AJO-DO

Validation of the American Board of Orthodontics Objective Grading System for assessing the treatment outcomes of Chinese patients

Guang-Ying Song,[a] Sheldon Baumrind,[b] Zhi-He Zhao,[c] Yin Ding,[d] Yu-Xing Bai,[e] Lin Wang,[f] Hong He,[g] Gang Shen,[h] Wei-Ran Li,[i] Wei-Zi Wu,[a] Chong Ren,[a] Xuan-Rong Weng,[a] Zhi Geng,[j] and Tian-Min Xu[i]
Beijing, Chengdu, Xi'an, Nanjing, Wuhan, and Shanghai, China, and San Francisco, Calif

Introduction: Orthodontics in China has developed rapidly, but there is no standard index of treatment outcomes. We assessed the validity of the American Board of Orthodontics Objective Grading System (ABO-OGS) for the classification of treatment outcomes in Chinese patients. **Methods:** We randomly selected 108 patients who completed treatment between July 2005 and September 2008 in 6 orthodontic treatment centers across China. Sixty-nine experienced Chinese orthodontists made subjective assessments of the end-of-treatment casts for each patient. Three examiners then used the ABO-OGS to measure the casts. Pearson correlation analysis and receiver operating characteristic curve analysis were conducted to evaluate the correspondence between the ABO-OGS cast measurements and the orthodontists' subjective assessments. **Results:** The average subjective grading scores were highly correlated with the ABO-OGS scores ($r = 0.7042$). Four of the 7 study cast components of the ABO-OGS score—occlusal relationship, overjet, interproximal contact, and alignment—were statistically significantly correlated with the judges' subjective assessments. Together, these 4 accounted for 58% of the variability in the average subjective grading scores. The ABO-OGS cutoff score for cases that the judges deemed satisfactory was 16 points; the corresponding cutoff score for cases that the judges considered acceptable was 21 points. **Conclusions:** The ABO-OGS is a valid index for the assessment of treatment outcomes in Chinese patients. By comparing the objective scores on this modification of the ABO-OGS with the mean subjective assessment of a panel of highly qualified Chinese orthodontists, a cutoff point for satisfactory treatment outcome was defined as 16 points or fewer, with scores of 16 to 21 points denoting less than satisfactory but still acceptable treatment. Cases that scored greater than 21 points were considered unacceptable. (Am J Orthod Dentofacial Orthop 2013;144:391-7)

Various orthodontic indexes that aim to assess orthodontic treatment outcomes objectively have been developed since the 1970s.[1] Derived from prior subjective evaluations by groups of authorities, objective rating or categorizing systems generally assign numeric scores and provide a threshold for evaluating successful treatment.[2-4] In 1994, the American Board of Orthodontics (ABO) began to develop its Objective Grading System (OGS) to standardize and increase the precision and reliability of

[a] Postgraduate student, Department of Orthodontics, School and Hospital of Stomatology, Peking University, Beijing, China.
[b] Professor, Department of Orthodontics, and director, Craniofacial Research Instrumentation Laboratory, School of Dentistry, University of the Pacific, San Francisco, Calif.
[c] Professor, Department of Orthodontics, and associate dean, West China School of Stomatology, Sichuan University, Chengdu, China.
[d] Professor and chair, Department of Orthodontics, School of Stomatology, Fourth Military Medical University, Xi'an, China.
[e] Professor, Department of Orthodontics, and dean, School of Stomatology, Capital Medical University, Beijing, China.
[f] Professor, Department of Orthodontics, and dean, Institute of Stomatology, Nanjing Medical University, Nanjing, China.
[g] Professor and chair, Department of Orthodontics School, Hospital of Stomatology, and Key Laboratory for Oral Biomedical Engineering, Wuhan University, Wuhan, China.
[h] Professor and chair, Department of Orthodontics of Orthodontics, School of Stomatology, Shanghai Jiao Tong University, Shanghai, China.
[i] Professor, Department of Orthodontics, School and Hospital of Stomatology, Peking University, Beijing, China.
[j] Professor and director, School of Mathematical Sciences, Peking University, Beijing, China.

All authors have completed and submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest, and none were reported.

Supported by the Specific Research Project of Health Pro Bono Sector, Ministry of Health, China.

Reprint requests to: Tian-Min Xu, Department of Orthodontics, School and Hospital of Stomatology, Peking University, 22 Zhongguancun South Street, Haidian District, Beijing, China; e-mail, tmxuortho@gmail.com.

Submitted, January 2013; revised and accepted, April 2013.
0889-5406/$36.00

http://dx.doi.org/10.1016/j.ajodo.2013.04.018

dental cast and panoramic radiograph measurements after treatment. This system was introduced in 1999 as a component of the examination to determine whether completed cases met the ABO standard.[3] The ABO-OGS is now widely accepted and has recently been renamed the Cast/Radiograph Evaluation tool by the ABO.[5]

As used by the ABO, the Cast/Radiograph Evaluation scores the results of objective measurements of the final study casts and radiographs of completed patients. The cast measurements are made using a physical measuring tool that has been devised based on evaluations by groups of experienced ABO examiners in previous tests. The casts are scored in 7 categories (alignment, marginal ridges, buccolingual inclinations, occlusal relationships, occlusal contacts, overjet, and interproximal contacts), and panoramic radiographs are scored according to the single category of root angulation.

In each category, points are scored characterizing discrepancies from a standard developed by the ABO. There is a limit to the total number of discrepancy points that can be scored against a case in each category. There is also a limit to the number of discrepancy points that can be scored against each tooth in each category. The ABO score for the case is calculated by summing the scores for the 8 categories. If fewer than 20 points are scored overall, the case is considered to meet the ABO standard. If 20 to 29 points are scored, then the standard of work is undetermined. If more than 30 points are scored, the case is considered unacceptable.[3] In a study that assessed how well the OGS measured the quality of treatment in a sample of adult orthodontic patients, the cutoff value for a case that met the ABO standard was 27 points when the score for root angulation was excluded.[6]

When cutoff values are determined by an aggregation of professional opinions, the diagnostic specificity and sensitivity of any index used for evaluation are optimized.[4] Thus, the validity of any orthodontic treatment index is influenced by local conditions of treatment and judging.[7,8] Hence, any objective index requires a comparison with subjective evaluations made by a group of experienced orthodontists in a specific geographic region to determine the optimal threshold for treatment standards in that region.

This consideration is particularly relevant in a region as large and diffuse as China. Orthodontics has developed rapidly in China over the past 20 to 30 years.[9] As the number of patients grows, it is important to evaluate the effectiveness of orthodontic treatment provided by the various orthodontic services. The aims of this study were to assess the validity of the ABO-OGS tool as an index of treatment outcomes in China and to investigate the optimum cutoff scores for the Chinese population.

MATERIAL AND METHODS

This article was based on a multicenter study joining 6 orthodontic treatment centers in different parts of China. The participants included the Peking University School of Stomatology, the West China College of Stomatology at Sichuan University, the School of Stomatology at the Fourth Military Medical University, the Beijing Stomatological Hospital and School of Stomatology at the Capital Medical University, the Stomatological Hospital at Nanjing Medical University, and the Hospital of Stomatology at Wuhan University. Each center collected the complete medical records of at least 300 patients who had completed treatment between July 2005 and September 2008. From the combined total of 2383 patients' records, a stratified random sample of 108 subjects was drawn and balanced to include 18 from each collaborating center, consisting of equal numbers of Angle Class I, Class II, and Class III subjects. This sample was then randomly allocated to produce 9 groups that contained 12 subjects each. Each group included 4 Class I subjects, 4 Class II subjects, and 4 Class III subjects. Seventy-two of the 108 patients were less than 18 years of age; the remaining 36 were 18 years or older. There were 30 male and 78 female subjects. All markings that could identify the patient, the clinician, and the treatment center of origin were removed from the casts.

A panel of judges was formed for making subjective assessments of the 108-patient sample. It consisted of 69 experienced orthodontic specialists recommended by the 6 participating treatment centers to represent the different districts of mainland China; they assessed the patients subjectively. The criteria for the inclusion of each judge were (1) more than 10 years of clinical experience in orthodontics, (2) an MS or a PhD degree in orthodontics or experience as a research supervisor of orthodontic postgraduates, and (3) an academic rank of associate professor or above. Thirty-eight judges were men, and 31 were women.

To standardize the responses of the judges, a pilot examination was conducted in each center. Each judge evaluated 4 groups of cases treated locally over a dedicated period of 2 days. Two to 4 months later, the entire sample of 108 cases was evaluated over a 3-day period by all judges gathered at 1 location in Beijing.

For each case, each judge was invited to examine the physical upper and lower study casts individually and in occlusion. For each group of records, 2 separate assessments were made. In the first

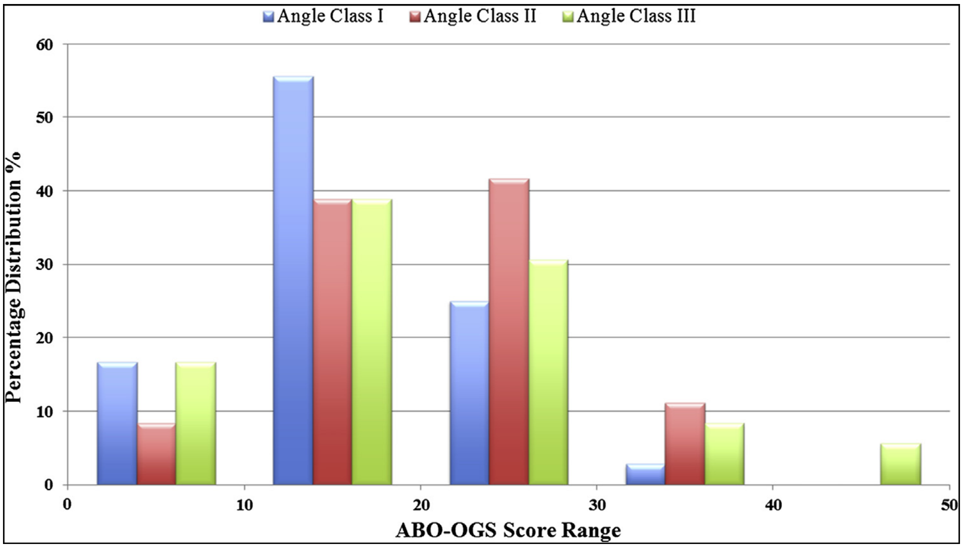

Fig 1. Distribution of the ABO-OGS scores of 108 cases according to Angle classification. Of the 108 cases that were assessed, 36 had Angle Class I, 36 had Angle Class II, and 36 had Angle Class III occlusal relationships.

assessment (ranking), each judge ranked and ordered the 12 study casts in each group numerically from 1 (most favorable) to 12 (least favorable) with respect to treatment outcome. In the second assessment (grading), the judge then identified the highest numerically ranked study casts in each group of 12 with a treatment outcome considered satisfactory. Then, beyond the highest numbered satisfactory casts, the judge identified the highest numbered casts considered acceptable. Cases with casts that had ranking numbers above the highest numbered acceptable casts were considered unacceptable. This procedure helped control for the chance aggregation of more or fewer acceptable cases in any group of 12 cases. Satisfactory cases were assigned a value of 1 point, acceptable cases were given 2 points, and unacceptable cases had 3 points. Over the entire sample, the cutoff points for satisfactory, acceptable, and unacceptable were based on the average scores of all 69 judges.

Under strict adherence to the ABO-OGS guidelines, 3 second-year postgraduate students (W.-Z.W., C.R., X.-R.W.) were invited to measure the study casts. They were asked to record the measurements in 7 ABO-OGS cast assessment categories. In the first session, a random set of 10 cases was measured by the 3 students for standardization. Four weeks later, each student assessed all 108 cases in the second session, including the 10 cases graded previously. Seven ABO-OGS categories were scored, and the grades of the 3 examiners were averaged.

Table I. Statistical analysis of the differences in ABO-OGS scores between Angle classifications, as assessed by 1-way ANOVA

Pretreatment Angle classification	n	*ABO-OGS scores Mean ± SD*	*F value*	P *value*
Class I	36	17.13 ± 6.21	1.585	0.210
Class II	36	20.56 ± 8.40		
Class III	36	19.53 ± 10.02		
Total	108	19.13 ± 8.40		

Statistical analysis

All statistical analyses were performed with SPSS software (version 20.0; SPSS, Chicago, Ill). Spearman correlation coefficients and kappa coefficients were calculated to assess the reliability between judges who undertook the subjective evaluations. Intraclass correlation coefficients (ICC) were computed to evaluate the intraexaminer and interexaminer reliabilities of the examiners who undertook the objective assessments. Stepwise linear regression and Pearson correlation analyses were conducted to assess validity. Receiver operating characteristic (ROC) curves were created to assess the sensitivity and specificity of the ABO-OGS tool and to determine the cutoff points for satisfactory, acceptable, and unacceptable cases. One-way analysis of variance (ANOVA) was used to determine whether the ABO-OGS scores differed systematically between Angle Class I, Class II, and Class III cases. Graphs were generated using MATLAB (R2011b; MathWorks, Natick,

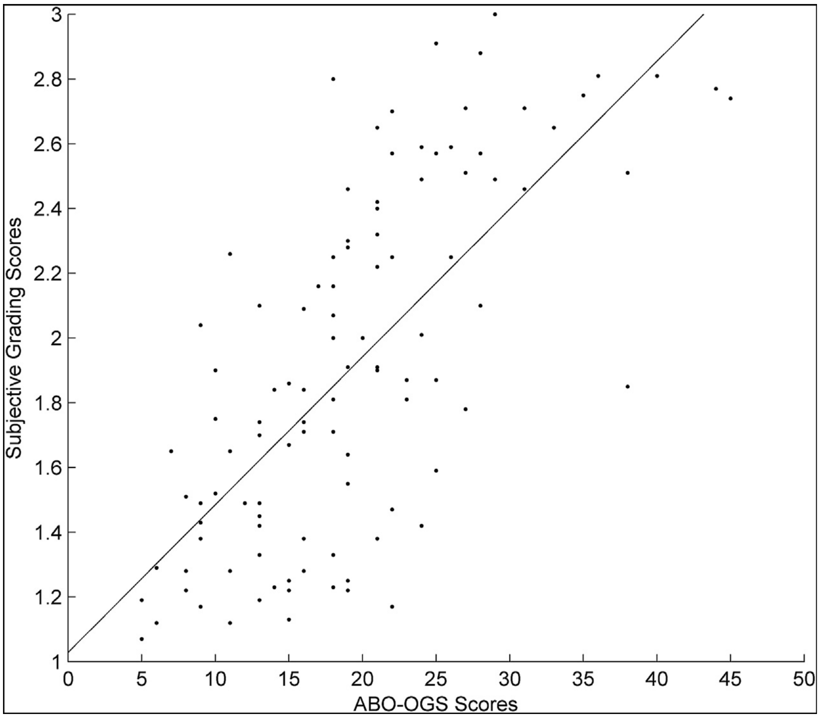

Fig 2. Scatter plot comparing the correlations between the subjective grading scores and the ABO-OGS scores. The correlation was *r* = 0.70 (*P* <0.05).

Table II. Summary of stepwise linear regression

Step	*Variable entered*	*Partial* R^2	*Model* R^2	P *value*	*Partial regression coefficient*	*Standardized partial regression coefficient*
1	X4	0.4291	0.4291	<0.0001	0.0766	0.5060
2	X6	0.0953	0.5244	0.0011	0.0595	0.2338
3	X7	0.0313	0.5557	0.0017	0.1524	0.2193
4	X1	0.0278	0.5835	0.0101	0.0426	0.1780

Regression equation: Y = 1.309 + 0.0766 * X4 + 0.0595 * X6 + 0.1524 * X7 + 0.0426 * X1.
X4, Occlusal relationships; *X6*, overjet; *X7*, interproximal contacts; *X1*, alignment.

Mass), Excel (Excel for Mac 2011; Microsoft, Redmond, Wash), and SPSS software.

RESULTS

The ABO-OGS scores of the 108 cases ranged from 5 to 45, with a mean value of 19.13 ± 8.40. The results of the 1-way ANOVA showed no statistically significant differences in the ABO-OGS scores between Class I, Class II, or Class III cases (Fig 1, Table I). The subjective grading scores of the 108 cases ranged from 1.07 to 3.00, with a mean value of 1.90 ± 0.54.

The mean value of the Spearman correlation coefficient was 0.64 ± 0.10 for all judge pairs of ranking score. The mean value of the kappa coefficient was 0.58 ± 0.06 for the subjective grading results of the 69 judges. An assessment of interexaminer reliability found that the ICC of the ABO-OGS scores of the 3 examiners was 0.74. For intraexaminer reliability, the ICC values of the ABO-OGS scores of the 3 examiners were 0.79, 0.81, and 0.77.

The average subjective grading scores correlated strongly with the ABO-OGS scores (r = 0.70, *P* <0.05; Fig 2). Validity testing selected the highly correlated categories and determined the weights of the components (Table II). Among the 7 categories, "occlusal relationship" was the first to enter into the regression equation, accounting for an R^2 value of 0.4291. "Overjet" entered next, adding an R^2 value of 0.0953. "Interproximal contacts" then added an R^2 value

Table III. Cutoff values for satisfactory and acceptable outcomes

Outcome	*Subjective grading score*	*ABO-OGS score*	*Sensitivity*	*Specificity*	*Kappa coefficient*	*AUC*
Satisfactory	1.51	16	0.82	0.72	0.53	0.84
Acceptable	2.16	21	0.83	0.81	0.61	0.89

AUC, Area under the receiver operating characteristic curve.

of 0.0313, followed finally by "alignment," which added a small but statistically significant increment of 0.0278. The overall R^2 value was 0.5835, implying that 58% of the variability in the average subjective grading scores was accounted for by the 4 categories of ABO-OGS scores.

The average grading scores of the subjective evaluation judging panel were used to determine the cutoff values for the ABO-OGS scores, according to a 3-grade outcome scale for the 33.3 and 66.7 percentiles on the mean outcome scale of 1.51 and 2.16 (Table III). The ROC curve analysis indicated that the cutoff values for satisfactory and acceptable outcomes were ABO-OGS scores of 16 and 21, respectively. The areas under the ROC curves values for these cutoffs were 0.84 and 0.89 (Figs 3 and 4).

DISCUSSION

In 2000, a study found that the incidence of malocclusion at the early permanent dentition stage among Chinese children was as high as 73%.[10] In 2008, it was reported that more than 300,000 Chinese patients received orthodontic treatment each year.[11] More than 2000 orthodontic specialists and thousands of general practitioners currently perform orthodontic treatments in China. However, few orthodontic assessment indexes have been validated for use in the large geographic area of China.

In this multicenter study, the sample was randomly selected from a large pool of 2383 patients from 6 parts of China. Their ages ranged from 12 to 35 years at the start of treatment. The 69 members of the subjective evaluation judging panel were recruited to represent all 6 treatment centers in the study. These experts had similar orthodontic training backgrounds and experiences. The interexaminer and intraexaminer levels of reliability were good for both ranking and grading; therefore, it would appear reasonable to consider the panel a homogeneous entity. Hence, we believe that our findings are widely applicable to the Chinese population, and the mean values of the subjective assessment reflect the gold standard.

In addition to being a clinical examination tool, the ABO-OGS has been used to increase the reliability, validity, and precision of the assessment of treatment

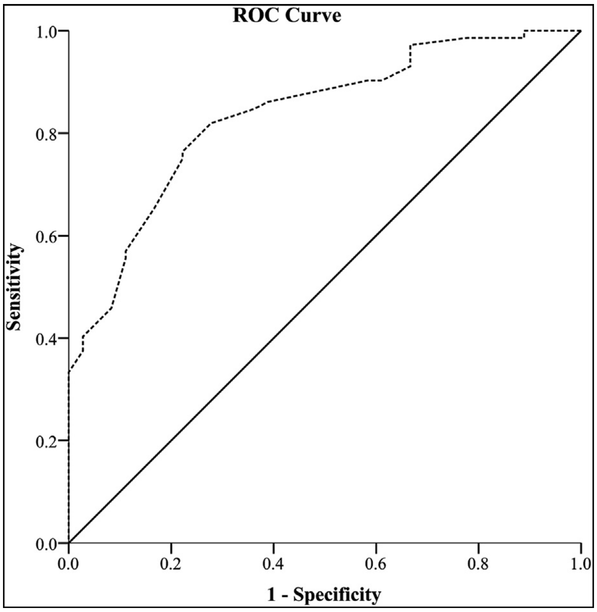

Fig 3. ROC curve showing the cutoff for a satisfactory outcome according to the ABO-OGS scores.

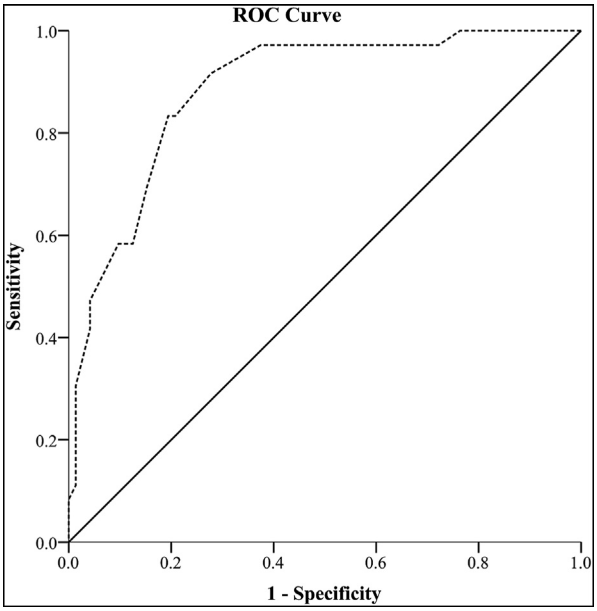

Fig 4. ROC curve showing the cutoff for an acceptable outcome according to the ABO-OGS scores.

outcomes in graduate programs.[12,13] The tool is also widely used in clinical studies to enable comparisons of outcomes achieved by different treatment modalities. Cook et al[14] used the ABO-OGS to compare university and private-practice orthodontic treatment outcomes and reported no significant differences between the overall scores. Hsieh et al[15] found that the ABO-OGS scores did not differ significantly between early and late treatment approaches. Kuncio et al[16] compared the postretention outcomes between Invisalign (Align Technology, Santa Clara, Calif) and traditional orthodontic treatments. The ABO-OGS scores showed that the patients treated with Invisalign had more relapse than those treated with conventional fixed appliances. In our study, we found no statistically significant differences between the ABO-OGS scores of patients with different pretreatment Angle classifications.

In previous studies, the reliability of the ABO-OGS was tested with parametric statistics (ICC)[12,13] or nonparametric statistics (Spearman rank coefficient, Wilcoxon, Kruskal-Wallis, and Mann-Whitney tests)[17,18] before its use. In our study, we included an opportunity for the postgraduate students to familiarize themselves with the ABO-OGS and to calibrate their scores. The 10 cases used in this part of the analysis were also assessed in the entire study sample; this enabled us to investigate intraexaminer reliability. We found relatively high interexaminer and intraexaminer agreement for the ABO-OGS panel. These outcomes are presumably related to the similar orthodontic training of the graduate students.

We assessed the validity of the scores with correlation analyses and ROC curve analysis. The subjective evaluation scores correlated well with the objective evaluation scores. Stepwise linear regression demonstrated that the components, when combined, complemented each other to predict the subjective perceptions of the judges. The best model included the scores for occlusal relationship, overjet, interproximal contact, and alignment. In a previous study in which the quality of treatment in adult orthodontic patients was assessed with the ABO-OGS, similar results were found.[6] Intercuspation was the only factor agreed upon by the 4 examiners as highly important in determining the quality of completed cases.

As a fundamental tool for the evaluation of diagnostic tests, ROC curve analysis has been used previously in orthodontic research to visualize and determine the optimal cutoff values for indexes of treatment outcomes and treatment need.[4,7] On the ROC curve, each point represents the sensitivity and specificity of different cutoff values in relation to a particular decision threshold.[19] The value of the area under the curve is between 0 and 1.0; values closer to 1.0 represent greater efficacy. We found that cutoff ABO-OGS scores of 16 points for satisfactory outcome and 21 points for acceptable outcome had good sensitivity, specificity, kappa values, and area-under-the-curve values. These findings suggest that this system has high validity for the classification of outcomes in Chinese patients. Thus, in China, we propose that cases with a total ABO-OGS score less than 16 should be deemed satisfactory, scores between 16 and 21 are acceptable, and scores greater than 21 are unacceptable. These cutoff values are lower than those currently recommended by the ABO. This difference might be attributable to the exclusion of the category of root angulation from the final model in this study, in addition to differences between the gold standards used in China and the United States. For the comprehensive assessment of treatment outcomes, the appropriateness of treatment plans and cephalometric measurements of skeletal, dental, and soft-tissue structures should also be taken into account.

With advances in digital technology, digital dental models are gradually replacing traditional plaster casts, as a result of limitations in storage, retrieval, transferability, durability, and remote diagnosis.[17] Many studies have confirmed the feasibility of the 3-dimensional measurement of digital casts.[20,21] The use of the ABO-OGS to make digital measurements is promising if the reliability and validity of these measurements are assessed.

CONCLUSIONS

Compared with the subjective evaluations of 69 experienced Chinese orthodontists, the objective ABO-OGS tool showed a high degree of validity as an index of treatment outcome in Chinese patients. The most important predictive components were occlusal relationship, overjet, interproximal contact, and alignment. With the root angulation score excluded, the cutoff value for satisfactory treatment outcome has been defined as a total ABO-OGS score of less than 16 points, with acceptable treatments having scores between 16 and 21 points. We believe that ABO-OGS scores greater than 21 points can indicate unacceptable treatment outcomes in a Chinese population.

ACKNOWLEDGMENTS

We thank the 69 orthodontists who participated in this study, which would not have been possible without their gracious cooperation. We also thank Edward L. Korn for his invaluable assistance in designing this study.

REFERENCES

1. Pickering EA, Vig P. The occlusal index to assess orthodontic treatment. Br J Orthod 1972;2:47-51.
2. Richmond S, Shaw WC, O'Brien KD, Buchanan IB, Jones R, Stephens CD, et al. The development of the PAR index (peer assessment rating): reliability and validity. Eur J Orthod 1992; 14:125-39.
3. Casko JS, Vaden JL, Kokich VG, Damone J, James RD, Cangialosi TJ, et al. Objective grading system for dental casts and panoramic radiographs. American Board of Orthodontics. Am J Orthod Dentofacial Orthop 1998;114:589-99.
4. Daniels C, Richmond S. The development of the index of complexity, outcome and need (ICON). J Orthod 2000;27: 149-62.
5. Greco PM, English JD, Briss BS, Jamieson SA, Kastrop MC, Castelein PT, et al. Posttreatment tooth movement: for better or for worse. Am J Orthod Dentofacial Orthop 2010;138:552-8.
6. Chaison ET, Liu X, Tuncay OC. The quality of treatment in the adult orthodontic patient as judged by orthodontists and measured by the objective grading system. Am J Orthod Dentofacial Orthop 2011;139(Supp):S69-75.
7. Liao ZY, Jian F, Long H, Lu Y, Wang Y, Yang Z, 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;4:88-93.
8. DeGuzman L, Bahiraei D, Vig KW, Vig PS, Weyant RJ, O'Brien K. The validation of the peer assessment rating index for malocclusion severity and treatment difficulty. Am J Orthod Dentofacial Orthop 1995;107:172-6.
9. Yeweng SJ, Huang SF, Ren LJ. Orthodontics in China. J Orthod 2002;29:62-5.
10. Fu MK, Zhang D, Wang BK, Deng Y, Wang FH, Ye XY. The prevalence of malocclusion in China—an investigation of 25,392 children. Zhonghua Kou Qiang Yi Xue Za Zhi 2002;37:371-3.
11. Lin JX, Xu TM. History and development of Chinese orthodontics. Beijing Da Xue Xue Bao 2008;18:11-4.
12. Pinskaya YB, Hsieh TJ, Roberts WE, Hartsfield JK. Comprehensive clinical evaluation as an outcome assessment for a graduate orthodontics program. Am J Orthod Dentofacial Orthop 2004; 126:533-43.
13. Knierim K, Roberts WE, Hartsfield JK. Assessing treatment outcomes for a graduate orthodontics program: follow-up study for the classes of 2001-2003. Am J Orthod Dentofacial Orthop 2006;130:648-55.
14. Cook DR, Harris EF, Vaden JL. Comparison of university and private-practice orthodontic treatment outcomes with the American Board of Orthodontics objective grading system. Am J Orthod Dentofacial Orthop 2005;127:707-12.
15. Hsieh TJ, Pinskaya YB, Roberts WE. Assessment of orthodontic treatment outcomes: early treatment versus late treatment. Angle Orthod 2005;75:162-70.
16. Kuncio D, Maganzini A, Shelton C, Freeman K. Invisalign and traditional orthodontic treatment postretention outcomes compared using the American Board of Orthodontics objective grading system. Angle Orthod 2007;75:864-9.
17. Okunami TR, Kusnoto B, BeGole E, Evans CA, Fadavi S. Assessing the American Board of Orthodontics objective grading system: digital vs plaster dental casts. Am J Orthod Dentofacial Orthop 2007;131:51-6.
18. Lieber WS, Carlson SK, Baumrind S, Poulton DR. Clinical use of the ABO-scoring index: reliability and subtraction frequency. Angle Orthod 2003;73:556-64.
19. Fawcett T. An introduction to ROC analysis. Pattern Recognition Lett 2006;27:861-74.
20. Costalos PA, Sarraf K, Cangialosi TJ, Efstratiadis S. 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;128:624-9.
21. Hildebrand JC, Palomo JM, Palomo L, Sivik M, Hans M. Evaluation of a software program for applying the American Board of Orthodontics objective grading system to digital casts. Am J Orthod Dentofacial Orthop 2008;133:283-9.