Original Article

Clinical Use of the ABO-Scoring Index: Reliability and Subtraction Frequency

William S. Lieber, DMD, MSD[a]; Sean K. Carlson, DMD, MS[b]; Sheldon Baumrind, DDS, MS[c]; Donald R. Poulton, DDS[d]

Abstract: This study tested the reliability and subtraction frequency of the study model–scoring system of the American Board of Orthodontists (ABO). We used a sample of 36 posttreatment study models that were selected randomly from six different orthodontic offices. Intrajudge and interjudge reliability was calculated using nonparametric statistics (Spearman rank coefficient, Wilcoxon, Kruskal-Wallis, and Mann-Whitney tests). We found differences ranging from 3 to 6 subtraction points (total score) for intrajudge scoring between two sessions. For overall total ABO score, the average correlation was .77. Intrajudge correlation was greatest for occlusal relationships and least for interproximal contacts. Interjudge correlation for ABO score averaged $r = .85$. Correlation was greatest for buccolingual inclination and least for overjet. The data show that some judges, on average, were much more lenient than others and that this resulted in a range of total scores between 19.7 and 27.5. Most of the deductions were found in the buccal segments and most were related to the second molars. We present these findings in the context of clinicians preparing for the ABO phase III examination and for orthodontists in their ongoing evaluation of clinical results. (*Angle Orthod* 2003;73:556–564.)

Key Words: Study models; Scoring indices

INTRODUCTION

The evaluation of final study models is essential for orthodontists who are interested in improving their clinical results. To date there is no objective evaluation system that has been widely accepted for this purpose.

Although several model scoring indices exist[1–4] and have been employed by many investigators,[3,5–18] most were developed to score pretreatment study models, usually for the purpose of determining eligibility for orthodontic services in insurance plans. Such scoring indices usually fall short when measuring final study models due to the fact that they do not measure the slight deviations from an ideal occulation that are typically found in a finished case.

The recently introduced American Board of Orthodontists (ABO)–scoring index was designed specifically to critique final study models.[19] It is one of the most detailed indices in use. It consists of seven distinct model-scoring criteria and one panoramic radiographic criterion. The index focuses on posttreatment study models and is designed to overcome deficiencies in other indices. The ABO recommends that clinicians use their scoring index to evaluate cases before submitting them to the board. They propose that this will help clinicians determine whether their cases are likely to pass that part of the examination.

A detailed document describing the ABO-scoring index was published in 1998.[19] However, little has been published on its use. We set out to quantify how reliably four clinicians performed when using the index. We were also interested in determining where the greatest frequency of subtractions was found. We elected not to include the panoramic criterion in this investigation in order to focus solely on the study models. By reporting our answers to the above questions, we hope to provide the clinical orthodontist with a better understanding of the strengths and weaknesses of the ABO-scoring index and how best to use it. We also hope to provide the reader with an example of those criteria for which they are most likely to have the greatest number of subtractions.

MATERIALS AND METHODS

Thirty-six posttreatment study models were selected from six different orthodontic offices. These thirty-six models

[a] Associate Professor of Orthodontics, University of the Pacific.
[b] Assistant Professor of Orthodontics, University of the Pacific.
[c] Professor of Orthodontics, University of the Pacific; Clinical Professor of Orthodontics, University of Medicine and Dentistry of New Jersey; Professor Emeritus, University of California, San Francisco, Calif.
[d] Professor of Orthodontics, University of the Pacific.
Corresponding author: Sean K. Carlson, DMD, MS, Department of Orthodontics, School of Dentistry, University of the Pacific, 2155 Webster Street, San Francisco, CA 94115.
(e-mail: skc@speakeasy.net).

Accepted: December 2002. Submitted: October 2002.

TABLE 1. Sample Demographics

	Mean	SD	Number	Percent
Sex				
Male			9	25
Female			27	75
Patient Age	16.36	2.83		
Treatment time	2.32	0.44		
Extraction Pattern				
Nonextraction			24	50
Extraction			24	50
Angle Class				
Class I			24	50
Class II			24	50

TABLE 2. Demographic Data for all Judges

	Age	Sex	Years in Practice	Years in Teaching	ABO Certification Year
Judge 1	61	M	37	36	1991
Judge 2	33	M	4	4	N/A
Judge 3	60	M	27	26	1980
Judge 4	38	F	3	6	N/A
Average	48		17.75	18	

were a subgroup selected from a larger sample that was collected without conscious bias as part of a concurrent outcomes study and fit the purposes of this study well. It was an equal representation of different types of finished cases because it was stratified by sex, pretreatment Angle Class, and extraction pattern. The demographics of the sample are shown in Table 1. All study casts were trimmed and polished to have a similar appearance.

Four judges were selected from the faculty in the Orthodontic Department at the University of Pacific School of Dentistry. The selection was made on the basis of familiarity with the ABO-scoring index, willingness to participate, and their availability. Information about the four judges is shown in Table 2.

All four judges underwent a four-stage calibration procedure before the data acquisition phase of the study. Stage one involved study of the ABO article, discussion of the ABO procedures, and consultation with a past ABO president and key investigator of the ABO-scoring index, Dr Vincent Kokich. Stage two involved scoring three sets of final study casts selected from an independent sample. This set was also scored by Dr Kokich. Stage three involved discussion of the results of the first scoring session and comparison with the scores of Dr Kokich. Modifications were made to our scoring procedures before beginning stage four that involved the scoring of a second set of three casts. The standard deviation for total score among the four judges decreased from 6.99 to 4.40 after the second scoring session, and data acquisition was started soon thereafter.

The data acquisition phase consisted of two scoring sessions separated by four weeks. For each session, each judge scored all 36 models in a single sitting. Each judge worked independently of the others. Data was collected manually on a scoring sheet (Figure 1) and later transcribed into Microsoft Excel. The same ABO measurement tool was used by all judges (Figure 2). Transcription of the data was done twice by two separate assistants to check for transcription errors.

Statistics were done in Microsoft Excel using a statistical add-on package called WinSTAT. Because the resulting ABO-scoring index data was ordinal, nonparametric statistical tests were used. Descriptive statistics (namely mean, median, and mode) were used to outline the results.

Three calculations were made to assess intrajudge reliability. First, differences between the two scoring sessions were reported for each judge for each criterion. Second, the Spearman rank correlation coefficient was calculated for each judge to assess the degree of association between that judge's two sessions. Third, the Wilcoxon rank sum test was used to detect differences between the two scoring sessions for the four judges.

Four calculations were made to assess interjudge reliability. First, scored subtractions were reported for each judge for each criterion. Second, the Spearman rank correlation coefficient was calculated for each two-judge combination to assess their degree of association with each other. Third, the Kruskal-Wallis nonparametric analysis of variance was used to detect differences in scores between the four judges. And fourth, the Mann-Whitney *U*-test was used to determine which judges differed from each other. Note that when the Kruskal-Wallis statistic was not significant, we did not proceed to the Mann-Whitney test for that criterion.

Descriptive statistics were used to report the frequency of subtractions for each tooth in each criterion. All values were considered statistically significant when $P < .05$ except for the results of the Mann-Whitney test where a conservative Bonferroni correction was used. For these tests, values were considered statistically significant when $P < .008$.

RESULTS

Intrajudge reliability

Results of the statistical tests evaluating intrajudge reliability are shown in Table 3. For all judges, for each criterion, the difference between the two scoring sessions was between 1 and 2 subtraction points. For the total score, the differences ranged from 4 to 5.5 subtraction points. That is, for total ABO score, judge 1 differed from himself by an average of 4.03 subtractions. Judges 2, 3, and 4 differed from themselves by an average of 5.19, 4.17, and 5.50 subtractions, respectively. For all criteria and for all judges, correlation coefficient values (r) between the two scoring

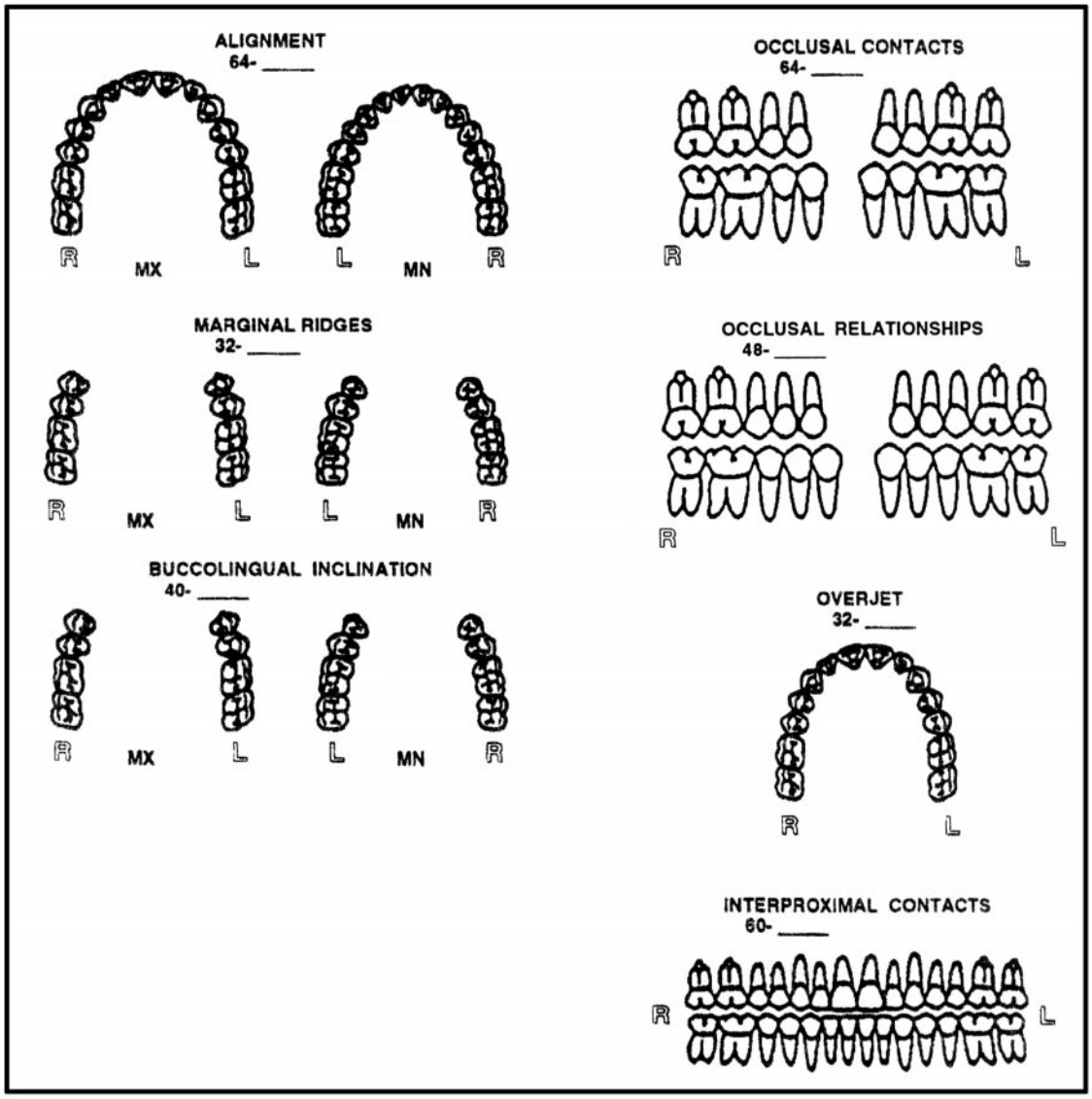

FIGURE 1. ABO-scoring index data collection sheet.

FIGURE 2. ABO measurement tool.

sessions were statistically significant except the interproximal contact comparison of judge 4. Although most of these correlations are statistically significant, the majority is of moderate strength (below .75). The strongest correlation for all judges between sessions 1 and 2 was observed in the occlusal relationship criterion (.83), whereas the weakest was observed in the interproximal contacts criterion (.52). For the overall ABO score, the average correlation value was .77.

Table 3 also shows the results of the Wilcoxon rank sum tests. There was at least one judge with a statistically significant difference between session 1 and session 2 for

TABLE 3. Results for all Statistical Tests of Intra-Judge Reliability. All tests Compared Session 1 Data Against Session 2 Data. Values were Considered Statistically Significant When $P < .05$

Intra-Judge Reliability		Differences Between Session 1 and Session 2				Spearman Rank Correlation		Wilcoxon Rank Sum Test	
		Mean	SD	Median	Mode	*r*	*P*	*Z*	*P*
Alignment	Judge 1	1.36	0.90	1	1	0.61	.00	−0.87	.19
	Judge 2	1.36	1.40	1	0	0.73	.00	−5.11	.01
	Judge 3	1.67	1.12	1	1	0.65	.00	−1.65	.05
	Judge 4	2.03	1.72	2	2	0.61	.00	−2.49	.01
						Average $r = .65$			
						Significant differences 3			
Marginal ridges	Judge 1	2.00	2.06	2	1	0.40	.01	−1.32	.09
	Judge 2	1.42	1.44	1	0	0.73	.00	−3.86	.00
	Judge 3	1.00	1.04	1	1	0.78	.00	−1.96	.02
	Judge 4	1.08	1.08	1	1	0.61	.00	−0.63	.26
						Average $r = .63$			
						Significant differences 2			
Buccolingual inclination	Judge 1	0.83	0.88	1	1	0.80	.00	−0.80	.21
	Judge 2	0.92	1.02	1	0	0.72	.00	−0.21	.42
	Judge 3	1.14	1.20	1	1	0.86	.00	−0.66	.26
	Judge 4	1.08	1.02	1	1	0.83	.00	−1.70	.05
						Average $r = .80$			
						Significant differences 1			
Occlusal relationship	Judge 1	1.06	1.04	1	1	0.87	.00	−0.86	.20
	Judge 2	2.00	1.88	2	2	0.84	.00	−3.83	.00
	Judge 3	1.11	1.06	1	1	0.87	.00	−1.09	.14
	Judge 4	1.75	2.70	1	1	0.74	.00	−0.07	.47
						Average $r = .83$			
						Significant differences 1			
Occlusal contacts	Judge 1	1.72	1.68	1	1	0.59	.00	−4.23	.00
	Judge 2	2.28	2.09	2	1	0.79	.00	−2.28	.01
	Judge 3	1.22	1.49	1	0	0.87	.00	−1.64	.05
	Judge 4	1.56	1.68	1	1	0.69	.00	−0.75	.23
						Average $r = .73$			
						Significant differences 2			
Overjet	Judge 1	1.22	1.27	1	1	0.59	.00	−0.34	.37
	Judge 2	1.56	1.52	1	1	0.63	.00	−1.69	.05
	Judge 3	1.72	1.58	1	1	0.64	.00	−0.62	.27
	Judge 4	1.53	1.38	1	1	0.70	.00	−2.08	.02
						Average $r = .64$			
						Significant differences 2			
Interproximal contacts	Judge 1	0.22	0.48	0	0	0.62	.00	−0.59	.28
	Judge 2	0.17	0.51	0	0	0.54	.00	−0.55	.29
	Judge 3	0.19	0.47	0	0	0.66	.00	−0.94	.17
	Judge 4	0.42	0.55	0	0	0.25	.07	−0.66	.25
						Average $r = .52$			
						Significant differences 0			
Total score	Judge 1	4.03	2.87	5	5	0.71	.00	−0.66	.25
	Judge 2	5.19	4.44	4	4	0.80	.00	−3.58	.00
	Judge 3	4.17	4.02	3	3	0.91	.00	−1.84	.03
	Judge 4	5.50	3.72	6	2	0.67	.00	−1.38	.08
						Average $r = .77$			
						Significant differences 2			

all criteria except interproximal contacts. The poorest agreement between sessions was found in the alignment criterion. Three of the four judges did not agree with themselves when evaluating this criterion. The best agreement was found in the interproximal contact criterion. In this study, all the judges agreed with themselves about the two sessions. For the overall score, two of the four judges agreed with themselves about the two sessions, whereas the other two showed statistically significant differences.

Interjudge reliability

Results for interjudge reliability are shown in Table 4. The correlation coefficients (r) differed for each criterion. The highest average correlation coefficient was observed in the buccolingual inclination criterion ($r = .85$). The lowest correlation was observed in the overjet criterion ($r = .50$). Total score had an average correlation value of $r = .85$. Results of the Kruskal-Wallis statistical test are also shown in Table 4. Five of the eight tests performed revealed a statistically significant difference between the four judges. Three criteria (buccolingual inclination, overjet, and interproximal contacts) showed no statistically significant differences between the four judges. In these instances, we did not proceed to the Mann-Whitney test. For the Mann-Whitney test, two criteria (alignment and occlusal relationship) showed statistically significant differences between four of the six possible combinations of judges. These criteria showed the greatest disagreement. Regarding overall score, there were statistically significant differences for two of six judge combinations.

Frequency of errors

Figures 3 through 10 show the frequency of subtractions by tooth for all criteria. Figure 3 shows the frequency of alignment subtractions. There were slightly more subtractions made in the maxilla than in the mandible. Second bicuspids and second molars were also deducted more frequently than other teeth. Central incisors were the tooth type with the lowest subtraction frequency.

Marginal ridge subtraction frequency is shown in Figure 4. Subtractions between the first and second bicuspids were less frequent than those of more posterior contacts. Subtractions tended to be slightly more frequent in the maxilla.

Figure 5 shows the subtraction frequency for buccolingual inclination. Both the right and left maxillary and mandibular second molars had a strikingly higher subtraction frequency compared with other teeth. Very few subtractions were made for bicuspids. There appeared to be no obvious trend of maxillary over mandibular subtractions.

The frequency of occlusal relationship subtractions is shown in Figure 6. The most anterior teeth (cuspids and bicuspids) showed the greatest frequency of subtractions. The dip in frequency at the first bicuspid site can be partly explained by the fact that 50% of the cases were treated with extractions. There was an interesting trend of all the left teeth having a slightly greater subtraction frequency than the right teeth.

Figure 7 shows the occlusal contact subtractions by tooth. The mandibular teeth had a significantly greater frequency of subtractions than did the maxillary teeth. For both jaws, there was a trend of a gradual increase in subtractions toward the posterior, with the second molars showing the greatest number of subtractions.

Overjet subtraction frequency is shown in Figure 8. The bicuspids and first molars had significantly less subtractions than did the other types of teeth. In fact, all other teeth had a similar frequency of subtractions. Again, the low first bicuspid frequency can be partly explained by the extraction patterns during treatment.

Figure 9 shows the frequency of interproximal contact subtractions by contact. Generally, there were very few subtractions for this criterion. Most of the subtractions were given because of spaces around the bicuspids. Very few spaces were found in the incisor area.

Figure 10 shows the total subtractions by tooth. This figure reflects the contribution of each tooth to the total amount of subtractions in the ABO-scoring procedure for this sample. Subtractions in the maxilla far outweighed those in the mandible. Subtractions tended to increase as we moved toward the posterior. Incisor teeth showed very few subtractions compared with posterior teeth. Maxillary second molars received the greatest amount of subtractions overall. Mandibular centrals received the fewest amount of subtractions.

DISCUSSION

In designing this experiment, we attempted to mimic what we felt the situation would be if a clinician sat down in his or her office to use the ABO index. However, our experiment was still quite ''controlled,'' and it did include a calibration session. However, even with these advantages, we discovered that both intrajudge and interjudge reliability were surprisingly low. Roughly half of the statistical tests performed showed significant differences between the two sessions and significant differences between the four judges. Some criteria showed better reliability than did others, and some judges showed better reliability than did others. But on the whole, there were far more statistically significant differences than expected.

Although statistically significant differences were found, this does not suggest that the ABO index is a poor index to use. A slightly different interpretation of the data reveals that it is still a very powerful index if its reliability is understood. Regarding intrajudge reliability, our data shows that all four judges subtracted differently over the entire index (total score) by roughly five points between sessions. That is, if one was scoring the same model twice on two different occasions, one might expect on average to subtract five more points during one of the sessions. If the total case score is near a clinician's target score (ie, an ABO passing score), one of the scoring sessions might put the case over the target score, whereas the other might keep it under. If the total case score is far from the target score, this difference matters less. On the basis of this observation, we recommend that those planning to use the ABO index as a measuring tool conduct their own assessment of reliability before starting to use the index fully. This will give them a better understanding of their own intrajudge reliability.

TABLE 4. Results for all Statistical Tests of Inter-Judge Reliability. Values were Considered Statistically Significant When $P < .05$

		Deductions (By Judge)				Spearman Rank Correlation (Judge vs Judge)					Kruskal-Wallis (All Judges)			Mann-Whitney (*P* values)			
Inter-Judge Reliability		Mean	SD	Median	Mode		Judge 1	Judge 2	Judge 3	Judge 4	*H*	*P*	Significance		Judge 1	Judge 2	Judge 3	Judge 4
Alignment	Judge 1	3.8	1.9	4	4	Judge 1		0.68	0.73	0.58	22.87	.00	S	Judge 1		0.00	0.35	0.48
	Judge 2	6.2	2.4	6	5	Judge 2			0.82	0.71				Judge 2			0.00	0.00
	Judge 3	4.1	2.2	4	3	Judge 3				0.62				Judge 3				0.24
	Judge 4	3.7	2.0	4	5	Judge 4								Judge 4				
									Average $r = .69$						Significant differences 3			
Marginal ridges	Judge 1	4.6	1.9	5	3	Judge 1		0.73	0.64	0.66	30.46	.00	S	Judge 1		0.28	0.00	0.00
	Judge 2	4.2	2.0	4	3	Judge 2			0.67	0.85				Judge 2			0.00	0.00
	Judge 3	2.5	1.8	3	0	Judge 3				0.54				Judge 3				0.46
	Judge 4	2.6	1.5	3	2	Judge 4								Judge 4				
									Average $r = .68$						Significant differences 4			
Buccolingual inclination	Judge 1	3.0	2.0	3	3	Judge 1		0.84	0.83	0.83	3.70	.30	NS	Judge 1				
	Judge 2	2.5	1.7	3	3	Judge 2			0.92	0.86				Judge 2				
	Judge 3	3.7	2.4	3	2	Judge 3				0.84				Judge 3				
	Judge 4	3.3	2.3	3	4	Judge 4								Judge 4				
									Average $r = .85$						Significant differences 0			
Occlusal relationship	Judge 1	3.9	3.0	4	0	Judge 1		0.74	0.81	0.73	16.68	.00	S	Judge 1		0.01	0.43	0.00
	Judge 2	5.9	3.4	6	4	Judge 2			0.84	0.86				Judge 2			0.00	0.17
	Judge 3	3.9	3.2	3	3	Judge 3				0.82				Judge 3				0.00
	Judge 4	6.8	3.9	7	7	Judge 4								Judge 4				
									Average $r = .80$						Significant differences 4			
Occlusal contacts	Judge 1	1.8	1.8	2	2	Judge 1		0.81	0.82	0.77	28.70	.00	S	Judge 1		0.00	0.00	0.00
	Judge 2	5.6	4.3	5	3	Judge 2			0.89	0.88				Judge 2			0.21	0.19
	Judge 3	4.9	4.3	4	2	Judge 3				0.84				Judge 3				0.46
	Judge 4	4.5	2.9	5	6	Judge 4								Judge 4				
		4.2							Average $r = .84$						Significant differences 3			
Overjet	Judge 1	2.3	1.6	2	2	Judge 1		0.48	0.34	0.46	7.34	.06	NS	Judge 1				
	Judge 2	3.0	2.0	3	5	Judge 2			0.53	0.54				Judge 2				
	Judge 3	3.6	2.2	4	2	Judge 3				0.62				Judge 3				
	Judge 4	3.5	2.3	4	5	Judge 4								Judge 4				
									Average $r = .50$						Significant differences 0			
Interproximal contacts	Judge 1	0.3	0.8	0	0	Judge 1		0.72	0.90	0.65	6.63	.08	NS	Judge 1				
	Judge 2	0.2	0.6	0	0	Judge 2			0.92	0.77				Judge 2				
	Judge 3	0.4	1.1	0	0	Judge 3				0.75				Judge 3				
	Judge 4	0.4	0.5	0	0	Judge 4								Judge 4				
									Average $r = .79$						Significant differences 0			
Total Score	Judge 1	19.7	6.9	19	22	Judge 1		0.84	0.89	0.81	13.67	.00	S	Judge 1		0.00	0.09	0.00
	Judge 2	27.5	9.4	26	26	Judge 2			0.84	0.86				Judge 2			0.02	0.13
	Judge 3	23.1	9.8	22	37	Judge 3				0.84				Judge 3				0.17
	Judge 4	24.8	8.0	25	18	Judge 4								Judge 4				
									Average $r = .85$						Significant differences 2			

Regarding interjudge reliability, our data shows that some judges were on average much more lenient than others. For example, judge 1 had an average total score of 19.7 ± 6.9, whereas judge 2 had an average total score of 27.5 ± 9.4. This reveals that it is not only important to know the intrajudge reliability but also important to know that scoring will likely differ between individual judges. Therefore, before concluding that cases are of passing caliber, it is important to know where individual judges stand in the way of scoring leniency. One suggestion to assist new judg-

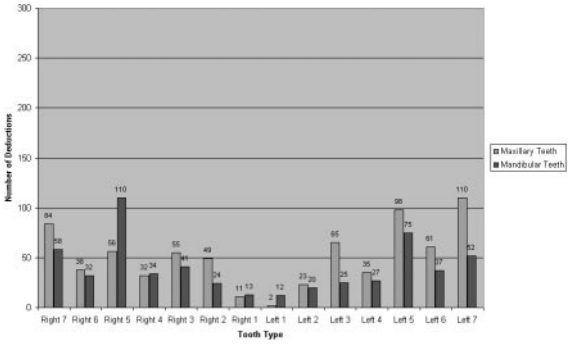

FIGURE 3. Alignment deductions by tooth.

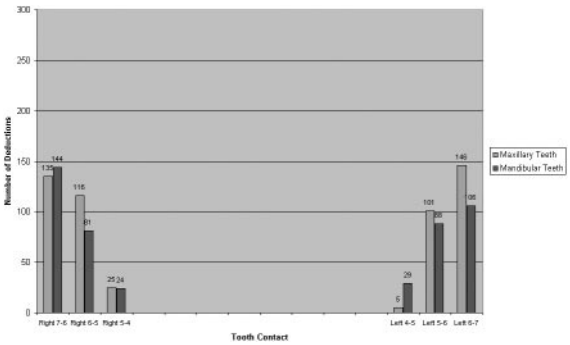

FIGURE 4. Marginal ridge deductions by contact.

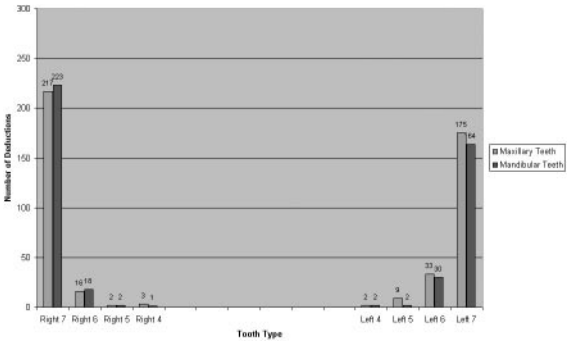

FIGURE 5. Buccolingual inclination deductions by tooth.

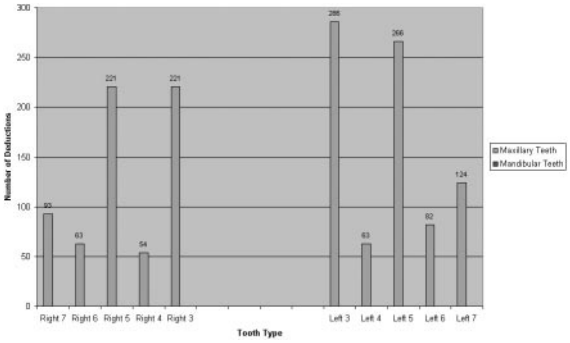

FIGURE 6. Occlusal relationship deductions by tooth.

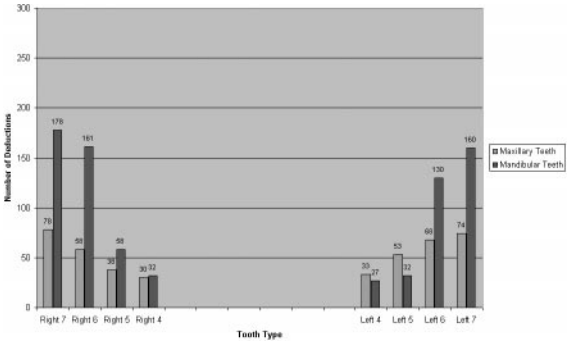

FIGURE 7. Occlusal contact deductions by tooth.

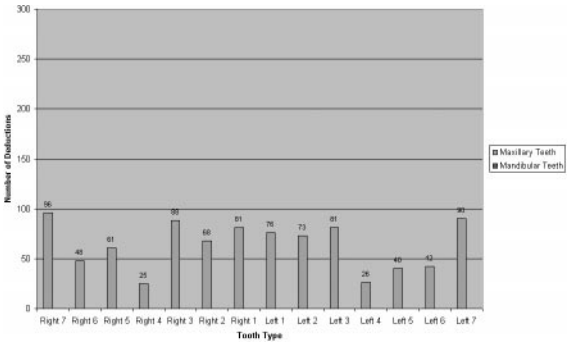

FIGURE 8. Overjet deductions by tooth.

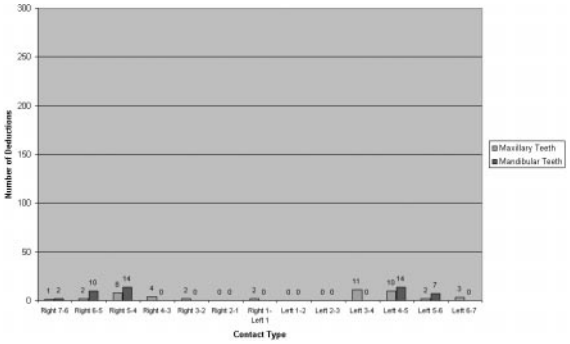

FIGURE 9. Interproximal contact deductions by contact.

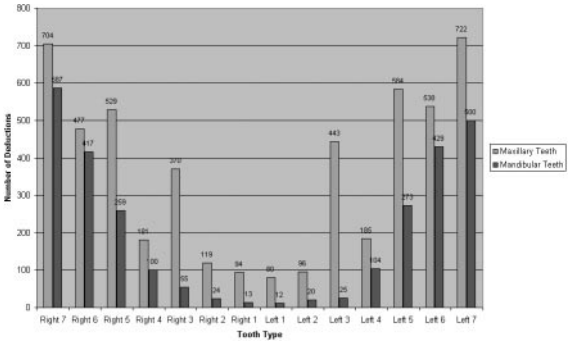

FIGURE 10. Total deductions by tooth.

es in determining where they stand is to have the ABO provide a calibration set of models that have been previously scored by Board members. This way, new judges can find out whether on average they score higher or lower than the Board members. This would certainly help prospective candidates prepare for the phase-III ABO examination.

Before starting our project, verbal discussions with Board members revealed that both their intrajudge and interjudge reliability was roughly $r = .85$. Our values were slightly lower for intrajudge reliability ($r = .77$) and the same for interjudge reliability ($r = .85$). One explanation for the lower score may be found in our calibration methods. Although we consulted with the ABO during our calibration process, our calibration methods were purposefully different than those of the ABO. Because we attempted to mimic as closely as possible how the average clinician would approach such a process, we chose to have less discussion among the prospective judges and less feedback from ''scoring experts'' than is usually available to the members of the Board. We feel our reliability values are representative of what one might encounter in his or her office while preparing to take the ABO examination or while examining posttreatment models.

Reliability in our study was generally lower than that found for the other indices. For the occlusal index, Summers[1] found interjudge and intrajudge reliability to be $r = .881$ and .903, respectively. Buchanan[20] studied reliability of the PAR index and found intrajudge reliability to be between $r = .95$ and .98. He also found interjudge reliability to be $r = .91$. DeGuzman et al[21] found a high intraexaminer reliability in his validation study of PAR and occlusal index of $r = .98$. Richmond found intrajudge correlations ranging between $r = .74$ and 1.00 within four examiners who measured a subset of his main sample, and interjudge correlations for the main sample of the unweighted PAR were $r = .91$ and of the weighted PAR $r = .93$.[4] Our results for intrajudge and interjudge reliability were $r = .77$ and .85, respectively. This is lower than all those mentioned except the lowest of Richmond's intrajudge range.

It is important for those pursuing the phase-III Board examination to understand how even their own reliability scores may be deceiving. Although one can repeat the grading over two sessions, and get similar, or even identical total scores, the correlation between the two scoring sessions can still be low. This is because the subtractions can come from different criteria. For example, judge 3 had a Spearman rank correlation coefficient of .91 for total scores (quite high) but was found to have statistically significant different total scores between sessions 1 and 2 ($P < .03$). However, judge 1 had a lower correlation coefficient of .71, but his total scores between sessions 1 and 2 were not different ($P = .25$). Who is the better scorer? Candidates should look at each criterion scored, not just total score, even though that is what the Board looks at as a pass/fail measure for a candidate.

We believe that the greatest limitation of the ABO index, in its current form, is its dependence on landmark identification. Most of the scoring involved measuring ''landmark-to-landmark'' linear distances using the ABO-scoring tool. This is not difficult, but when the judge's estimates of the landmarks differ, reliability suffers. We believe that by establishing better methods of obtaining the landmark data, we could greatly improve the reliability. For example, if automatic landmark location on digital models were incorporated, reliability would no longer be an issue. Currently, the ABO locates each landmark with pencil markings on the casts. These markings remain on the casts for all subsequent judges. We did not do this in our study. We avoided it in an attempt to represent a ''real world'' situation better. But it has become apparent that this step may be necessary for the time being to ensure better reliability between multiple judges.

Regarding subtraction frequency, our data clearly show that posterior teeth, especially second molars, are the sites of the most subtractions. They are also the sites with the greatest potential for subtractions. For example, a maxillary right second molar is scored eight times over six criteria, whereas a central incisor is scored only two times over two criteria. This means that one tooth's malalignment is often scored multiple times. That is, if a tooth is out of place, its poor placement gets evaluated from many perspectives and then gets deducted many times over multiple criteria. On the other hand, certain teeth may not be scored enough. For example, imagine a set of study models with an otherwise adequate occlusion but with four mandibular incisor teeth rotated 45°. If these incisors were the only problem, this case would still appear quite good when looking at the ABO index numbers alone, but it is clearly an unacceptable treatment. This exemplifies that there is an imbalance in the scoring criteria. Perhaps one solution to this ''over'' or ''under'' scoring of certain teeth would be to develop a weighting system similar to those previously developed for the PAR index analysis.[22]

Despite the ABO system's limitations, it is still clearly a step in the right direction. We feel that the seven measurement criteria are very appropriate and that they divide the model-scoring task into easily ''digestible'' sections. The index also provides a checklist for clinicians to use when evaluating their own cases. Such movements that bring objectivity to evaluation of the treatment outcomes should be encouraged. The ABO has provided a measurement system that adds a much needed objectivity to the analysis of final study models. Although our data give us the impression that this system is still highly subjective, future improvements and the appropriate use of the data presented in this study will render the ABO index highly valuable.

CONCLUSIONS

We have presented the results of intrajudge and interjudge reliability tests of the ABO-scoring index as well as subtraction frequency by tooth for each of the seven ABO model criterions. Reliability was lower than expected, suggesting that the ABO index may still be overly subjective. Subtraction frequency revealed a significant emphasis on second molars.

Methods that will assist prospective judges in using the ABO-scoring index are presented. Application of the data presented in this article is suggested to help clinicians achieve more accurate interpretations of their own ABO index scores. Suggestions are made for future improvements of the ABO-scoring index. Although this study revealed some current limitations of the index, the authors believe that such movements toward the objective analysis of treatment outcomes should be encouraged.

ACKNOWLEDGMENTS

The authors would like to thank Dr Roger Boero and Dr Vicki Vlaskalic for their participation in the study and also Dr Vincent Kockich for his generous guidance.

REFERENCES

1. Summers C. The occlusal index: a system for identifying and scoring occlusal disorders. *Am J Orthod.* 1971;59:552–567.
2. Andrews L. The six keys to normal occlusion. *Am J Orthod.* 1972;62:296–309.
3. Pickering E, Vig P. The occlusal index to assess orthodontic treatment. *Br J Orthod.* 1976;2:47–51.
4. Richmond S, Shaw W, O'Brien K, Buchanan I, Jones R, Stephens C. The development of the PAR index (Peer Assessment Rating): reliability and validity. *Eur J Orthod.* 1992;14:125–139.
5. Eismann D. Reliable assessment of morphological changes resulting from orthodontic treatment. *Eur J Orthod.* 1980;2:19–25.
6. Elderton R, Clark J. Orthodontic treatment in the general dental services assessed by the occlusal index. *Br J Orthod.* 1984;11:178–185.
7. Fox N. The first 100 cases: a personal audit of orthodontic treatment assessed by the PAR index. *Br Dent J.* 1993;174:290–297.
8. Birkeland K, Furevik J, Boe O, Wisth P. Evaluation of treatment and post-treatment changes by the PAR index. *Eur J Orthod.* 1997;19:279–288.
9. Buchanan I, Russell J, Clark J. Practical application of the PAR index: an illustrative comparison of the outcome of treatment using two fixed appliance techniques. *Br J Orthod.* 1996;23:351–357.
10. Chew M, Sandham A. Effectiveness and duration of two-arch fixed appliance treatment. *Aust Orthod J.* 2000;16:98–103.
11. Kerr W, McColl J. Use of the PAR index in assessing the effectiveness of removable orthodontic appliances. *Br J Orthod.* 1993;20:351–357.
12. Ngan P. Evaluation of treatment and post treatment changes of protraction facemask treatment using the PAR index. *Am J Orthod Dentofacial Orthop.* 2000;118:414–420.
13. Otuyemi O, Jones S. Long-term evaluation of treated Class II Division 1 malocclusions utilizing the PAR index. *Br J Orthod.* 1995;22:171–178.
14. Pangrazio-Kulbersh V, Kaczynski R, Shunock M. Early treatment outcome assessed by the PAR index. *Am J Orthod Dentofacial Orthop.* 1999;115:544–550.
15. Tang E, Wei S. Assessing treatment effectiveness of removable and fixed orthodontic appliances with the occlusal index. *Am J Orthod Dentofacial Orthop.* 1990;99:550–556.
16. Turbill E, Richmond S, Andrews M. A preliminary comparison of the DPB's grading of completed orthodontic cases with the PAR index. *Br J Orthod.* 1994;21:279–285.
17. Wijayaratne D, Harkess M, Herbison P. Functional appliance treatment assessed using the PAR index. *Aust Orthod J.* 2000;16:118–126.
18. Woods M, Lee D, Crawford E. Finishing occlusion, degree of stability and the PAR index. *Aust Orthod J.* 2000;16:15.
19. Casko J. Objective grading system for dental casts and panoramic radiographs. *Am J Orthod Dentofacial Orthop.* 1998;114:589–599.
20. Buchanan I, Shaw W, Richmond S, O'Brien K. A comparison of the reliability and validity of the PAR index and Summers' occlusal index. *Eur J Orthod.* 1993;15:27–31.
21. DeGuzman L, Bahiraei D, Vig K, Vig P, Weyant R, O'Brien K. The validity of the PAR index for malocclusion severity and treatment difficulty. *Am J Orthod Dentofacial Orthop.* 1995;107:172–176.
22. Hamdan A, Rock W. An appraisal of the PAR index and a suggested new weighting system. *Eur J Orthod.* 1999;21:181–192.