European Journal of Orthodontics, 2017, 1–7
doi:10.1093/ejo/cjx043

Original article

Interpreting weightings of the peer assessment rating index and the discrepancy index across contexts on Chinese patients

Siqi Liu[1], Heesoo Oh[2], David William Chambers[3], Tianmin Xu[4] and Sheldon Baumrind[2]

[1]First clinical division, Peking University School and Hospital of Stomatology, Beijing, China, Departments of [2]Orthodontics and [3]Dental Practice, University of the Pacific, Arthur A. Dugoni School of Dentistry, San Francisco, CA, USA and [4]Department of Orthodontics, Peking University School and Hospital of Stomatology, Beijing, China

Correspondence to: David William Chambers, Department of Dental Practice, University of the Pacific, Arthur A. Dugoni School of Dentistry, 155 Fifth Street, San Francisco, CA94103, USA. E-mail: dchamber@pacific.edu

Summary

Objective: Determine optimal weightings of Peer Assessment Rating (PAR) index and Discrepancy Index (DI) for malocclusion severity assessment in Chinese orthodontic patients.

Methods: Sixty-nine Chinese orthodontists assessed a full set of pre-treatment records from a stratified random sample of 120 subjects gathered from six university orthodontic centres. Using professional judgment as the outcome variable, multiple regression analyses were performed to derive customized weighting systems for the PAR index and DI, for all subjects and each Angle classification subgroup.

Results: Professional judgment was consistent, with an Intraclass Correlation Coefficient (ICC) of 0.995. The PAR index or DI can be reliably measured, with ICC = 0.959 and 0.990, respectively. The predictive accuracy of PAR index was greatly improved by the Chinese weighting process (from $r = 0.431$ to $r = 0.788$) with almost equal distribution in each Angle classification subgroup. The Chinese-weighted DI showed a higher predictive accuracy, at $P = 0.01$, compared with the PAR index ($r = 0.851$ versus $r = 0.788$). A better performance was found in the Class II group ($r = 0.890$) when compared to Class I ($r = 0.736$) and III ($r = 0.785$) groups.

Conclusions: The Chinese-weighted PAR index and DI were capable of predicting 62 per cent and 73 per cent of total variance in the professional judgment of malocclusion severity in Chinese patients. Differential prediction across Angle classifications merits attention since different weighting formulas were found.

Introduction

It is an ideal to have a perfect measurement system for determining the extent of malocclusion. In practice, this project would be complicated by multiple sources of variation, as schematically presented in Figure 1. A perfect agreement of professional judgments for the presenting condition of malocclusion, depicted as the thin line leading from 'presenting condition' to 'professional judgments' in the figure, is the ideal situation. In statistical terms, this would be reflected as a coefficient of generalizability, a general type of the Intraclass Correlation Coefficient (ICC), of 1.000 (1). However, when different standards are applied, disagreement has been found in professional judgment for presenting malocclusions under specific conditions. Efforts to identify and control these multiple sources of variation have been made through previous research (2).

A meta-analysis by Meehl (3) demonstrated that computers using diagnostic rules generated by psychiatrists were consistently superior at predicting the presence of various mental conditions and behavioural disturbances than the psychiatrists themselves. The analogue

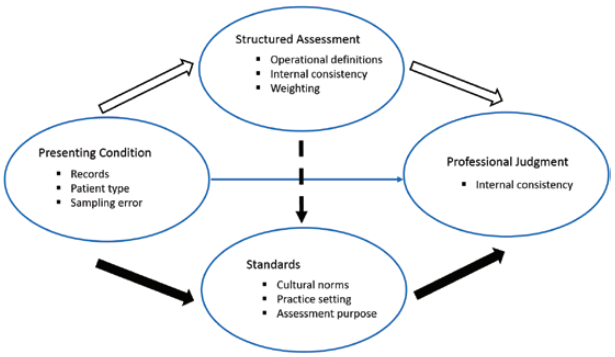

Figure 1. Schematic representation of determining professional judgment from presenting condition.

in orthodontics is to develop one or more 'objective' indices of malocclusion. In Figure 1, 'objective' scoring is represented by the hollow arrows leading from 'Presenting Condition' to 'Professional Judgment' through 'Structured Assessment'. When organizing structured assessments, researchers may face issues such as 1. identifying and unambiguously defining the right operational components, 2. determining the optimal weighting of the components for creating a scale (index), and 3. achieving strong calibration for internal consistency.

The Peer Assessment Rating (PAR) index (4) and the Discrepancy Index (DI) (5) are examples of protocols for combining scores on operationally defined characteristics of various patient records. The PAR index has been reported as having strong internal consistency and has been found to predict clinical assessments, especially when its components are weighted (4, 6, 7). Similarly, the DI has been demonstrated to be a reliable index to assess malocclusion complexity but has not been widely validated against professional judgment (8). Even when these assessment systems are well managed, they often account for no more than two-thirds of the variance in the standard (4, 6). To the extent that structured assessments use only a subset of available records, draw unrepresentative samples, use the same sample to both develop and verify the structured system, and generalize a single system across clinically important differences in patient types, they will produce estimates of malocclusion that are suboptimal.

Other sources of unwanted variance may include malalignment of the measurement system with the target population. Sources of variance related to the 'Standards' are grouped in the oval at the bottom of Figure 1. An index that works well in one situation or for a particular purpose is not guaranteed to be equally effective in other settings. The literature contains examples demonstrating the existence of various cultural norms (9), diversity across types of practitioner (4), and preferences for types of scales depending on whether one is using them to determine treatment plans or for public health monitoring or for research on comparative delivery systems or treatment modalities (2). Context-specific scoring is represented by the solid arrows leading from 'Presenting Condition' to 'Professional Judgment' through 'Standards'.

It is hypothesized in this study that incorporating the objective scoring path with the context-specific scoring path, by following the broken solid arrow in Figure 1, would increase the amount of explained variance between 'Presenting Condition' and 'Professional Judgment'. In order to systematically explore the relationship among the sources of variance identified in the figure, a dataset of professional judgments of malocclusion in a Chinese population was used to compare differences between two structured measurements (PAR index and DI) across patients with different Angle classifications using weighting formulas developed in different countries. It is anticipated that the differential distribution of Angle classifications in the Chinese population, with its higher prevalence of Angle Class III malocclusions (10) and possibly a different esthetic standard among Chinese orthodontists, will be better reflected in a modification of component weighting in the PAR index and DI compared with previously published standards.

Materials and Methods

The present study was designed as a multicentre, prospective investigation. Following a protocol suggested by Song, *et al.* (11), six orthodontic treatment centres located in various parts of China participated in this study from January to June 2010. To form a diverse patient sample, each centre collected complete medical records for at least 250 patients who had treatment completed between July 2005 and September 2008. A sample randomly stratified on Angle classification documented in the patient chart was used (Supplemental Figure 1). From a combined total of 2383 records, 21 were drawn from each centre to create a sample of 126 subjects that consisted of equal numbers of Angle Class I, Class II, and Class III subjects. Patient records included study casts, lateral cephalometric radiographs, panoramic radiographs, facial photographs (front, lateral, and front smile views), and treatment charts. Six subjects were excluded before data acquisition because their casts were accidentally damaged during the clinician evaluation session. The final sample consisted of 120 subjects.

A panel of judges was used to obtain a standard for determining the overall malocclusion severity of patients presenting for treatment. Twelve judges were recruited from each of the participating centres basing on these inclusion criteria: 1. more than 8 years of clinical experience in orthodontics; 2. an M.S. or Ph.D. degree in orthodontics; and 3. an academic rank of associate professor or above. Three judges were dropped because of schedule conflicts, leaving a total of 69 judges. Each of the 120 subjects in the study was examined by each of the 69 judges. Given a full set of diagnostic records, each judge rated the malocclusion severity of each subject using a 5-point rating scale (1-mild, 2-mildly moderate, 3-moderate, 4-severely moderate, and 5-severe). The scale was anchored in the clinicians' overall professional judgment, intentionally avoiding the introduction of bias on behalf of the researchers that might have directed clinicians to favour one dimension or another. The average of all 69 orthodontists' scores on the 5-point scale was used as the clinical judgment score for each subject.

Three second-year orthodontic residents scored the 120 sets of initial study casts and standardized cephalometric and panoramic radiographs that had undergone a pixel conversion process. Preliminary calibration sessions were carried out using 10 randomly selected subjects. Intra- and inter-rater reliability were tested by ICC. The components with an ICC value of less than 0.750 were discussed and measured again. After repeating three calibration sessions, there was no component with an ICC value of less than 0.750. Having been calibrated, the three raters independently recorded the raw score for each of the 11 components of PAR index (4) and 12 target disorders of DI (5). The raw scores for each component or disorder were averaged across the three raters for further analysis.

Data analysis

- Normality, uniformity of variances, and absence of collinearity of the data and other assumptions required for intraclass correlation and regression analysis were verified before statistical tests were performed.

- Intraclass correlation coefficients (ICC), which were determined by Cronbach's generalizability analysis (1), were calculated for the 69 judges to test inter-examiner reliability. Consistency across raters for overall index consistency and the consistency of ratings of individual components for PAR index and DI were determined in the same fashion.
- Predictive accuracy of unweighted PAR index and DI were tested by multiple regression using overall and component scores as predictors and clinical judgment scores as predicted values, both overall and by each Angle Class subgroup.
- Customized weighting systems of the indices were derived through multiple regression analysis by entering the clinical judgment score as the dependent variable and the unweighted component scores for PAR index and DI as the independent variable. Components that failed to reach a P value of 0.100 or smaller were excluded from further consideration on the assumption that they added no statistically significant predictive power. The regression analysis was repeated using the reduced set of significant predictor components. The resulting partial regression coefficients were multiplied by 10 and rounded to the nearest 0.5 to determine the component weights. This weighting procedure was performed separately for PAR index and DI components for all 120 subjects as an overall group and for Angle Class I, II, and III subjects as separate subgroups. Age and Angle classification were used as control variables for the entire group and age as a control variable for subgroups by Angle classification.
- Differences between the strength of inter-examiner reliability and predictive r values were tested with Fisher's test for independent associations and the test for related correlations, as appropriate.
- Monte Carlo analysis was performed post hoc to explore the effect of sampling on the stability of component weighting across indices (12). For the subsample with the highest predicted value, samples were randomly generated by a computer algorithm, each containing a set of subjects used for developing weightings and a separate set of subjects of approximately the same size to be used to determine the predictive value of the weightings. The regression procedure described above was used to generate weighting values with one set of subjects, and these weights were verified by separate regression analysis on cases not included to create the weightings. This procedure was performed 50 times and results were averaged.

Results

The final sample was composed of 33 male and 87 female patients and distributed as 38 Angle Class I subjects, 40 Angle Class II subjects, and 42 Angle Class III subjects. The mean age was 17.1 ± 6.3 years, ranging from 10 to 41. There were no statistically significant differences between Angle classifications with regard to sex and age. The panel of 69 orthodontists was comprised of 35 male and 34 female orthodontists. The mean age was 45.2 ± 7.3 years, ranging from 35 to 72 years. The orthodontists had an average of 19 ± 7.2 years of practice experience, ranging from 8 to 49 years. The judges did not show statistically significant differences related to their age, sex, years of experience, or residing centres.

Inter-examiner reliability, ICC, of the 69 judges on overall clinical assessment of malocclusion severity was 0.995. Overall, the inter-examiner reliability of the three raters for the PAR index and the DI scale were 0.959 and 0.990, respectively. The inter-examiner reliability on each separate PAR index components ranged from 0.912 to 0.999 and for DI components from 0.835 to 0.999, as shown in Tables 1 and 2.

Predictive accuracy of current PAR index and DI

A modest correlation was found between the unweighted PAR index and overall average clinical judgments (r = 0.431) for the total sample (Table 3). Using weighting formulas from various countries previously reported in the literature improved the predictive accuracy of PAR index for the present sample: the Australian (AU) weightings (7) raised the unweighted predictive r value from 0.431 to 0.554; United Kingdom (UK) weightings (4) raised the r value to 0.647; and USA weighting (6) raised the r value to 0.701. When compared to the PAR index, the correlation between unweighted DI scores and clinical judgment was relatively higher (r = 0.680) (Table 3).

Customized PAR index and DI weights for Chinese sample

The r values for all variables (regardless of individual statistical significance) are shown in the preliminary weighted row, and the r values for the selected and weighted formulas are shown in the custom weighted row in Table 3 for PAR index and DI. It can be seen in Table 3 that custom weighted PAR index improved prediction of clinical judgments from r = 0.431 to r = 0.788, and custom weighted DI gained predictive power from r = 0.680 to r = 0.851. The predictions using custom weights were statistically superior to the unweighted scores at $P < 0.001$ for PAR index and $P < 0.01$ for DI.

Tables 1 and 2 display the individual component weightings, respectively for custom weighted PAR index and custom weighted DI. Overjet and occlusion relationship were common to PAR index and DI as significant weighting components. Other components such

Table 1. Custom weighted components of Peer Assessment Rating (PAR) index

	Multiple regression analysis				
		Angle classification			
Components	Whole sample (n = 120)	Class I (n = 38)	Class II (n = 40)	Class III (n = 42)	ICC
Overjet	3	4	5.5	2.5	0.999
Overbite	3.5	0	2.5	3.5	0.999
Centreline	1.5	0	3	0	0.999
R and L buccal occlusion*	1.5	2.5	0	1.5	0.921
Upper anterior displacement#	0	0	0	0	0.912
Lower anterior displacement#	0	0	0	0	0.999

*The Right buccal occlusion and Left buccal occlusion components of unweighted PAR index were merged following Richmond *et al.*'s protocol (4).

#The Upper right and left displacement and Lower right and left displacement components of unweighted PAR index were omitted without being weighted in any of weighting systems.

Table 2. Custom weighted components of Discrepancy Index (DI)

	Multiple regression analysis				
		Angle classification			
Components	Whole sample (n = 120)	Class I (n = 38)	Class II (n = 40)	Class III (n = 42)	ICC
Overjet	3	3.5	3.5	2	0.928
Overbite	0	0	0	0	0.912
Crowding	1	1	0	1	0.878
Occlusion relationship	1	0	1.5	0	0.931
Anterior open bite	2	0	0	2.5	0.922
Lateral open bite	0	0	3	0	0.835
Lingual posterior crossbite	0	0	2.5	1	0.930
Buccal posterior crossbite	1	0	1	0	0.879
ANB	1	1	0	0	0.999
SN-MP	0	0	1	0	0.999
L1 –MP	0	0	0	0	0.999
Other	1.5	1	1	1	0.999

Table 3. Predictive accuracy of Peer Assessment Rating (PAR) index and Discrepancy Index (DI)

		Multiple regression analysis			
			Angle classification		
Indices		Whole sample (n = 120)	Class I (n = 38)	Class II (n = 40)	Class III (n = 42)
PAR index	Unweighted PAR (r)	0.431	0.443	0.152	0.559
	Preliminary weighted PAR (r)*	0.790	0.716	0.751	0.767
	Custom weighted PAR (r)#	0.788	0.652	0.736	0.758
DI	Unweighted DI (r)	0.680	0.518	0.780	0.654
	Preliminary weighted DI (r)*	0.852	0.885	0.910	0.828
	Custom weighted DI (r)#	0.851	0.736	0.890	0.785

*Regression correlation coefficient (r) from all components.
#Regression correlation coefficient (r) from the weighted components only.

as displacement in PAR index and overbite in DI achieved no significant weightings in either system. ANB, SN-MP, and 'Other' were contributing components in the DI index, although they were not measured in the PAR index.

There was a weak correlation between unweighted PAR and DI overall indices (r = 0.315). When weights appropriate to the Chinese sample were applied, the correlation between the two indices rose to r = 0.801. It should be noted that components with similar names (overbite and overjet, for example) are operationally defined differently in the PAR and DI systems and cannot be considered as 'objectively' equivalent. Correlations between components with similar names but different operational definitions have only modest or even weak correlations; Overjet (r = 0.590), Overbite (r = 0.744), Occlusion relationship (r = 0.144), and Displacement (r = 0.582).

Differential prediction across Angle classifications

Significantly differentiated predictive accuracy was revealed among Angle Class subgroups for unweighted PAR index, with the least predictive power being captured in the Angle Class II subgroup. With the customized weighting of PAR index derived from this sample, superior predictive accuracy was found in Class II and Class III groups (r = 0.736 and r = 0.758) when compared with that of the Class I group (r = 0.652) (Table 3). A different pattern of components entered the weighting formulas for each Angle Class subgroup. In particular, there was no weight assigned to centreline discrepancy in Class I or Class III, to buccal occlusion in Class II, or to overbite in Class I (Table 1).

Similarly, both identified components and accuracy of predicting clinical judgment differed across Angle classifications when using the DI, as seen in Table 3. The highest correlation was found in the Class II group (r = 0.890) and the lowest in the Class I group (r = 0.736), with the Class III group being intermediate (r = 0.785). Only overjet and 'Other' components were related to the clinical judgment of malocclusion in all Angle classifications. A unique pattern of other components differentiated each of the Angle classifications (Table 2). Generally, properly weighted DI was a better measurement system than PAR index for identifying malocclusion among various Angle classification types, especially for Class II, where the difference is statistically significant at $P < 0.001$.

Measuring sampling bias

The group with the highest observed predictive accuracy (DI for Class II patients, r = 0.890) was chosen for testing as likely representing the largest positive sampling bias. The Monte Carlo correlation value for Class II cases using the DI was reduced from r = 0.890 to r = 0.864, a notable but statistically insignificant shrinkage. The Monte Carlo simulation has the added advantage of testing the configuration of components reaching the $P < 0.100$ threshold for inclusion as predictor components. In every randomly generated model, overjet, lateral open bite, either occlusal relation or lingual posterior crossbite, and 'Other' were picked as significant factors. Overbite, crowding, ANB, and L1-MP were never picked.

Discussion

The development of weighting formulas for PAR index and DI appropriate to a representative sample of Chinese patients illustrates how the multiple and interrelated sources of variance must be managed to move from the 'Presenting Condition' to 'Clinical Judgment'. 'Structured Assessment' aids, such as PAR index and DI, can help to identify the extent of malocclusion (increase the *r* value). But, it is equally apparent that predictive accuracy can be further improved by making adjustments for the target population and the use structured assessment is intended to play.

Sensitivity to Sample

When culturally appropriate weights were developed for both the PAR index and DI in this study of a Chinese population, weak to moderate predictions from the indices were elevated into the $r = 0.800$ range and higher, which is consistent with the hypothesis of this investigation. This degree of predictive accuracy was comparable to other studies reporting weightings appropriate to their contexts (4, 6, 7). But, the weights of components changed across populations (Table 4). Changes of weights in the present study support the criticism by Hamdan, *et al.* that the UK weighting system excessively emphasizes overjet and insufficiently weights overbite (13). In this study, not only did a culturally appropriate weighting system improve prediction, it outperformed weighting systems developed for other population groups (Table 5). In other words, borrowing weightings from other countries was not as successful as developing new population-specific ones. A similar example can be found in customized weighting procedures of the Index of Complexity, Outcome and Need in Chinese orthodontic patients (9). Thus, population sampling and ethnic esthetic norms were found to affect the usefulness of assessment systems for malocclusion.

Sensitivity to Diagnostic records

The type of records used in the presenting condition also plays a role when using indices. Han, *et al.* reported that study models alone provided adequate information for treatment planning in about 55 per cent of their Class II patients (14). A complete set of records was available to clinical judges in this study, including study casts, lateral cephalometric radiographs, panoramic radiographs, facial photographs, and treatment charts. Where measures are constrained to include only similar kinds of records, (alternatively, where relevant data are not used), the correlations are subject to common-method bias (15). This may explain the finding that predictions made from unweighted PAR were less than $r = 0.500$. The DI, which included cephalometric data, was a better predictor, even when unweighted. The 'Other' items category in the DI was weighted in all three Angle Class subgroups and the whole sample. The 'Other' items include supernumerary and congenitally missing teeth and impactions that are not accounted for in the PAR index (16, 17). This discrepancy in the record set from 'Presenting Condition' to 'Professional Judgment' was partially overcome by appropriate weighting, as demonstrated by the improvement of predictive accuracy after the PAR index was properly weighted, further confirming the major hypothesis in this study.

Sensitivity to Angle classification

The differential weighting of both PAR index and DI with respect to Angle classification of subjects illustrates that structured measurement systems are enhanced when adjusted for the context in which they will be used (see Tables 1 and 2). This suggests that the overall weighting generated for the whole sample would be less accurate than the customized weightings for each Angle Class subgroup. However, one can argue that the weighting formulas for the three Angle Class subgroups are less nuanced since they exclude low-level components that did not make the statistical cut due to a smaller sample used in developing the weighting formulas.

Table 4. Comparison of weightings for the components of the Peer Assessment Rating (PAR) index for malocclusion severity

Component	Present study (2017, CH)	Vlaskalic, V (1994, AU) (7)	DeGuzman, *et al.* (1995, US) (6)	Richmond, *et al.* (1992, UK) (4)*				
				All	Consultant	Specialist	G.D.P	Community
Overjet	3	5	5	6	5	4	5	4
Overbite	3.5	2	3	2	3	1	2	2
Centerline	1.5	2	3	4	6	–0.54	4	–0.25
R and L buccal occlusion	1.5	2	2	1	1	1	1	1
Upper anterior displacement	0	3	1	1	1	2	1	1
Lower anterior displacement	0	0	0	1	1	1	1	1

*Dentists of various groups carrying out orthodontic treatment in England and Wales were enrolled in Richmond, *et al.*'s study: All refers to considering them as a whole group; Consultant refers to Consultant orthodontists; Specialist refers to Specialist practitioner possessing specialist orthodontic qualifications; G.D.P. refers to General Dental Practitioners; and Community refers to Community dentists without orthodontic qualifications.

Table 5. Comparison of correlation coefficients (*r*) between the unweighted and weighted Peer Assessment Rating (PAR) indices

PAR weighting system	Present study (2017, CH)	Vlaskalic, V (1994, AU) (7)	Richmond, *et al.* (1992, UK) (4)	DeGuzman, *et al.* (1995, US) (6)
Unweighted PAR	0.431	0.74	0.74	0.69
AU-weighted PAR	0.554	0.80		
UK-weighted PAR	0.647		0.85	
US-weighted PAR	0.701			0.83
CH-weighted PAR	0.787			

Both indices exhibited the weakest predictive power for subjects in Class I. Besides the nature of Angle Class I malocclusions presenting with various occlusal traits such as severe crowding, impaction, and others (18, 19), another possible factor may be bimaxillary protrusion, which is more common in the Asian population than in the Caucasian population (20). It is possible that bimaxillary protrusion was considered a more severe form of malocclusion by Chinese orthodontists than what was captured by the unweighted indices (21–24).

Sensitivity to Sampling bias

The psychometric issues in this kind of study, such as internal consistency and sampling bias, drew the researchers' attention. Internal consistency of both clinical judgment and PAR or DI indices are required to demonstrate a high level of association among them. The popular, but somewhat incomplete, saying 'reliability is a precondition for validity' is applicable. It is possible to estimate the upper limit of association between clinical judgment and indices used in this study by applying the appropriate attenuation formula (25). An r values of 0.960, not 1.000, is the maximum value possible given the consistency of the current data. Internal consistency also increases proportionally with the number of judges, despite constant consistency across particular cases. Indices suitable for use in research or epidemiological contexts may not be satisfactory for use by individual clinicians. In the case of a single clinician and a single rater using one or the other the Spearman–Brown formula projects that the maximum r value would fall to $r = 0.634$.

The potential for sampling bias is real, especially when this is compounded by using the same sample to develop weights for structured measurements and to gauge their utility. The use of Monte Carlo techniques provided an estimate of the likely extent of such bias. In the present study, it was noticeable, approaching statistical significance. The Monte Carlo technique has the additional advantage of confirming the details of the weighting structure.

Conclusions

This study demonstrated that inter-examiner reliability of judges in evaluating malocclusion severity was excellent and that either the PAR index or the DI can be reliably scored by raters in a new population of patients. Customized sets of weightings for the PAR index and DI were developed by incorporating Chinese clinician judgments of malocclusion severity for Chinese orthodontic patients. The results of this study suggest that both the new weightings for PAR index and DI provide good predictive accuracy, with 62 per cent and 73 per cent, respectively, of the total variance in the clinical judgment of malocclusion severity being explained by these weighted indices. The present study also revealed that there are different optimal weight distributions for different Angle classifications, suggesting that it may not be appropriate to use the same weighting formula for all malocclusion types. In addition, psychometric issues of internal consistency, sampling bias, record bias, and the match between instrument and context of application were addressed.

Supplementary Material

Supplementary data are available at *European Journal of Orthodontics* online.

Funding

Specific Research Project of Health Pro Bono Sector, Ministry of Health, China [200802056].

Acknowledgements

We are indebted to the 69 orthodontists who kindly gave their time and expertise in the subjective evaluation session. And, we appreciate the three residents' hard work.

Conflict of Interest

None to declare.

References

1. Cronbach, L.J., Gleser, G.C., Nanda, H. and Rajaratnam, N. (1972) *The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles*. John Wiley & Sons, New York, NY.
2. Shaw, W.C., Richmond, S., O'Brien, K.D., Brook, P. and Stephens, C.D. (1991) Quality control in orthodontics: indices of treatment need and treatment standards. *British Dental Journal*, 170, 107–112.
3. Meehl, P.E. (1954/2013) *Clinical versus Statistical Prediction: A Theoretical Analysis and a Review of the Evidence*. Echo Point Books & Media, [No city of publication].
4. Richmond, S., *et al.* (1992) The development of the PAR Index (Peer Assessment Rating): reliability and validity. *European Journal of Orthodontics*, 14, 125–139.
5. Cangialosi, T.J., *et al.* (2004) The ABO discrepancy index: a measure of case complexity. *American Journal of Orthodontics and Dentofacial Orthopedics*, 125, 270–278.
6. DeGuzman, L., Bahiraei, D., Vig, K.W., Vig, P.S., Weyant, R.J. and O'Brien K. (1995) The validation of the Peer Assessment Rating index for malocclusion severity and treatment difficulty. *American Journal of Orthodontics and Dentofacial Orthopedics*, 107, 172–176.
7. Vlaskalic, V. (1994) *Australian weightings for the PAR Index* [Master's thesis]. School of Dental Science, The University of Melbourne, Melbourne, Australia.
8. Reagin, K.B. (2006) *The American Board of Orthodontics Discrepancy Index: an evaluation of its validity* [Master's thesis]. School of Dentistry, University of Alabama at Birmingham, Birmingham, AL.
9. Liao, Z.Y., *et al.* (2012) Validity assessment and determination of the cutoff value for the Index of Complexity, Outcome and Need among 12-13 year-olds in Southern Chinese. *International Journal of Oral Science*, 4, 88–93.
10. Hardy, D., Cubas, Y. and Orellana, M. (2012) Prevalence of Angle class III malocclusion: a systematic review and meta-analysis. *Open Journal of Epidemiology*, 2, 75–82.
11. Song, G.Y., *et al.* (2013) Validation of the American Board of Orthodontics Objective Grading System for assessing the treatment outcomes of Chinese patients. *American Journal of Orthodontics and Dentofacial Orthopedics*, 144, 391–397.
12. Geisser, S. (1993) *Predictive Inference*. Chapman and Hall, New York, NY.
13. Hamdan, A.M. and Rock, W.P. 1999 An appraisal of the Peer Assessment Rating (PAR) Index and a suggested new weighting system. *European Journal of Orthodontics*, 21, 181–192.
14. Han, U.K., Vig, K.W., Weintraub, J.A., Vig, P.S. and Kowalski, C.J. (1991) Consistency of orthodontic treatment decisions relative to diagnostic records. *American Journal of Orthodontics and Dentofacial Orthopedics*, 100, 212–219.
15. Campbell, D.T. and Fiske, D.W. (1959) Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56, 81–105.
16. Fox, N.A. (1993) The first 100 cases: a personal audit of orthodontic treatment assessed by the PAR (peer assessment rating) index. *British Dental Journal*, 174, 290–297.
17. Daniels, C. and Richmond, S. (2000) The development of the index of complexity, outcome and need (ICON). *Journal of Orthodontics*, 27, 149–162.
18. Bonk, R.T. (1996) Case report: adult Class I, constricted arches, crowding and impacted cuspid. *Journal of General Orthodontics*, 7, 23–27.
19. Ritter, D.E. (2014) Class I malocclusion with anterior crossbite and severe crowding. *Dental Press Journal of Orthodontics*, 19, 115–125.

20. Solem, R.C., Marasco, R., Guiterrez-Pulido, L., Nielsen, I., Kim, S.H. and Nelson, G. (2013) Three-dimensional soft-tissue and hard-tissue changes in the treatment of bimaxillary protrusion. *American Journal of Orthodontics and Dentofacial Orthopedics*, 144, 218–228.
21. Soh, J., Chew, M.T. and Wong, H.B. (2005) Professional assessment of facial profile attractiveness. *American Journal of Orthodontics and Dentofacial Orthopedics*, 128, 201–205.
22. Pae, E.K., McKenna, G.A., Sheehan, T.J., Garcia, R., Kuhlberg, A. and Nanda, R. (2001) Role of lateral cephalograms in assessing severity and difficulty of orthodontic cases. *American Journal of Orthodontics and Dentofacial Orthopedics*, 120, 254–262.
23. Bills, D.A., Handelman, C.S. and BeGole, E.A. (2005) Bimaxillary dentoalveolar protrusion: traits and orthodontic correction. *Angle Orthodontist*, 75, 333–339.
24. Huang, Y.P. and Li, W.R. (2015) Correlation between objective and subjective evaluation of profile in bimaxillary protrusion patients after orthodontic treatment. *Angle Orthodontist*, 85, 690–698.
25. Spearman, C. (1904) The proof and measurement of association between two things. *American Journal of Psychology*, 15, 72–101.