic treatment of 108 Chinese patients. Three different information sources: study casts (SC), lateral cephalometric X-ray images (LX) and facial photographs (PH) were generated at the end of treatment for 108 patients selected randomly from six orthodontic treatment centers throughout China. Six different assessments of treatment outcome were made by each orthodontist using data from the three information sources separately and in combination. Each assessment included both ranking and grading for each patient. The rankings of each of the 69 judges for the 108 patients were correlated with the rankings of each of the other judges yielding 13 873 Spearman r_s values, ranging from -0.08 to $+0.85$. Of these, 90% were greater than 0.4, showing moderate-to-high consistency among the 69 orthodontists. In the combined evaluations, study casts were the most significant predictive component ($R^2=0.86$, $P<0.0001$), while the inclusion of lateral cephalometric films and facial photographs also contributed to a more comprehensive assessment ($R^2=0.96$, $P<0.0001$). Grading scores for SC+LX and SC+PH were highly significantly correlated with those for SC+LX+PH ($r_{(SC+LX)vs.(SC+LX+PH)}=0.96$, $r_{(SC+PH)vs.(SC+LX+PH)}=0.97$), showing that either SC+LX or SC+PH is an excellent substitute for all three combined assessment.

International Journal of Oral Science (2014) 6, 50–55; doi:10.1038/ijos.2013.72; published 18 October 2013

Keywords: correlation analysis; orthodontic treatment outcome; reliability; subjective assessment

INTRODUCTION

Since orthodontics was first introduced into China in the 1950s, it has experienced an early stage (from 1950s to late 1970s) and a rapidly developing later stage (1980s till now).¹ With recent economic and social developments, increasing numbers of Chinese people are seeking orthodontic treatment for improved facial appearance and oral health. There are now more than 300 000 Chinese patients accepting orthodontic treatment each year,² and government spending on orthodontic education is increasing to meet this great demand. At present, there are 37 universities conducting postgraduate orthodontic programs.³ More than 2 000 orthodontic specialists and thousands of general practitioners are delivering orthodontic treatment around China. Hence, it is opportune to evaluate the effectiveness of orthodontic treatment provided by different orthodontic services.

Assessment of orthodontic treatment outcome is an important issue that has been discussed since the 1970s.^{4–5} An objective evaluation

system is necessary to determine whether a treatment outcome is successful or unsuccessful. Several orthodontic treatment outcome indicators have been used over the past 10–20 years, including the Peer Assessment Rating (PAR)^{6–7} established in 1992 in Europe, the Objective Grading System^{8–9} drawn up in 1998 by the American Board of Orthodontics and the Index of Complexity, Outcome and Need (ICON)^{10–12} proposed by a collaboration of centers in Germany, Greece, Hungary, Italy, The Netherlands, Norway, Spain, the United Kingdom and the United States. As a general rule, objective measures are developed from prior subjective evaluations by a group of authorities whose judgment is regarded as the ‘gold standard’.¹³ Thus, the average perception of experienced orthodontists is the cornerstone of systems for the evaluation of orthodontic treatment outcome.

Study casts, cephalometric images and facial photographs are the three essential elements of the evaluation of orthodontic treatment outcome. However, most research has focused mainly on study casts

¹Department of Orthodontics, School and Hospital of Stomatology, Peking University, Beijing, China; ²State Key Laboratory of Oral Diseases, Department of Orthodontics, West China Hospital of Stomatology, Sichuan University, Chengdu, China; ³Department of Orthodontics, School of Stomatology, Fourth Military Medical University, Xi’an, China; ⁴Department of Orthodontics, School of Stomatology, Capital Medical University, Beijing, China; ⁵Department of Orthodontics, Institute of Stomatology, Nanjing Medical University, Nanjing, China; ⁶Department of Orthodontics, School and Hospital of Stomatology, Wuhan University, Wuhan, China; ⁷Key Laboratory for Oral Biomedical Engineering, Wuhan University, Wuhan, China; ⁸Department of Orthodontics, School of Stomatology, Shanghai Jiao Tong University, Shanghai, China; ⁹Department of Orthodontics, School of Dentistry, University of the Pacific, San Francisco, USA and ¹⁰School of Mathematical Sciences, Peking University, Beijing, China

Correspondence: Dr TM Xu, Department of Orthodontics, School and Hospital of Stomatology, Peking University, 22 Zhongguancun South Street, Beijing 100081, China

E-mail: tmxuortho@gmail.com

Accepted 13 August 2013

or dental esthetics.^{5–14} Owing to the difficulty of measuring and integrating data from cephalometric images and facial photographs, these information sources have often been ignored. In the present study, to achieve a comprehensive evaluation, study casts, lateral cephalogram and facial photographs were all involved in the assessments.

The aims of this study were to assess the agreements among experienced Chinese orthodontists when evaluating the treatment outcome in Chinese patients, to quantify the contribution of each of the three information sources to the combined evaluation and to determine the correlations among the three information sources in the combined assessments.

MATERIALS AND METHODS

Six orthodontic treatment centers in different parts of China collaborated in this study, including the Peking University School of Stomatology, the West China School of Stomatology at Sichuan University, the School of Stomatology at the Fourth Military Medical University, the Beijing Stomatological Hospital and School of Stomatology at Capital Medical University, the Stomatological Hospital and College of Nanjing Medical University and the Hospital of Stomatology at Wuhan University. At each center, a list of no fewer than 300 patients with full records for whom orthodontic treatment had been completed between 2005–2008 was collected. The initial lists from the six centers, which totaled 2 383 patients, were forwarded to the Peking University School of Stomatology for further processing including randomization. A stratified random sample of 108 patients was drawn from the larger sample, balanced to contain 18 cases from each collaborating center and equal numbers of Angle Class I, Class II and Class III malocclusions. This sample was further randomized with adaptive allocation to produce nine groups containing 12 patients each. Each group contained four Angle Class I, four Angle Class II and four Angle Class III treated patients. Seventy-two patients were aged less than 18 years; the remaining 36 were adults 18 years of age and older. There were 30 males and 78 females. Signed informed consents of the patients and institutional permissions from all six participating orthodontic treatment centers for the study were obtained.

For each patient, the post-treatment information sources evaluated in this study included (i) standardized plaster study casts (SC); (ii) a lateral cephalometric X-ray image (LX); and (iii) facial photographs (PH) that included profile, frontal and frontal smiling views (Figure 1). These information sources were evaluated both singly and in combination by a panel of 69 judges assembled for the purpose at Peking University School of Stomatology for a dedicated period of 3 days. The panel of judges included 69 experienced orthodontic specialists recommended by the six participating treatment centers in consideration of their representativeness of different districts of mainland China. The inclusion criteria for judges were that each had:

- (i) an MS or PhD degree in orthodontics or experience as a research supervisor of orthodontic postgraduates;
- (ii) no less than 10 years of clinical experience in orthodontics;
- (iii) the academic rank of Associate Professor or above.

Thirty-eight judges were males and 31 were females.

The three information sources were evaluated individually and in the following combinations: study casts and lateral cephalogram (SC+LX); study casts and facial photographs (SC+PH); study casts, cephalogram and facial photographs combined (SC+LX+PH). Judges were invited to examine the upper and lower study casts in

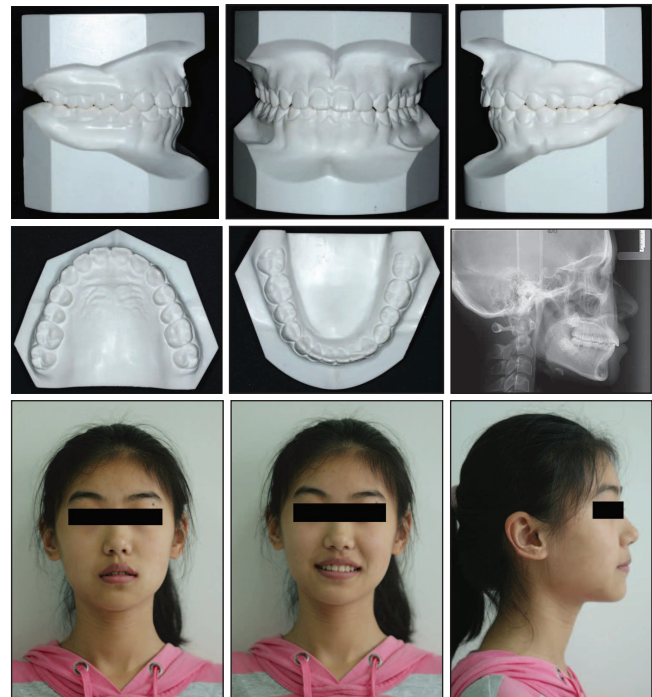


Figure 1 Representative case records of one patient. Each patient's records contained study casts, a lateral cephalometric X-ray image and a triplet of facial photographs.

occlusion and separately, manipulating them by hand. They made the assessment for 'facial attractiveness' from the end-of-treatment lateral cephalometric X-ray images and facial photographs. They were not instructed which component should be prioritized, but made the judgments based on their own clinical experience.

For each group of records, two separate assessments were made. In the first assessment (ranking), each judge ordered the 12 records in each group from 1 (most favorable) to 12 (least favorable) with respect to treatment outcome. In the second assessment (grading), the judge divided the group of 12 cases into three categories: satisfactory, acceptable and unsatisfactory. This procedure served as a protection against chance concentration of more acceptable or less acceptable cases into any individual group of 12 cases. The satisfactory cases were assigned a value of one point, the acceptable cases were assigned two points and the unsatisfactory cases were assigned three points to obtain an average value for each patient. The lower the ordinal rating score, the better the treatment outcome.

Two to four months before commencing the main study, over a dedicated 2-day period and using the proposed criteria, each judge evaluated four groups of cases treated locally.^{15–16}

Statistical analysis

All statistical analyses were performed using Statistical Product and Service Solutions (SPSS) software (V20.0; IBM SPSS Statistics, IBM, Armonk, NY, USA). Spearman rank-order correlation coefficients (r_s) analyses were computed to show agreement in ranking between all judge pairs. The Kendall coefficients of concordance (W) were calculated to determine the overall agreement in rankings and gradings among all judges. Stepwise linear regression and Pearson product-moment correlation coefficient (r) analyses were conducted to assess the contribution of single information source assessments to the combined assessment and to determine correlations among the three

combined assessments. Graphs were generated using MATLAB (R2011b; MathWorks, Natick, MA, USA), Visio (Microsoft Visio Premium 2010; Microsoft, Redmond, WA, USA) and SPSS software. One judge's data for the SC group and two judges' data for the SC+LX group were not obtained. The level of probability for statistical significance was set at $\alpha=0.05$.

RESULTS

Our first task was to determine agreement between and among the subjective evaluations using the Spearman correlation coefficient for pairs of ranking scores and the Kendall coefficient of concordance for group ranking and grading scores.

This produced 2 346 (C_{69}^2) pairs for each comparison. Overall, for all six comparisons, there were 13 873 Spearman r_s values, ranging from as high as +0.852 to as low as -0.077. The maximum, minimum and mean values, and the upper, median and lower quartiles of the r_s values for each ranking are displayed in Figure 2. In total, 103 pairs of r_s failed to reach statistical significance at the 0.05 level, including 0 pairs for SC and SC+LX+PH, 75 pairs for LX, 13 pairs for PH, 7 pairs for SC+LX and 8 pairs for SC+PH (statistical significance at the 0.05 level required an r_s value greater than 0.190). From Figure 2, it can be seen that the judges' average agreement for ranking was relatively high for SC, relatively low for LX and PH and moderate for SC+LX, SC+PH and SC+LX+PH.

The percentages of highly correlated ($r_s>0.7$), moderately correlated ($0.4\leq r_s<0.7$), slightly correlated ($0\leq r_s<0.4$) and negatively correlated ($r_s<0$) pairs are listed for each group in Table 1. Of the 13 873 Spearman r_s values, 10.92% were above 0.70, 79.3% were between 0.40 and 0.70, 9.77% were between 0 and 0.40 and 0.04% were below 0, showing that agreement between individual pairs of judges was highly variable. Agreement also varied among the six test groups.

All Kendall W values were statistically significant ($P<0.05$), indicating that the overall agreement between the judges was good in both ranking and grading (Table 2).

Table 1 Distribution of Spearman r_s values for 108 cases /%

Groups	$r_s\geq 0.7$	$0.4\leq r_s<0.7$	$0\leq r_s<0.4$	$r_s<0$
SC	28.97	68.83	2.19	0.00
LX	2.47	76.17	21.14	0.21
PH	3.32	82.01	14.66	0.00
SC+LX	10.81	82.00	7.19	0.00
SC+PH	10.57	80.14	9.29	0.00
SC+LX+PH	9.89	86.36	3.75	0.00
Average	10.92	79.30	9.77	0.04

LX, lateral cephalometric X-ray image; PH, facial photograph; SC, study cast.

Table 2 Kendall W values of ranking and grading ($P<0.05$, $n=108$)

Kendall W	SC	LX	PH	SC+LX	SC+PH	SC+LX+PH	Mean \pm s.d.
Ranking	0.65	0.50	0.52	0.58	0.57	0.59	0.57 \pm 0.05
Grading	0.59	0.44	0.46	0.52	0.51	0.53	0.51 \pm 0.06

LX, lateral cephalometric X-ray image; PH, facial photograph; SC, study cast; s.d., standard deviation.

Our second task was to assess the contribution of the evaluation of single information source assessments to the combined assessment. First, it was important to determine the correlation between each single information source assessment and the combined effect of all three sources, as illustrated in Figure 3. Here, the average grading scores for each single information source assessment are seen to be correlated with the average grading scores for SC+LX+PH for the same case as assessed by the same judge in each plot.

The relationship between the average grading scores for SC+LX+PH and those for each single information source assessment are further explored using Pearson r analyses in Table 3, which showed that the grading scores for SC predicted those for SC+LX+PH with

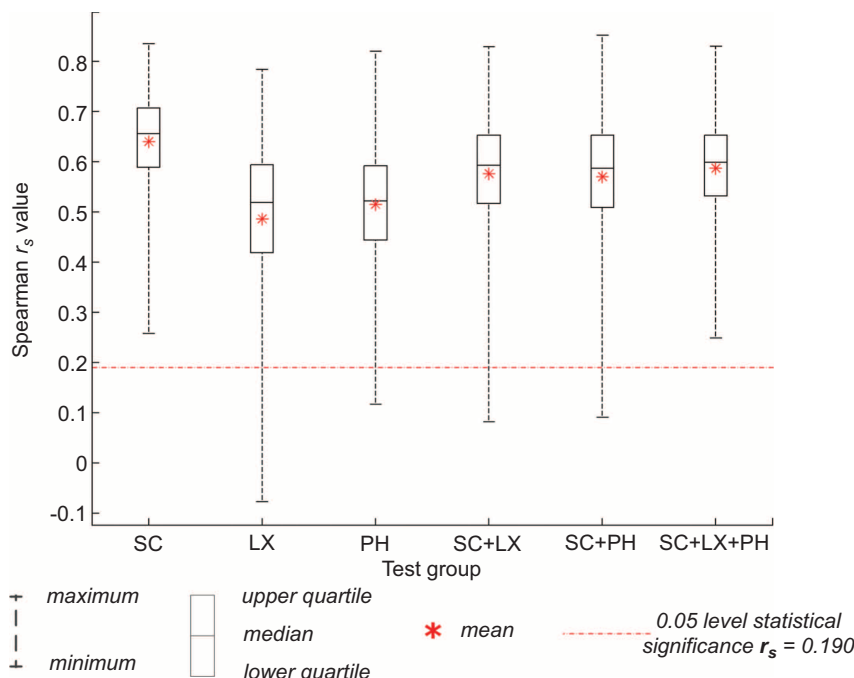


Figure 2 Box and whisker plots showing the judges' agreements for the ranking in each test group ($n=108$). LX, lateral cephalometric X-ray image; PH, facial photograph; SC, study cast.

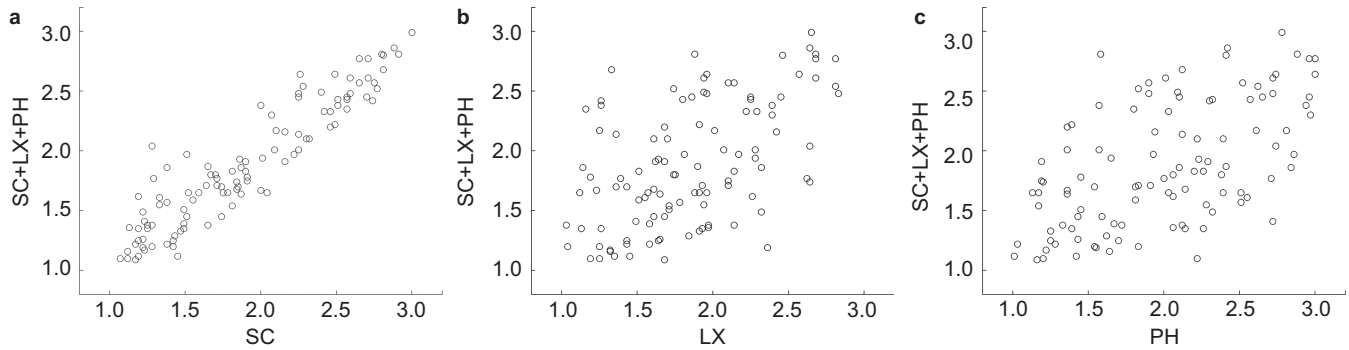


Figure 3 Scatter plots comparing associations between SC, LX, PH and SC+LX+PH. For each scatter plot, $n=108$ and $P<0.05$. LX, lateral cephalometric X-ray image; PH, facial photograph; SC, study cast.

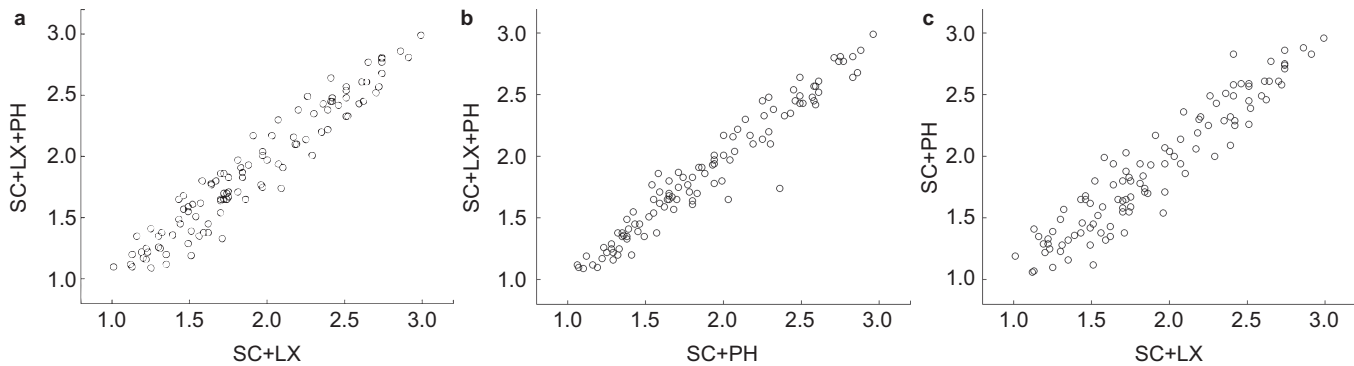


Figure 4 Scatter plots comparing associations among SC + LX, SC+PH and SC+LX+PH. For each scatter plot, $n=108$ and $P<0.05$. LX, lateral cephalometric X-ray image; PH, facial photograph; SC, study cast.

Table 3 Pearson correlation (r) and R^2 value ($n=108$)

Parameters	SC	LX	PH
Pearson r			
LX	0.34		
PH	0.30	0.52	
SC+LX+PH	0.93	0.53	0.57
R^2 values			
SC+LX+PH	0.86	0.28	0.32

LX, lateral cephalometric X-ray image; PH, facial photograph; SC, study cast.

Table 4 Stepwise regression analysis (dependent variable, SC+LX+PH; independent variables, SC, LX, PH; $n=108$)

Step	Variable entered	Partial R^2	Cumulative R^2	P value to entered variable
1	SC	0.863 0	0.863 0	<0.000 1
2	PH	0.091 6	0.954 6	<0.000 1
3	LX	0.009 8	0.964 5	<0.000 1

LX, lateral cephalometric X-ray image; PH, facial photograph; SC, study cast.

extremely high significance ($r=0.93$). However, the grading scores for lateral cephalometric films ($r=0.53$) and facial photographs ($r=0.57$) predicted merely a small portion of those for SC+LX+PH. SC and LX ($r=0.34$) and SC and PH ($r=0.30$) exhibited low correlations. LX and PH were moderately correlated ($r=0.52$).

The results of the linear stepwise regression analysis for the grading scores are presented in Table 4. SC entered into the analysis first, accounting for an R^2 value of 0.86. PH entered next, adding 0.09, and LX entered last, adding a small but statistically significant increment of

0.01, to obtain a total R^2 of 0.96. From these results, it is reasonable to assume that 96% of the variability in the average grading scores for SC+LX+PH can be accounted for by the scores for SC, LX and PH. $[SC+LX+PH] = -0.29 + 0.76 \times [SC] + 0.24 \times [PH] + 0.13 \times [LX] + \text{error}$.

Our third task was to determine correlations among the three information sources in the combined evaluations. As shown in Table 5 and illustrated in Figure 4, in each respective scatter plot, the average grading scores demonstrated extremely high correlations for SC+LX vs. SC+LX+PH ($r=0.96$), SC+PH vs. SC+LX+PH ($r=0.97$) and SC+LX vs. SC+PH ($r=0.95$).

Table 5 Pearson correlation (r) and R^2 values ($n=108$)

Parameters	SC+LX	SC+PH
Pearson r		
SC+PH	0.95	
SC+LX+PH	0.96	0.97
R^2 values		
SC+LX+PH	0.92	0.94

LX, lateral cephalometric X-ray image; PH, facial photograph; SC, study cast.

DISCUSSION

The Visual Analog Scale¹⁷ and Q-Sort assessment¹⁸ are the two methods most commonly used for subjective evaluation. The Visual Analog Scale was used in the development of PAR⁶⁻⁷ and ICON,¹⁰⁻¹¹ the subjective perspective being described by several scales. Q-Sort assessment is used predominantly in psychology,¹⁹ though several studies of occlusal index and smile esthetics have used Q-Sort to achieve a

logical subjective evaluation.^{20–21} Sixty to 120 items are placed in order of representativeness, the most characteristic being given a high score and the least characteristic a low score.

When subjectively evaluating end-of-treatment cases, because the differences between patients tend to be small, judges are inclined to mark only approximately, using a moderate scale on which it is difficult to distinguish details. Moreover, it is impracticable to rank a large number of cases at one time. In this study, 12 cases were assessed in each information source, using study casts, cephalometric films and/or facial photographs. For ranking, the judges were asked to examine each patient's case records carefully and to compare them one-by-one within each test group to produce a sequence of 12 cases. For grading, the judges were asked to divide the 12 ordered cases into three ordinal categories—satisfactory, acceptable and unsatisfactory—to allow comparisons among the groups. By using both ranking and grading assessments, detailed and reliable results were obtained.

Validity²² and reliability²³ are two basic necessities for any evaluation system or index. In this study, we were concerned with the agreement among 69 experienced orthodontists of their subjective perceptions of orthodontic treatment outcome. Subjective perceptions can be averaged to obtain a 'gold standard' only when the reliability is good.

In previous studies on agreement among orthodontists, many measures of association and concordance have been used, including the intraclass correlation coefficient (ICC) statistics and kappa statistics, the Pearson and Spearman correlation tests, the Kendall correlation analysis and the root mean square error. Except for the root mean square error, the coefficient or score for each test is a value between 0 and +1.0 or -1.0 and +1.0, where values closer to +1.0 represent higher reliability. The root mean square error is an estimate of the standard deviation representing the measurement error of a single measurement.^{23–24}

The ICC statistic employs repeated measures analysis of variance, use of which is appropriate only when the underlying model is assumed to be a one-way random effects model.²⁵ ICCs are regularly used in the analysis of measurement data such as the objective scores of PAR,⁶ American Board of Orthodontics-Objective Grading System²⁶ and ICON.¹³ Pinskaya *et al.*²⁷ reported excellent measuring repeatability, with ICCs of 0.97 and 0.98 for intra-examiner reliability, and an ICC of 0.98 for inter-examiner reliability.

Kappa statistics²⁸ are widely used to assess the reliability of categorical measurement; for example, the Dental Health Component of the Index of Orthodontic Treatment Need²⁹ and judges' subjective perception in the PAR and ICON.^{6–7,13} When developing the PAR index in 1992, Richmond *et al.*⁶ reported kappa scores of 0.39–0.87 for intra-judge reliability and 0.43–0.58 for inter-judge reliability. In a study validating the complexity and treatment outcome components of ICON, Nicholas *et al.*¹³ reported that the inter-rater kappa agreements were moderate (0.50) for complexity, slight (0.18) for outcome and poor (0.04) for degree of improvement.

The Pearson and Spearman correlation tests are used to assess reliability between pairs of judges ranking data that are distributed normally or non-normally, respectively.^{30–34} Xu *et al.*³⁰ used the Pearson correlation test to assess agreement and disagreement among pairs of Chinese and US orthodontists. They found that the correlations ranged from +0.004 to +0.96 with a median of +0.54.

The Kendall correlation analysis evaluates agreement in a group of judges making ordered categorical measurements.³⁵ When testing the agreement among six judges in an early study developing the occlusal

index,³⁶ the high Kendall coefficient obtained of 0.881 indicated that the judges were applying essentially the same criteria in their assessments.

The present study addressed inter-judge reliability related to subjective evaluation; that is, the extent to which ranking and grading evaluations were repeatable among the 69 judges. Two aspects of judge homogeneity were considered: agreement between judge pairs and agreement among the whole panel. Spearman analysis indicated that, for ranking, the agreement between judge pairs tended to be strongly significant, which was similar to our previous study.¹⁶ All Kendall coefficients of concordance were statistically significant, demonstrating high consistency among the panel of Chinese judges for both the ranking assessment and the grading assessment.

However, Spearman correlations varied among the six test groups. The mean Spearman correlation was the highest for SC compared with the other five comparisons, and most of the correlations for SC were above 0.4. Orthodontists aim to achieve as near perfect occlusion as possible, as defined in Andrew's six keys to occlusion.³⁷ Due to their similar background knowledge and clinical experience, the 69 judges provided an almost standardized subjective assessment of the study casts, with high inter-judge reliability. The mean Spearman r_s value was the lowest for LX, where 21.4% of correlations were below 0.4 and 75 pairs of Spearman r_s failed to reach statistical significance at the 0.05 level. This considerable disagreement could be attributed to the diversity of cephalometric analytical methods and contrary opinions about orthodontic camouflage treatment. The mean Spearman r_s values for SC+LX+PH, SC+LX and SC+PH were comparable (Figure 1 and Table 1). However, all pairs of Spearman r_s values for SC+LX+PH were statistically significant at the 0.05 level, whereas seven pairs of those for SC+LX and eight pairs of those for SC+PH were statistically non-significant. The Spearman correlation results for PH in this study were similar to those of serial studies of facial attractiveness.^{30–31}

Our study is the first attempt to use combinations of study casts, cephalometric films and facial photographs for the subjective evaluation of orthodontic treatment outcome. In the past, these different materials have usually been assessed separately. It was important to determine the relationship between the different sources and the contribution of single information source assessments to the combined evaluation. The Pearson correlation test and stepwise regression were conducted for this purpose.

Table 3 showed that, when assessed separately, the three types of end-of-treatment records were correlated with each other to some extent. The Pearson correlation between SC and LX was low, as was that between SC and PH ($r_{SC\text{vs.}LX}=0.34$, $r_{SC\text{vs.}PH}=0.30$, $P<0.001$). The correlation between LX and PH was moderate ($r_{LX\text{vs.}PH}=0.52$, $P<0.001$), being similar to the finding of Oh *et al.*³⁸ This higher, moderate association between the cephalometric films and facial photographs is probably related to their shared information about the facial profile.

Stepwise regression analysis illustrated how the three single information sources complemented one another when combined. Table 4 showed that the grading scores for SC predicted 86.3% of the grading scores for SC+LX+PH, with the scores for PH increasing this to 95.5% and LX to 96.5% ($P<0.001$). In the regression model, though SC entered into the regression equation first, PH and LX both added statistically significantly to the R^2 value. Tables 3 also showed that the grading scores for LX separately predicted 28% of the variability for SC+LX+PH ($P<0.001$), while the grading scores for PH separately predicted 32% of the variability for SC+LX+PH ($P<0.001$).

Regarding correlations between the three combined sources, Table 5 showed that the grading scores for SC+LX, SC+PH and SC+LX+PH were highly significantly correlated with each other ($r_{(SC+LX) \times (SC+PH)}=0.95$, $r_{(SC+LX) \times (SC+LX+PH)}=0.96$, $r_{(SC+PH) \times (SC+LX+PH)}=0.97$), which explains the similar mean agreement between the combined evaluations. R^2 values indicated that the grading scores for SC+LX and SC+PH separately predicted 92% and 94%, respectively, of the variability of the grading scores for SC+LX+PH. In view of this, it is reasonable to conclude that the grading scores for SC+LX and SC+PH predicted the grading scores for SC+LX+PH extremely well.

Based on the above discussed data, it is highly desirable to include both lateral cephalometric films and facial photographs for a comprehensive evaluation of orthodontic treatment outcome, though study casts play a significant role in the combined assessment.

CONCLUSIONS

A panel of 69 Chinese orthodontists demonstrated mainly moderate to high consistency for each group of end-of-treatment study casts, cephalometric films and facial photographs from 108 Chinese patients. In the combined group evaluations, study casts were the most significant predictive component, while lateral cephalometric films and facial photographs also contributed to a comprehensive assessment. Since the grading scores for SC+LX and SC+PH were highly significantly correlated with those for SC+LX+PH, either study casts plus lateral cephalometric films or study casts plus facial photographs is an excellent substitute for all three combined assessment methods.

ACKNOWLEDGEMENTS

We thank all the patients and 69 orthodontists who agreed to take part in this study. We also thank the West China School of Stomatology at Sichuan University, the School of Stomatology at the Fourth Military Medical University, the Beijing Stomatological Hospital and School of Stomatology at Capital Medical University, the Stomatological Hospital and College of Nanjing Medical University and the Hospital of Stomatology at Wuhan University. This study would not have been possible without their participation. We thank Dr Edward L Korn from the National Institute of Health for his invaluable assistance in designing this study. This work was supported by the Specific Research Project of Health Pro Bono Sector, Ministry of Health, China (200802056).

- 1 Lin JX, Xu TM. [History and development of Chinese orthodontics.] *Beijing Da Xue Xue Bao* 2008; **40**(1): 11–14. Chinese.
- 2 Fu MK. [An insight into orthodontics in China.] *Zhonghua Kou Qiang Yi Xue Za Zhi* 2004; **39**(2): 89–90. Chinese.
- 3 Yeweng SJ, Huang SF, Ren LJ. Orthodontics in China. *J Orthod* 2002; **29**(1): 62–65.
- 4 Pickering EA, Vig P. The occlusal index to assess orthodontic treatment. *Br J Orthod* 1975; **2**(1): 47–51.
- 5 Elderton RJ, Clark JD. Orthodontic treatment in the general dental services assessed by the Occlusal Index. *Br J Orthod* 1983; **10**(4): 178–186.
- 6 Richmond S, Shaw WC, O'Brien KD *et al*. The development of the PAR Index (Peer Assessment Rating): reliability and validity. *Eur J Orthod* 1992; **14**(2): 125–139.
- 7 Richmond S, Shaw WC, Roberts CT *et al*. The PAR index (Peer Assessment Rating): methods to determine outcome of orthodontic treatment in terms of improvement and standards. *Eur J Orthod* 1992; **14**(3): 180–187.
- 8 Casko JS, Vaden JL, Kokich VG *et al*. Objective grading system for dental casts and panoramic radiographs. American Board of Orthodontics. *Am J Orthod Dentofacial Orthop* 1998; **114**(5): 589–599.
- 9 Costalos PA, Sarraf K, Cangialosi TJ *et al*. Evaluation of the accuracy of digital model analysis for the American Board of Orthodontics objective grading system for dental casts. *Am J Orthod Dentofacial Orthop* 2005; **128**(5): 624–629.

- 10 Richmond S, Daniels CP. International comparisons of professional assessments in orthodontics: Part 1—Treatment need. *Am J Orthod Dentofacial Orthop* 1998; **113**(3): 180–185.
- 11 Richmond S, Daniels CP. International comparisons of professional assessments in orthodontics: Part 2—Treatment outcome. *Am J Orthod Dentofacial Orthop* 1998; **113**(3): 324–328.
- 12 Daniels C, Richmond S. The development of the index of complexity, outcome and need (ICON). *J Orthod* 2000; **27**(2): 149–162.
- 13 Savastano NJ, Firestone AR, Beck FM *et al*. Validation of the complexity and treatment outcome components of the index of complexity, outcome, and need (ICON). *Am J Orthod Dentofacial Orthop* 2003; **124**(3): 244–248.
- 14 Pangrazio-Kulbersh V, Kaczynski R, Shunock M. Early treatment outcome assessed by the Peer Assessment Rating index. *Am J Orthod Dentofacial Orthop* 1999; **115**(5): 544–550.
- 15 Liu SQ, Shen G, Bai D *et al*. [Consistency of the subjective evaluation of malocclusion severity by the Chinese orthodontic experts.] *Beijing Da Xue Xue Bao* 2012; **44** (1): 98–102. Chinese.
- 16 Song GY, Li WR, Geng Z *et al*. [Agreement analysis of subjective evaluation of orthodontic treatment outcome.] *Beijing Da Xue Xue Bao* 2012; **44** (1): 103–107. Chinese.
- 17 Phillips C, Tulloch C, Dann C. Rating of facial attractiveness. *Community Dent Oral Epidemiol* 1992; **24**(4): 214–220.
- 18 Stephenson W. *The study of behavior: Q-technique and its methodology*. Chicago: University of Chicago Press, 1953: 376.
- 19 Block J. *The Q-sort method in personality assessment and psychiatric research*. Philadelphia: Charles C Thomas Publisher, 1961: 5.
- 20 Airton OA. Occlusal indexes as judged by subjective opinions. *Am J Orthod Dentofacial Orthop* 2008; **134**(5): 671–675.
- 21 Cox NH, Frans PG. Facial harmony. *Am J Orthod* 1971; **60**(2): 175–183.
- 22 Liao ZY, Jian F, Long H *et al*. Validity assessment and determination of the cutoff value for the index of complexity, outcome and need among 12–13 year-olds in southern Chinese. *Int J Oral Sci* 2012; **4**(2): 88–93.
- 23 Roberts CT, Richmond S. The design and analysis of reliability studies for the use of epidemiological and audit indices in orthodontics. *Br J Orthod* 1997; **24**(2): 139–147.
- 24 Onyeaso CO, Begole EA. Relationship between index of complexity, outcome and need, dental aesthetic index, peer assessment rating index, and American Board of Orthodontics objective grading system. *Am J Orthod Dentofacial Orthop* 2007; **131**(2): 248–252.
- 25 Bartko JJ. The intraclass correlation coefficient as a measure of reliability. *Psychol Rep* 1966; **19**(1): 3–11.
- 26 Lieber WS, Carlson SK, Baumrind S *et al*. Clinical use of the ABO-scoring index: reliability and subtraction frequency. *Angle Orthod* 2003; **73**(5): 556–564.
- 27 Pinskaya YB, Hsieh TJ, Roberts WE *et al*. Comprehensive clinical evaluation as an outcome assessment for a graduate orthodontic program. *Am J Orthod Dentofacial Orthop* 2004; **126**(5): 533–543.
- 28 Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1997; **33**(1): 159–174.
- 29 Bowden DE, Davies AP. Inter- and intraexaminer variability in assessment of orthodontic treatment need. *Community Dent Oral Epidemiol* 1975; **3**(4): 198–200.
- 30 Xu TM, Korn EL, Yan L *et al*. Facial attractiveness: ranking of end-of-treatment facial photographs by pairs of Chinese and US orthodontists. *Am J Orthod Dentofacial Orthop* 2008; **134**(1): 74–84.
- 31 Liu Y, Korn EL, Oh HS *et al*. Comparison of Chinese and US orthodontists' averaged evaluations of "facial attractiveness" from end-of-treatment facial photographs. *Am J Orthod Dentofacial Orthop* 2009; **135**(5): 621–634.
- 32 Evans R, Shaw W. Preliminary evaluation of an illustrated scale for rating dental attractiveness. *Eur J Orthod* 1987; **9**(4): 314–318.
- 33 Woolass KF, Shaw WC. Validity and reproducibility of rating dental attractiveness from study casts. *Br J Orthod* 1987; **14**(3): 187–190.
- 34 Brown R, Richmond S. An update on the analysis of agreement for orthodontic indices. *Eur J Orthod* 2005; **27**(3): 286–291.
- 35 Kendall MG. A new measure of rank correlation. *Biometrika* 1938; **30**(2): 81–93.
- 36 Summers CJ. The occlusal index: a system for identifying and scoring occlusal disorders. *Am J Orthod* 1971; **59**(6): 553–567.
- 37 Andrews LF. The six keys to normal occlusion. *Am J Orthod Dentofacial Orthop* 1972; **62**(3): 296–309.
- 38 Oh HS, Korn EL, Zhang XY *et al*. Correlations between cephalometric and photographic measurements of facial attractiveness in Chinese and US patients after orthodontic treatment. *Am J Orthod Dentofacial Orthop* 2009; **136**(6): 762.e1–762.e14.



This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivative Works 3.0 Unported License. To view a copy of this license, visit <http://creativecommons.org/licenses/by-nc-nd/3.0>