Archives of Orofacial Sciences
The Journal of the School of Dental Sciences
Universiti Sains Malaysia

Arch Orofac Sci (2014), **9**(1): 25-33.

Original Article

An investigation of inter-examiner reproducibility in recording malocclusion parameters during orthodontic epidemiologic studies

Grace Ang[a]*, Jacqueline Maryam Kamaluddin[a], Wizziyiane Ahmad [a], Uday Kumar Umesan[a], Siti Waznah Wahab [a], Lin Naing [b]

[a] Department of Dental Services, Ministry of Health, Bandar Seri Begawan, Brunei Darussalam.
[b] Universiti Brunei Darussalam, Bandar Seri Begawan, Brunei Darussalam.

* Corresponding author: angjess@yahoo.com

Revised edition: 19/02/2014. Accepted: 15/05/2014. Published online: 15/05/2014.

Abstract This study assesses inter-examiner reproducibility in recording various malocclusion parameters and Index of Orthodontic Treatment Need (IOTN) grade during patient examination by utilising the kappa statistic. Five previously calibrated orthodontists clinically examined 233 non-orthodontically treated schoolchildren aged 14-17 years for recording various malocclusion parameters. The examination was repeated twice, thirty days apart and precluded the use of study-models or radiographs. Although good inter-examiner reproducibility was observed in recording incisor class, IOTN dental health grade, type of posterior crossbite, and excellent for parameters with absolute criteria like erupted supernumeraries, etc, substantial examiner variation resulted in only fair reproducibility for recording IOTN esthetic category, canine class, overbite category, traumatic overbite and upper centre-line shift of two millimetres or more from the facial midline. Reproducibility for detecting occlusal displacement in the presence of crossbite was poor, and kappa statistic was incalculable for recording openbite and number of upper incisors rotated 30° or more. Kappa was also incalculable for recording IOTN dental health subcategory due to the creation of asymmetric tables caused by rarely chosen subcategory options. Despite prior agreement between previously calibrated examiners on evaluation criteria, detection of certain malocclusion parameters during an epidemiological examination can prove to be challenging. Epidemiological studies that report on prevalence of malocclusion in the population should always report on the kappa reproducibility, especially if the study is carried out by multiple examiners.

Keywords: kappa statistic, multiple examiners, malocclusion.

Introduction

Determining the prevalence of malocclusion and orthodontic treatment need for a population is essential in planning public health services. Any index or measurement used to quantify criteria within a population should show good inter-examiner reproducibility to produce valid results representative of that population, and to facilitate meaningful comparisons with similar epidemiological studies.

Many a time, reproducibility is not reported in studies apart from reassurance that examiner calibration had been performed (Gábris *et al.*, 2006; Perillo *et al*., 2010). In instances where reproducibility has been reported upon; percentage agreements (Du *et al.*, 1998; Svedström-Oristo *et al*., 2002) or even correlations (Onyeaso, 2004) have been used. Among various measures used to assess inter-examiner reproducibility, the kappa statistic has been recommended as it takes into account agreement due to chance (Hunt, 1986).

Previous studies of malocclusion or orthodontic treatment need have focused on Index of Orthodontic Treatment Need (IOTN), and have reported excellent kappa reproducibility (Abdullah and Rock, 2001; Ngom *et al*., 2007; Soh *et al*., 2005) when assessed by a single examiner. Pair *et al*. (2001) found good inter-examiner reproducibility amongst thirty orthodontists when assessing accurately trimmed study casts using standardized definitions for diagnostic subcategories.

Imprecision of recording malocclusion parameters in epidemiologic studies by multiple examiners was raised by Keeling *et al*. (1996), who suggested that differing population results may be a reflection of examiner discrepancies rather than actual population differences. Although it is ideal to assess malocclusion under optimal clinical conditions or from accurate study models; when large samples are required to represent a population, logistics and resources sometimes dictate that clinical assessments be made in-the-field employing multiple examiners, without using study models.

This study was designed to assess inter-examiner reproducibility in recording malocclusion parameters and IOTN in-the-field using the kappa statistic.

Materials and methods

A cross-sectional, observational, prospective study was planned to evaluate the reproducibility of recording malocclusion parameters between multiple examiners in a selected sample. The sample subjects constituted 248 non-orthodontically treated school children with ages ranging from 14 to 17 years derived from ten classes. The students were selected from a randomly chosen secondary school in Brunei Darussalam. After preliminary agreement on diagnostic criteria to assess malocclusion, five public service orthodontists were calibrated against each other twice, thirty days apart, on a set of thirty randomly chosen study casts. Matters of disagreement in recording malocclusion parameters were discussed and consensus arrived at which defined the final diagnostic criteria for use in this study (Table 1).

The five examiners (1 to 5) were then randomly paired utilizing the random number generator of Microsoft Excel to generate five examiner-pair combinations: 1&3, 1&5, 2&4, 2&3 and 4&5. Random pairing of examiners ensured that each examiner will ultimately be related to the other four, whether this is by a direct pairing or indirectly by the others. In this manner all the examiners are eventually compared with the others (Bianchi *et al.*, 2003). Two classes were randomly assigned to each pair, where each examiner of the pair, on the same day but independently of the other, conducted the first examination (T1). A repeat examination was conducted by the same examiner pair on the same subjects a month later (T2). This ensured that each pair examined a minimum of 30 subjects, which was deemed adequate for calculation of examiner reproducibility using the kappa statistic (Keeling *et al*., 1996).

Examinations were performed in a large multi-purpose hall with good lighting while the subject was seated on a chair with the head resting comfortably against a cushion placed on a high table-top behind. The examiner stood either directly in front or slightly to the side of the subject with a head mounted LED lamp providing better illumination of the intra-oral areas. Examination was effected primarily by direct vision while indirect vision of the upper teeth was facilitated by use of a disposable mouth mirror. The subject was asked to swallow and bite to record details in centric occlusion. Where there was clinical suspicion of a postured bite, attempts were made to reproduce the normal bite.

Disposable IOTN rulers (©Victoria Dental Hospital, Manchester) were made available during the examination, together with disposable acetate millimetre ruler portions cut to 30 mm lengths. No study casts or radiographs were used. The clinical examination and recording onto pro-forma sheets took on average five minutes per subject.

Statistical analysis

Data was entered into the IBM SPSS Version 20.0 by a trained assistant. Inter-examiner reproducibility was assessed using the kappa statistic by comparing the results recorded by the examiners in each pair, at both T1 and T2. Intra-examiner reproducibility was assessed using the kappa statistic, by comparison of the results obtained by each examiner at T1 and T2. A modified interpretation according to Hunt (1986) was applied where kappa ≤ 0.4 was designated as 'poor reproducibility (agreement)'; 0.41-0.58 as 'fair reproducibility'; 0.59-0.75 as 'good reproducibility' and $\geq$0.76 as 'excellent reproducibility'.

Table 1 Diagnostic criteria used in study

Diagnostic criteria	Definition
Incisor Class I	The lower incisors occlude onto the middle portion of the palatal surface of the upper incisors or on the borderline of this middle portion
Incisor Class II	The lower incisors occlude posterior to the middle third of the palatal surface of the upper incisor:
Class II division I	Class II incisors with upper incisors proclined or of normal inclination and the overjet is more than 4 mm.
Class II division II	Class II incisors with at least one of the upper central incisors retroclined or a Class II incisor relation could not be classed as II div 1.
Incisor Class III	The lower incisors occlude anterior to the middle third of the palatal surface of the upper incisor. In a bimaxillary proclination, this can exist even in the presence of a positive overjet
Canine Class I	The upper canine tip occludes in the contact embrasure between the lower canine and first premolar.
Canine Class II	Upper canine tip occlusion is mesial to Class I
Canine Class III	Upper canine tip occlusion is distal to Class I
IOTN DHC and EC	According to Brook and Shaw (1989)
Spacing *(extra subcategory)*	Aligned but spaced dentition In the absence of radiographic evidence: a) Premolars and permanent upper canines not erupted or showing signs of eruption by age 14 years is assumed to be impacted. b) Absence of upper central incisors is assumed as impaction of the tooth in question if a history of traumatic loss cannot be elicited c) Absence of upper lateral or lower incisors, where orthodontic treatment is required for space closure or restorative replacement, is recorded as congenital absence if palpation of the alveolar area does not reveal any clinical signs of the incisor being unerupted or impacted.
Orthodontic treatment recommended	IOTN 5, 4; or 3 with EC score more than 5
Overjet	Antero-posterior position of upper central incisor relative to the lower central incisor, measurement to the nearest mm, parallel to the occlusal plane from the labial surface of upper incisor to labial surface of lower incisor. Maximum overjet is recorded.
Overbite	Amount of vertical overlap of upper incisor to lower incisor, whereby the lower incisor is divided into vertical thirds. Presence or absence of trauma is recorded.
Openbite	Absence of vertical overlap in occlusion either anteriorly or in lateral segments, record maximum as less than or equal to 4 mm.
Crossbite	Recorded as anterior or posterior, where at least one tooth is in transverse discrepancy in occlusion.
Type of posterior crossbite	Also recorded as bilateral or unilateral if posterior crossbite is present. Recorded as upper teeth in relation to lower (Buccal/ lingual or tendency)
Displacement + crossbite	If displacement on occluding occurs in the presence of anterior crossbite or a unilateral posterior crossbite.
Midline diastema	Refers to presence of an open contact between the upper central incisors and the severity is measured in millimetres.
Spacing	Recorded in sextants where there is absence of crowding and at least one open contact in the sextant, except for the upper anterior segment where there has to be more than one open contact present.
Crowding	General assessment of crowding and recorded in sextants as a total of 2 mm or more in each sextant.
Excluded teeth	Erupted tooth completely excluded from the line of the arch
Rotations	Recorded only for the upper incisors with $\geq$ 30° rotation from the line of the arch
Upper Center-line Shift	Observed from the frontal view. Shift of $\geq$ 2mm from the facial midline
Clinically absent teeth	Any missing permanent tooth expected to be present for that dental age group. For premolars, upper canines and second molars, absence of tooth in question is assumed if subject is more than 14 years old and tooth is not palpable
Anomalous teeth	Developmental abberation of teeth e.g. Peg laterals, Leong's premolars
Transposition	Ectopic eruption of a permanent tooth which has resulted in exchange of normal tooth position with the adjacent permanent tooth

Table 2 Inter- and Intra-examiner kappa statistic results for parameters recorded

Parameters recorded	Inter-examiner kappa[a]	Intra-examiner kappa[b]
	Mean (T1, T2)	Mean (Min., Max.)
Incisor class	0.67 (0.64, 0.70)	0.74 (0.68, 0.80)
Left canine class	0.42 (0.41, 0.43)	0.66 (0.62, 0.74)
Right canine class	0.43 (0.40, 0.44)	0.64 (0.50, 0.87)
IOTN DHC Grade	0.73 (0.72, 0.74)	0.76 (0.74, 0.81)
IOTN Esthetic category	0.54 (0.55, 0.53)	0.66 (0.56, 0.72)
Overjet category	0.69 (0.64, 0.74)	0.76 (0.68, 0.83)
Overbite category	0.52 (0.50, 0.54)	0.69 (0.55, 0.80)
Presence of traumatic overbite	0.50 (0.52, 0.49)	* (*, 1.00)
Presence of openbite	* (* , *)	* (*, 1.00)
Crossbite(anterior and/or posterior)	0.84 (0.83, 0.85)	0.91 (0.82, 0.92)
Occlusal displacement with crossbite	0.22 (0.24, 0.19)	* (*, 0.66)
Number of anterior teeth in crossbite	0.79 (0.76, 0.81)	* (*, 0.86)
Unilateral/bilateral posterior crossbite	0.76 (0.77, 0.75)	0.87 (0.79, 0.96)
Buccal/Lingual/Tendency posterior crossbite	0.74 (0.71, 0.76)	* (*, 0.92)
Upper incisor teeth rotated 30^0 or more	* (* , *)	* (*, 0.83)
Presence of midline diastema	0.86 (0.88, 0.84)	0.84 (0.72, 0.95)
Presence of spacing in sextants	0.94 (0.92, 0.95)	0.95 (0.86, 1.00)
Presence of crowding in sextants	0.62 (0.59, 0.64)	0.81 (0.73, 0.89)
Presence of completely excluded teeth	0.96 (0.95, 0.97)	0.97 (0.95, 1.00)
Upper centerline shift by 2mm or more	0.58 (0.59, 0.56)	0.70 (0.64, 0.85)
Detection of absent teeth	0.96 (0.95, 0.97)	0.96 (0.90, 1.00)
Detection of partially erupted impacted teeth	0.84 (0.82, 0.86)	0.97 (0.85, 1.00)
Detection of erupted supernumeraries	1.00 (1.00, 1.00)	1.00 (1.00,1.00)
Detection of anomalous teeth	0.76 (0.75, 0.77)	0.83 (0.66, 1.00)
Detection of transposed teeth	1.00 (1.00,1.00)	1.00 (1.00, 1.00)
Recommendation for orthodontic treatment	0.83 (0.80, 0.85)	0.86 (0.84, 0.91)

*Kappa was not calculable

[a]Inter-examiner kappa statistics means and of T1, T2

[b]Intra-examiner kappa statistics mean, minimum, and maximum of the five examiners

Ethical approval

Ethical approval was obtained from the Medical and Health Research and Ethics Committee, Ministry of Health, Brunei Darussalam, Ref: MHREC/MOH/201/1(1) dated 14th March 2011. Patient information sheets and consent forms for the study were distributed two weeks before the first examination.

Results

Absences resulted in only 233 subjects being examined at both T1 and T2; however the number has surpassed the minimum required subjects for inter- and intra-examiner analysis. Table 2 presents the mean, inter-examiner kappa statistics at T1 and T2; and the mean, minimum and maximum intra-examiner kappa statistic results of the five examiners for the parameters recorded.

As no cases of obvious facial asymmetry were detected, examiner reproducibility in positively recording this parameter could not be ascertained.

Kappa statistic for the individual IOTN dental health components (DHC) and esthetic components (EC) could not be calculated due to the large number of categories available and the occasional recording of rarely chosen categories, which created asymmetric tables. Thus kappa was calculated for IOTN DHC grade

(five categories) and three EC categories (1-4, 5-6 and 7-10).

Excellent inter- and intra-examiner reproducibility

The parameters where both mean inter- and intra-examiner reproducibility have 'excellent agreement' as judged by the kappa statistic, were: decision on whether orthodontic treatment was recommended; detecting anterior and/or posterior crossbites, unilateral or bilateral crossbites, midline diastemas, spacing in sextants, erupted supernumeraries and teeth that were clinically absent, completely excluded from the arch, partially erupted and impacted, anomalous, or were transposed.

More examiner variability

Inter-examiner reproducibility was also rated as excellent in recording the number of anterior teeth in crossbite. However, individual examiner reproducibility varied for this parameter, whereby two examiners showed good to excellent reproducibility but kappa statistic was not calculable for the other three examiners.

Good inter-examiner reproducibility was noted for recording incisor class, IOTN DHC grade, overjet category, crowding in sextants, and the type of posterior crossbite (buccal/ lingual/ tendency). Intra-examiner reproducibility had ranged from good to excellent for recording most of these parameters, with the exception for recording the type of posterior crossbite, where kappa statistic was incalculable for one examiner.

Inter-examiner reproducibility was only fair for recording canine class, IOTN EC category group, overbite category, upper centreline shift (2 mm or more) and the presence of traumatic overbite. Intra-examiner reproducibility for these parameters varied from fair (one to two examiners) to good-excellent for most of the examiners, with the exception for recording traumatic overbite, where kappa statistic was incalculable for one examiner.

Inter-examiner reproducibility was rated as poor in recording occlusal displacement in the presence of crossbite. For this parameter, there was much individual examiner variability where one examiner showed good reproducibility, two examiners had poor reproducibility and kappa statistic was incalculable for the remaining two examiners.

Inter-examiner kappa statistic was incalculable for detecting the presence of openbite and upper incisors rotated 30° or more. For these parameters, although one or two examiners showed good or even excellent reproducibility, the kappa statistic was incalculable for the majority of the examiners.

Discussion

Although it is not clear to what extent examination conditions, recording, transcription errors, and patient variability might affect results in prevalence reports; nevertheless, good reproducibility between multiple examiners using standardized diagnostic criteria is desirable to validate results obtained in-the-field, and to facilitate meaningful comparisons with other population studies. The kappa statistic has been recommended as the gold standard to assess reproducibility of results (Hunt, 1986).

IOTN DHC instance

Where there are numerous choices in classifying a parameter, if a category is rarely chosen in the first instance, the probability of choosing this rarely chosen category again, at on a second occasion can be quite low. In such instances kappa becomes incalculable due to an asymmetric table caused by occurrence of the rare observation (Du *et al*., 1998).

The IOTN DHC has five grades with component-categories giving more than 30 possible recordable choices. The Kappa statistic was rendered incalculable when these sub-categories were used (as asymmetric tables were created when one examiner recorded a rare subcategory but the same subcategory was never recorded by the other examiners - five cases in total); or, when the examiner opted a rare subcategory at T1 but failed to record the same subcategory at T2 (one to three cases). Such discrepancies occurred in mild

malocclusions with minimal to no contact point displacement. Due to the hierarchical nature of IOTN DHC, one examiner recorded a mild reverse overjet or, crossbite with discrepancy between retruded and intercuspal position, but the chosen subcategory was not opted during the second examination. Undetected patient posture and difficulties in reproducing retruded contact position under field conditions may have contributed to these discrepancies. Interestingly, when the rarely chosen observations were replaced with more commonly occurring subcategories (e.g. contact displacement), while still retaining them as disagreements, intra- and inter-examiner kappa for the individual IOTN DHC subcategories were calculable and ranged from good to excellent.

Some examiners (Abdullah and Rock, 2001; Ngom *et al*., 2007), in single examiner studies, had reported almost perfect intra-examiner reproducibility for recording IOTN DHC in-the-field. However, it was not clear whether this was based on IOTN DHC grades or individual subcategories.

In the present study, examiner calibrations had been performed on study casts taken from a stock of past referral cases and therefore representative of malocclusions usually seen in our clinics. In retrospect, perhaps, had very mild malocclusions also been included in the calibration exercise so as to represent the rarely chosen IOTN DHC subcategories, kappa statistic for IOTN DHC subcategories may have been calculable. However, certain IOTN DHC subcategories cannot be reproduced by study cast calibration e.g. displacements during dynamic occlusion.

For in-the-field situation, inter- and intra-examiner reproducibility for IOTN DHC grades in the present study were good to excellent and comparable to those reported by Brook and Shaw (1989) and Souames *et al*. (2006). In another study, a calibrated single examiner had reported good but not excellent kappa results for IOTN DHC grades, even after he had re-recorded his in-the-field IOTN DHC results with reference to study casts when the in-the-field recording did not match that of the study casts (Abu Alhaija *et al*., 2004).

Where differences in recording IOTN DHC grades existed, these were often differences in recording severity of contacts displacement or overjet (affecting borderline choices), or a failure to detect a worse IOTN DHC grade in the more posterior regions of the mouth. Similar findings regarding disagreements in borderline decisions were reported by Buchanan *et al*. (1994) even for observations made under ideal clinical conditions by well calibrated examiners.

Burden *et al*. (2001) had questioned the benefit of recording IOTN DHC subcategories, as this does not reflect treatment complexity; however, data on IOTN DHC grades can be beneficial for the purposes of deciding on orthodontic treatment need and planning of services.

IOTN EC

Kappa statistic for IOTN EC utilizing the 10 subcategories was incalculable due to the creation of asymmetric tables, again, caused by a rarely chosen subcategory which was never utilized at the second observation. When the observations were categorized into three practically relevant categories (EC 1-4 as 'no need for treatment', 5-6 as 'borderline' and 7-10 as 'treatment advised'), intra-examiner reproducibility ranged from fair to good, with inter-examiner reproducibility for this parameter being only fair, reflecting the more subjective nature of the assessment.

Incisor and buccal segment relation

The incisor morphology in our ethnic Malay population often lacks a well-defined cingulum plateau; thus the palatal surface was divided into thirds (Mills, 1981) to facilitate classification into four categories. This resulted in good reproducibility. Kappa for recording incisor relation has not been previously reported.

Although individual reproducibility for recording canine relations ranged from fair to excellent, inter-examiner reproducibility was only fair at best. This could be attributed to difficulties in ascertaining relations in crowded arches, viewing angle, projection of

worn canine tips, undetected patient postured bites and field conditions. Difficulty in obtaining unobstructed views of the molar areas resulted in a joint decision by the examiners to exclude recording molar relationship.

It is possible that our reproducibility for recording buccal segment relationship could have been improved, had the examination been carried out utilizing study casts or in a clinic setting. Although good reproducibility between study cast assessments and measurements obtained intra-orally had been reported by Ovsenik *et al*. (2004) for the single examiner, the authors did not show evidence of reproducibility for each type of assessment by repeating observations at different time intervals. In any case, surveying a large population sample utilizing study casts or in a clinic setting would render this option impractical.

Overjet, overbite, traumatic overbite and openbite

Inter-examiner reproducibility for overjet category was good and comparable with Burden (1995). One examiner showed fair reproducibility in recording overbite category while the others were good-excellent, resulting overall in only fair inter-examiner reproducibility. Subjective assessments under non-ideal conditions may have played a role (Svedström-Oristo *et al*., 2002); instead of measurements for overbite in terms of millimetre overlap, the lower incisor crown had to be mentally divided into thirds and individual perceptions in borderline cases could have differed.

Recording traumatic overbites could not be calibrated during the study cast exercises and was subject to individual examiner interpretation in-the-field. Although there were very few positive recordings of traumatic overbite and four examiners showed good to excellent reproducibility in this parameter, for one examiner, kappa was not calculable as the single positive finding of palatal trauma at one examination was not reproduced at all in the second. Inter-examiner kappa was only fair for this parameter. The few positive cases, possible undetected patient posture at different examination sessions (affecting intra-examiner result), differing interpretations of what constitutes traumatic overbite and field conditions may have affected reproducibility.

Open bite was recorded in terms of location – anterior, and/ or posterior, as well as severity. There were few positive cases recorded, which were all mild. For both intra- and inter-examiner kappa analysis, asymmetrical tables were created for this parameter, as a category chosen at one examination was not ever utilized at the next, rendering kappa incalculable. It is likely the few findings, coupled with undetected patient posture at different examination sessions affected reproducibility for this parameter.

Crossbites and displacement on occluding

Although inter- and intra-examiner reproducibility for detecting the presence and location of crossbites and whether posterior crossbite was unilateral or bilateral, were generally good-excellent, for some examiners there were inconsistencies in recording the number of anterior teeth involved in anterior crossbite and whether the posterior crossbite was buccal, lingual or in-tendency. However inter-examiner results were better. It is possible the intra-examiner inconsistencies were due to undetected patient posturing (at either T1 or T2), whilst the subjects may have been more consistent in occluding during the same session, giving more reproducible inter-examiner results.

Detection of occlusal displacements in the presence of crossbites showed poor inter-examiner reproducibility in-the-field with much individual examiner variability. This may be related to the few positive findings and difficulties associated with examinations in-the-field. Examiner interpretations may have also played a role as this parameter could not be calibrated during the study cast exercises.

Presence of rotated upper incisors

For three examiners kappa was incalculable for recording the number of upper incisors rotated 30° or more, with resultant incalculable inter-examiner results as asymmetric tables were created. That a figure chosen at one observation was not chosen at all at the next observation session could be related to the difficulty in ascertaining the

degree of rotation in borderline cases where adjacent teeth were also rotated and mal-aligned. Perhaps revision of the criteria to 45°, or, the use of a transparent reference guide (Ovsenik, 2007) could have improved examiner reproducibility.

Upper centre-line

Although intra-examiner reproducibility for detecting an upper centre-line shift of two millimetres or more from the facial midline was good to excellent, inter-examiner reproducibility was only borderline fair-good. Perhaps inter-examiner reproducibility could have been improved had the criterion for recording this parameter been set at four millimeters (Kokich *et al*., 1999).

Crowding and spacing in sextants

The criteria of assessing crowding or spacing in terms of presence in sextants instead of absolute measurement of these parameters per arch probably played a major role in the good reproducibility of recording these parameters in-the-field. Had measurement of total crowding or spacing been considered, reproducibility in-the-field would likely have decreased (Keeling *et al*., 1996).

Excellent agreement for parameters with absolute criteria and decision whether treatment is recommended

Not surprisingly, where the diagnostic criteria was easily detectable and recorded as either present or absent, with no measurements or estimations required, intra- and inter-examiner reproducibility was excellent to near perfect. Recommendation for orthodontic treatment was based on IOTN grade severity and examiner agreement on this parameter was excellent, which is reassuring for consistency in public sector orthodontic decision for treatment.

Conclusions

Despite prior agreement on criteria and good calibration, recording of malocclusion parameters in-the-field can still show much examiner inconsistencies and variations, even among orthodontists. Kappa reproducibility should always be reported even for previously calibrated examiners.

Acknowledgements

The authors wish to thank the Department of Dental Services, Ministry of Health Brunei Darussalam and the Ministry of Education for the support given in carrying out this project.

References

Abdullah MSB, Rock WP (2001). Assessment of orthodontic treatment need in 5,112 Malaysian children using the IOTN and DAI indices. *Community Dent Health*, **18**(4): 242-248.

Abu Alhaija ES, Al-Nimri KS, Al-Khateeb SN (2004). Orthodontic treatment need and demand in 12-14 year old north Jordanian school children. *Eur J Orthod*, **26**:261-263.

Bianchi LI, Gallagher EJ, Korte R, Ham HP (2003). Interexaminer agreement on the American Board of Emergency Medicine oral certification examination. *Ann Emerg Med*, **41**(6): 859-864.

Brook PH, Shaw WC (1989). The development of an index of orthodontic treatment priority. *Eur J Orthod*, **11**(3): 309-320.

Buchanan IB, Downing A, Stirrups DR (1994). A comparison of the Index of Orthodontic Treatment Need applied clinically and to diagnostic records. *Br J Orthod*, **21**(2):185-188.

Burden DJ (1995). An investigation of the association between overjet size, lip coverage, and traumatic injury to maxillary incisors. *Eur J Orthod*, **17**(6): 513-517.

Burden DJ, Pine CM, Burnside G (2001). Modified IOTN: an orthodontic treatment need index for use in oral health surveys. *Community Dent Oral Epidemiol,* **29**(3): 220-225.

Du SQ, Rinchuse DJ, Zullo TG, Rinchuse DJ (1998). Reliability of three methods of occlusion classification *Am J Orthod Dentofacial Orthop,* **113**(4): 463-470.

Gábris K, Márton S, Madléna M (2006). Prevalence of malocclusions in Hungarian adolescents. *Eur J Orthod*, **28**(5): 467-470.

Hunt RJ (1986). Percent agreement, Pearson's correlation, and kappa as measures of inter-examiner reliability. *J Dent Res*, **65**(2): 128-130.

Keeling SD, McGorray S, Wheeler TT, King GJ (1996). Imprecision in orthodontic diagnosis: reliability of clinical measures of malocclusion. *Angle Orthod*, **66**(5): 381-391.

Kokich VO, Kiyak HA, Shapiro PA (1999). Comparing the perception of dentists and lay people to altered dental esthetics. *J Esthet Dent*, **11**(6): 311-324.

Mills JRE (1981). *Principles and Practice of Orthodontics* 2nd edn. Edinburgh: Churchill Livingstone. p. 132.

Ngom PI, Diagne F, Dieye F, Diop-Ba K, Thiam F (2007). Orthodontic treatment need and demand in Senegalese school children aged 12-13 years. An appraisal using IOTN and ICON. *Angle Orthod*, **77**(2): 323-330.

Onyeaso CO (2004). Prevalence of malocclusion among adolescents in Ibadan, Nigeria. *Am J Orthod Dentofacial Orthop*, **126**(5): 604-607.

Ovsenik M (2007). Assessment of malocclusion in the permanent dentition: reliability of intraoral measurements. *Eur J Orthod*, **29**(6): 654-659.

Ovsenik M, Farcnik F, Verdenik I (2004). Comparison of intra-oral and study cast measurements in the assessment of malocclusion. *Eur J Orthod*, **26**(3): 273-277.

Pair JW, Luke L, White S, Atchinson K, Englehart R, Brennan R (2001). Variability of study cast assessment among orthodontists. *Am J Orthod Dentofacial Orthop*, **120**(6): 629-638.

Perillo L, Masucci C, Ferro F, Apicella D, Bacetti T (2010). Prevalence of orthodontic treatment need in Southern Italian schoolchildren. *Eur J Orthod*, **32**(1): 49-53.

Soh J, Sandham A, Chan YH (2005). Malocclusion severity in Asian men in relation to malocclusion type and orthodontic treatment need. *Am J Orthod Dentofacial Orthop*, **128**(5): 648-652.

Souames M, Bassigny F, Zenati N, Riordan PJ, Boy-Lefevre ML (2006). Orthodontic treatment need in French schoolchildren: an epidemiological study using the Index of Orthodontic Treatment Need. *Eur J Orthod*, **28**(6): 605-609.

Svedström-Oristo AL, Helenius H, Pietilä T, Pietilä I, Alanen P, Varrela J (2002). Reproducibility of characteristics assessing the occlusion of young adults. *Angle Orthod*, **72**(4): 310-315.