Journal of Oral Rehabilitation 2003 **30**; 113–118

Reliability of clinician judgements of bruxism

J. J. MARBACH*,†,§, K. G. RAPHAEL*,†, M. N. JANAL*,† & R. HIRSCHKORN-ROTH‡ **New Jersey Dental School, University of Medicine and Dentistry of New Jersey, Newark, NJ, †New Jersey Medical School, University of Medicine and Dentistry of New Jersey, Newark, NJ and ‡New York University College of Dentistry, New York, NY, USA*

SUMMARY Bruxism is considered to be a parafunctional disorder requiring treatment and is viewed as a risk factor for the development of temporomandibular disorders (TMDs). The purpose of this investigation is to examine the reliability of clinician judgements of bruxism severity. Twenty dentists who are faculty members in a dental school examined 29 stone casts and gold-plated models of individual teeth for evidence of bruxism. Ordinal ratings of bruxism severity for the 29 augmented models were made on two occasions, approximately 3 months apart. Inter-rater reliability among all clinicians, evaluated using intraclass correlation coefficients (ICCs), was poor at both time one and time two (i.e. ICC = 0·33 and 0·32, respectively), with somewhat better reliability found among those clinicians with above-average time elapsed since completion of dental training (i.e. ICC = 0·48 and 0·50 for time 1 and time 2, respectively). Three-month test–retest reliabilities were fair (ICC = 0·46) for the full group of raters and were unrelated to clinicians' degree of confidence in their ratings. These results indicate a need to standardize methods for clinical assessment of bruxism. Additionally, they have implications for studies using clinical assessments of bruxism to test the association between bruxism and other conditions such as TMDs.

KEYWORDS: bruxism, tooth attrition, reliability, temporomandibular disorders

Introduction

'Bruxism', defined as forcible clenching or grinding of the teeth, or a combination of both, has long been regarded as a disorder requiring treatment (Nadler, 1957; Attanasio, 1997). The view of this behaviour as a parafunctional disorder is likely due to the fact that, in extreme cases, it can cause tooth structure breakdown (Nadler, 1957; Pavone, 1985; Rawlinson, 1991). Moreover, it is widely believed (Le Resche, Truelove & Dworkin, 1993) that bruxism plays a role in the development of temporomandibular disorders (TMDs). As some patterns of tooth wear are often considered to be a sign of bruxism, reliable assessment of such patterns is necessary for the reliable clinical diagnosis of bruxism.

The three most common methods to evaluate bruxism are self-report questionnaires, clinical oral examination and sleep laboratory assessments (Seligman, Pullinger & Solberg, 1988). Studies have appeared which show that sleep laboratory electromyography (EMG)-based assessments are reliable (Bowley, Stockstill & Pierce, 1993; Rivera-Morales & McCall, 1995; Gallo *et al.*, 1997), but these studies have limited diagnostic utility in the typical clinical setting. In both research and clinical settings, an individual's bruxism status is typically based on clinical oral examination (e.g. Droukas, Lindee & Carlsson, 1984; Richmond *et al.*, 1984), the participants'

§**Joe Marbach died on 22 July 2001. He was an outstanding clinician who developed and promoted an understanding of the significance of psychological and psychiatric disturbances which could interrelate with some elements of clinical dentistry. It was a field in which he achieved international distinction. In some ways, he was an iconoclast and, as many such do, he could provoke perplexity, particularly amongst those colleagues consumed by the technology of dental practice. However, his integrity and innovative approach as a clinician have made a lasting positive impact on our profession. It is a privilege to publish, in conjunction with his co-workers, this, his last contribution.**

Arnold Franks

self-report of bruxism (Droukas *et al.*, 1984; Glass *et al.*, 1993; Lavigne & Montplaisir, 1994; Aromaa *et al.*, 1998; Hublin *et al.*, 1998; Hicks *et al.*, 1999; Israel *et al.*, 1999; Michalowicz *et al.*, 2000) or a combination of the two (Molina *et al.*, 1997; Gavish *et al.*, 2000).

There is a paucity of information about the reliability of judgements of bruxism based on oral examination. Although an incomplete surrogate for such information, scoring systems for tooth wear severity have been employed (e.g. Smith & Knight, 1984; Johansson *et al.*, 1993; Donachie & Walls, 1996). A single study (Johansson *et al.*, 1993) reported on the reliability of one such system. It found good reliability between two raters in scoring severity and progression of tooth wear, on a tooth-by-tooth basis. It is unknown how these results apply to interrater reliability of an overall judgement of bruxism among a group of dentists who have not been trained to use a specific occlusal wear scale. Moreover, bruxism and tooth wear are not synonymous. Attrition caused by bruxism is believed to produce a recognizable pattern of wear with 'highly polished facets, matching facets, ridges between facets, grooves, ledges and thinning out and scooping out of incisal edges of anterior teeth' (Nadler, 1979, p. 344) and, in theory, can be distinguished from tooth wear as a result of abrasion or erosion (Bartlett, Phillips & Smith, 1999).

The primary aims of this study were to assess interrater reliability among a group of academic clinicians and 3-month test–retest reliability of clinical judgements of bruxism, based on examination of augmented dental models. We also tested whether there were characteristics of raters or models that influenced the extent of reliability.

Materials and methods

Design

After obtaining approval from the university Institutional Review Board, dental school faculty members were recruited to examine 29 augmented sets of dental models and rate each set for the extent of bruxism. After 3 months (±1 week), they were asked to rate the same 29 sets again.

Raters

Twenty-four candidates were approached and 21 agreed. Of the 21, one completed only the first rating session. The average age of completers was 57·15 years (s.d. = 10·46, range 32–74), with the average participant graduating from dental school 30·15 years prior to study commencement (s.d. = 10·83, range 6–47). Eighteen of the 20 were male. The majority (75%) indicated that they had moderate experience assessing bruxism, while 15% indicated extensive experience and 10% indicated no experience.

Procedures and materials

Subjects were tested individually, in a well-illuminated and private room. The tester (RH-R) presented the subject with 29 sets of maxillary and mandibular stone casts, of good to excellent quality. Each cast was presented separately in one of 10 random sequences, to control for potential order effects. The casts were augmented by individual gold-plated models of each tooth (Fig. 1). The latter are extremely accurate reproductions of the dental anatomy of each tooth.

Subjects were informed that the materials came from an all female, case–control study of myofascial face pain. The age of the case was provided, because age can influence degree of tooth wear (Ekfeldt *et al.*, 1990; Seligman & Pullinger, 1995). Case status was not revealed to the subjects, in order to minimize expectancies about the relationship between case status and bruxism.

Subjects were asked to rate each cast, in concert with the gold-plated tooth models, for severity of bruxism, i.e. 'none', 'mild', 'moderate' or 'severe'. In addition, subjects rated each set for confidence in their rating of bruxism, i.e. 'poor', 'fair' or 'good'. Three months (±1 week) later, subjects were asked to rate the same 29 sets for severity of bruxism and confidence in bruxism rating. The order of presentation was again randomized among 10 orders, with the restriction that no subject rated the casts in the same sequence twice. To minimize attempts to recall prior ratings, participants were not told that the sets at 3-month follow-up were identical to the initial sets.

Statistical analysis

All data were analysed using SPSS 9·0*. Inter-rater reliability was quantified as the intra-class correlation coefficient (ICC), a multivariate analogue of the Pearson correlation coefficient, which is interpreted similar to the

*SPSS Inc., Chicago, IL, USA.

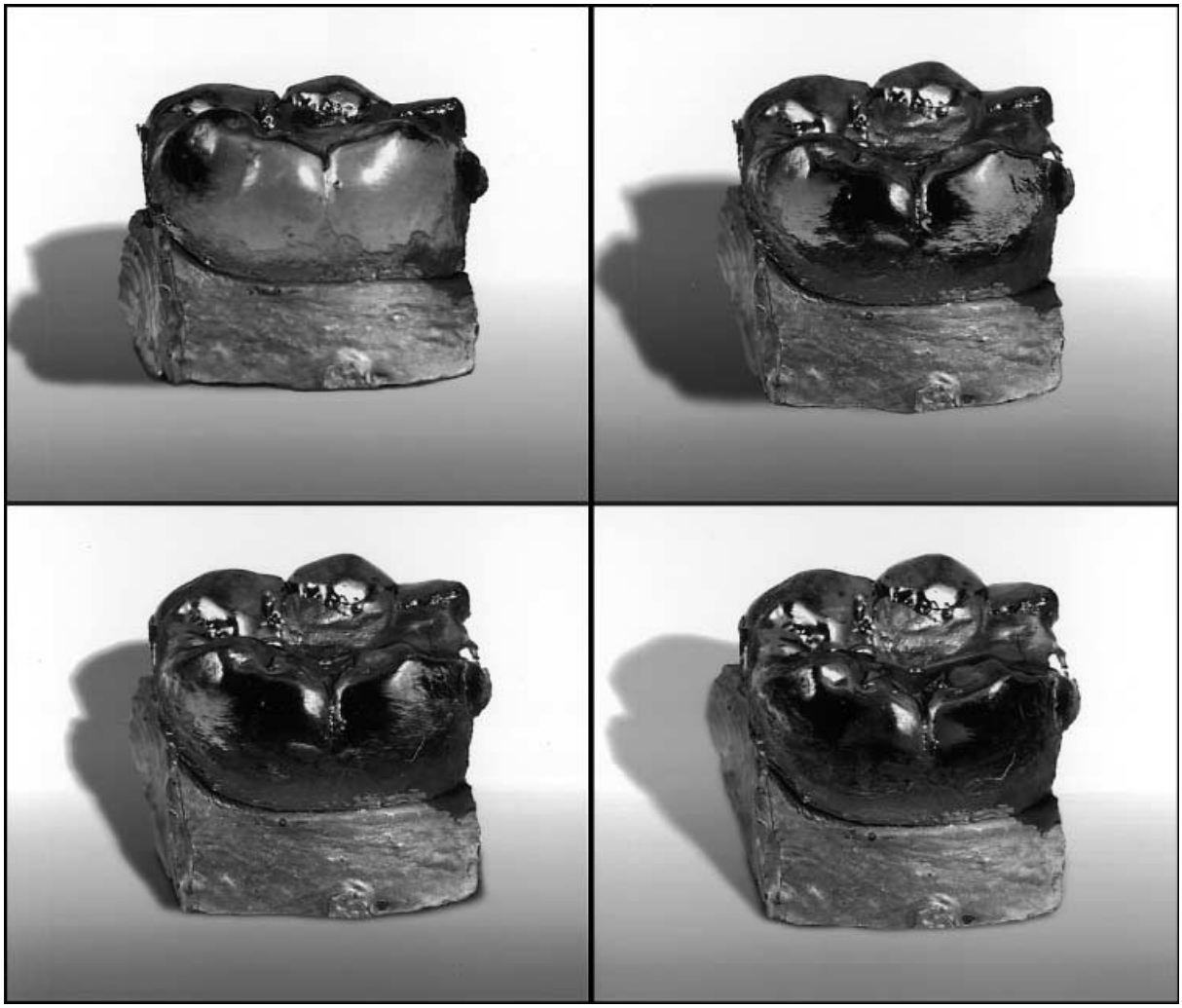

Fig. 1. Various lingual views of a gold-plated model of a mandibular molar, illustrating the fine detail provided by the plating method.

Pearson coefficient. The ICCs were calculated using a two-way random effects model (Shrout & Fleiss, 1979). Because there is controversy about whether a consistency approach (i.e. equal ordering) or an absolute agreement approach (i.e. equal rating) is correct (Shrout & Fleiss, 1979) in evaluating concordance in clinical decision making, both approaches were utilized for an initial examination of reliability. The consistency approach was used to examine whether rater characteristics affected inter-rater reliability, as this approach sets an upper limit to the estimate of inter-rater reliability.

Results

Inter-rater reliability coefficients among the 20 raters and across the 29 sets were first determined. At time 1, ICC = 0·33; at time 2, ICC = 0·32, using a two-way random effects model, consistency of agreement approach. As a more stringent test of agreement, ICC coefficients were also calculated using a two-way random effects model in which absolute agreement was required. Utilizing this form of the ICC, at time 1, ICC = 0·25, and at time 2, ICC = 0·21.

Next we tested whether there were rater characteristics affecting inter-rater reliability, using the consistency model. First, we compared reliability coefficients for those greater or equal to the median years since graduation (i.e. 33 years) to reliability coefficients for those raters who were below the median in years since graduation. At both time 1 and time 2, those with above-average years since graduation ($n = 11$) tended to show higher inter-rater reliability among one another

(time 1 ICC = 0·48; time 2 ICC = 0·50) than those with below-average years since graduation (time 1 ICC = 0·19, time 2 ICC = 0·19). Years of teaching experience did not affect the extent of inter-rater reliability, with those having less than 25 years teaching experience ($n = 10$) showing similar reliability at both time 1 (ICC = 0·37) and time 2 (ICC = 0·32) to those with 25 or more years of teaching experience (time 1 ICC = 31; time 2 ICC = 0·35). There was a trend towards an unexpected relationship between the rater's confidence in his or her ratings and agreement across other like-minded raters. Those raters ($n = 10$) whose average rating across all models equalled 'poor' or 'fair' confidence showed equal or better agreement on bruxism severity (time 1 ICC = 0·41; time 2 ICC = 0·37) than those raters whose average confidence rating was 'good' (time 1 ICC = 0·28; time 2 ICC = 0·30).

If a subset of the 29 models were selected to include only those models having the highest average bruxism severity ratings across raters ($n = 5$) and the models having the lowest average bruxism severity ratings ($n = 5$), the inter-rater reliability coefficients for these 10 sets was not better than for the full 29 models (i.e. time 1 ICC = 0·34; time 2 ICC = 0·35). Similarly, selecting the least variable items (where s.d. < 0·75, n of models = 9 and time 1 and 11 at time 2) at each test time failed to improve the level of agreement among judges (i.e. time 1 ICC = 0·32; time 2 ICC = 0·21). Thus, reliability was not improved by removing less clearly defined casts from the stimulus set.

Test–retest reliabilities were then computed, comparing the 20 judges' rating of all 29 models at time 1 to the ratings of all 29 models at time 2. The average test–retest correlation coefficient was better than the ICC among all raters and models (mean $r = 0·46$, s.d. = 0·16, median = 0·48, range = 0·20–0·94). When analysis was restricted to those models for which the rater had fair or good confidence at both time 1 and time 2, four raters who indicated poor confidence in rating all 29 models dropped out of the analysis. Nevertheless, the overall test–retest correlation among the remaining 16 raters was not markedly improved (mean $r = 0·48$, s.d. = 0·09, median = 0·51, range = 0·23–0·60).

Discussion

In studying the reliability of clinical judgements, the magnitude of the association or agreement is more salient than the statistical significance of the associations. Analyses here indicate that, among a group of experienced clinicians who are faculty members in a dental school, agreement about bruxism severity on the basis of augmented dental models is generally poor. Even among the subset of clinicians with the most experience, inter-rater reliability was only fair. Three-month test–retest reliability was somewhat better than inter-rater reliability, but there was great deal of variability in consistency among raters over this period. Confidence in one's judgements about the severity of bruxism bore little relationship to test–retest consistency.

These reliability coefficients compare poorly to the reliability of other clinical judgements made by dentists. For example, reliability estimates for bruxism are inferior to the moderate levels of agreement observed for dentists' treatment recommendation for individual teeth (Bader & Shugars, 1993) or the moderate to excellent agreement for judgement of caries or periodontal disease (Valachovic *et al.*, 1986; Langlais *et al.*, 1987; Flack *et al.*, 1996; Bader *et al.*, 1999).

There are several factors that may explain the poorer reliabilities for identification of bruxism, than caries or periodontal disease. First, most clinicians have little or no systematic training regarding signs of bruxism. Standards for detection of bruxism on the basis of wear patterns of models or teeth have not been widely established for clinical purposes. It is possible that, were such standards established and were clinicians trained, reliabilities could reach satisfactory levels. Secondly, raters in this study were asked to make a judgement about bruxism with less information than would be available in the usual clinical situation; i.e. in addition to examining the teeth, the clinician will typically interview the patient regarding his or her awareness of bruxism.

While the addition of patient self-report of bruxism might have improved the reliabilities observed here, we (Marbach *et al.*, 1990) and others (Seligman *et al.*, 1988) have previously questioned the impartiality of patient self-report, as such reports can be a reflection of a clinician telling a patient that she bruxes. This clinical judgement is especially likely to be made, even in the absence of characteristic wear patterns, for patients who experience facial pain. Other research (Droukas *et al.*, 1984; Seligman *et al.*, 1988; Ekfeldt *et al.*, 1990; Chung, Kim & Kim, 2000) has concurred in finding poor correspondence between wear patterns and self-reported bruxism.

Thus, the potential tautology inherent in reliance on patient self-report of bruxism dictates reliance on clinical judgements that do not involve patient self-report or clinical knowledge of TMD case status. The clinical judgements made here solely on the basis of dental models satisfy that condition. The examination of models and casts of individual teeth served as a surrogate for a clinical dental examination. The method was clearly superior to a clinical examination, with improved ability to examine the fine details of the dental anatomy. Nevertheless, clinical judgements of bruxism based on augmented models demonstrated poor levels of reliability.

Standardization of training to clinically diagnose bruxism is recommended. The first step in implementing such training will be to determine the specific characteristic patterns of teeth that are associated with 'true' bruxism. The gold standard will likely be derived from EMG and sleep studies. Yet, the relationship between tooth wear patterns and bruxism findings from EMG and sleep studies may be low, thwarting training efforts. Wear patterns on teeth invariably reflect a lifetime of functional and environmental influences, while data on bruxism collected from EMG and sleep studies can reflect only a current sample of behaviour, as bruxism has been shown to exhibit significant variability over time (Rugh & Harlan, 1988; Yap, 1998).

The external validity of the findings reported here might be compromised if the sample of dentists studied was atypical in terms of ability or training. Multiple analyses suggest that the participating dentists were not atypical. First, with the exception of years since doctoral training, demographical factors did not influence reliability estimates. Secondly, in analyses not detailed above, participants were asked four questions related to their knowledge and beliefs about TMDs and bruxism, derived from a previous survey of dentists (Le Resche *et al.*, 1993). Our participants' responses were virtually identical to those of dental specialists in the larger study.

One potential limitation to this investigation is that clinical ratings were made of augmented dental models rather than on the basis of full-mouth *in vivo* examination of patients. It is possible that reliabilities would have differed, had this latter procedure been used. Given the consistency of lighting conditions and the additional, fine detail afforded by the gold-plated models, we believe that the current study design created an optimal condition for judgement of wear patterns. Moreover, only four of the 20 raters consistently rated their confidence in their judgements of bruxism as 'poor'. This suggests that the large majority of raters considered the augmented models to provide sufficiently detailed information with which to make a clinical judgement of bruxism.

In sum, these data point to a pressing need for the development of reliable standards for the clinical assessment of bruxism. Current levels of unreliability limit the correct identification of those who may need preventive treatment to preserve tooth structure. Unreliability also reduces confidence in conclusions about the relationship between bruxism and TMDs, to the extent that such conclusions are based on clinical assessments of bruxism.

Acknowledgments

This investigation was supported in part by grant DE11714 from the National Institute of Dental and Craniofacial Research, National Institutes of Health, Bethesda, MD 20892. The authors thank Dr Mark Teaford for preparation of gold-plated models.

References

Aromaa, M., Sillanpaa, M.L., Rautava, P. & Helenius, H. (1998) Childhood headache at school entry: a controlled clinical study. *Neurology*, **50,** 1729.

Attanasio, R. (1997) An overview of bruxism and its management. *Dental Clinics of North America*, **41,** 229.

Bader, J.D. & Shugars, D.A. (1993) Agreement among dentists' recommendations for restorative treatment. *Journal of Dental Research*, **72,** 891.

Bader, J.D., White, B.A., Olsen, O. & Shugars, D.A. (1999) Dentist reliability in classifying disease risk and reason for treatment. *Journal of Public Health Dentistry*, **59,** 158.

Bartlett, D., Phillips, K. & Smith, B. (1999) A difference in perspective – the North American and European interpretations of tooth wear. *International Journal of Prosthodontics*, **12,** 401.

Bowley, J.F., Stockstill, J.W. & Pierce, C.J. (1993) Reliability and validity of instrumentation used to record nocturnal clenching and/or grinding. *Journal of Orofacial Pain*, **7,** 378.

Chung, S.C., Kim, Y.K. & Kim, H.S. (2000) Prevalence and patterns of nocturnal bruxofacets on stabilization splints in temporomandibular disorder patients. *Cranio*, **18,** 92.

Donachie, M.A. & Walls, A.W. (1996) The tooth wear index: a flawed epidemiological tool in an ageing population group. *Community Dentistry and Oral Epidemiology*, **24,** 152.

Droukas, B., Lindee, C. & Carlsson, G.E. (1984) Relationship between occlusal factors and signs and symptoms of mandibular dysfunction. A clinical study of 48 dental students. *Acta Odontologica Scandinavica*, **42,** 277.

Ekfeldt, A., Hugoson, A., Bergendal, T. & Helkimo, M. (1990) An individual tooth wear index and an analysis of factors correlated to incisal and occlusal wear in an adult Swedish population. *Acta Odontologica Scandinavica*, **48,** 343.

Flack, V.F., Atchison, K.A., Hewlett, E.R. & White, S.C. (1996) Relationships between clinician variability and radiographic guidelines. *Journal of Dental Research*, **75,** 775.

Gallo, L.M., Lavigne, G., Rompre, P. & Palla, S. (1997) Reliability of scoring EMG orofacial events: polysomnography compared with ambulatory recordings. *Journal of Sleep Research*, **6,** 259.

Gavish, A., Halachmi, M., Winocur, E. & Gazit, E. (2000) Oral habits and their association with signs and symptoms of temporomandibular disorders in adolescent girls. *Journal of Oral Rehabilitation*, **27,** 22.

Glass, E.G., McGlynn, F.D., Glaros, A.G., Melton, K. & Romans, K. (1993) Prevalence of temporomandibular disorder symptoms in a major metropolitan area. *Cranio*, **11,** 217.

Hicks, R.A., Lucero-Gorman, K., Bautista, J. & Hicks, G.J. (1999) Ethnicity and bruxism. *Perceptual and Motor Skills*, **88,** 240.

Hublin, C., Kaprio, J., Partinen, M. & Koskenvuo, M. (1998) Sleep bruxism based on self-report in a nationwide twin cohort. *Journal of Sleep Research*, **7,** 61.

Israel, H.A., Diamond, B., Saed-Nejad, F. & Ratcliffe, A. (1999) The relationship between parafunctional masticatory activity and arthroscopically diagnosed temporomandibular joint pathology. *Journal of Oral and Maxillofacial Surgery*, **57,** 1034.

Johansson, A., Haraldson, T., Omar, R., Kiliaridis, S. & Carlsson, G.E. (1993) A system for assessing the severity and progression of occlusal tooth wear. *Journal of Oral Rehabilitation*, **20,** 125.

Langlais, R.P., Skoczylas, L.J., Prihoda, T.J., Langland, O.E. & Schiff, T. (1987) Interpretation of bitewing radiographs: application of the kappa statistic to determine rater agreements. *Oral Surgery, Oral Medicine, and Oral Pathology*, **64,** 751.

Lavigne, G.J. & Montplaisir, J.Y. (1994) Restless legs syndrome and sleep bruxism: prevalence and association among Canadians. *Sleep*, **17,** 739.

Le Resche, L., Truelove, E.L. & Dworkin, S.F. (1993) Temporomandibular disorders: a survey of dentists' knowledge and beliefs. *Journal of the American Dental Association*, **124,** 97.

Marbach, J.J., Raphael, K.G., Dohrenwend, B.P. & Lennon, M.C. (1990) The validity of tooth grinding measures: etiology of pain dysfunction syndrome revisited. *Journal of the American Dental Association*, **120,** 327.

Michalowicz, B.S., Pihlstrom, B.L., Hodges, J.S. & Bouchard, T.J. Jr (2000) No heritability of temporomandibular joint signs and symptoms. *Journal of Dental Research*, **79,** 1573.

Molina, O.F., dos Santos, J., Nelson, S.J. & Grossman, E. (1997) Prevalence of modalities of headaches and bruxism among patients with craniomandibular disorder. *Cranio*, **15,** 314.

Nadler, S.C. (1957) Bruxism: a classification: critical review. *Journal of the American Dental Association*, **54,** 615.

Nadler, S.C. (1979) The treatment of bruxism – a review and analysis. *New York State Dental Journal*, **45,** 343.

Pavone, B.W. (1985) Bruxism and its effect on the natural teeth. *Journal of Prosthetic Dentistry*, **53,** 692.

Rawlinson, A. (1991) Treatment of root and alveolar bone resorption associated with bruxism. *British Dental Journal*, **170,** 445.

Richmond, G., Rugh, J.D., Dolfi, R. & Wasilewsky, J.W. (1984) Survey of bruxism in an institutionalized mentally retarded population. *American Journal of Mental Deficiency*, **88,** 418.

Rivera-Morales, W.C. & McCall, W.D. Jr (1995) Reliability of a portable electromyographic unit to measure bruxism. *Journal of Prosthetic Dentistry*, **73,** 184.

Rugh, J.D. & Harlan, J. (1988) Nocturnal bruxism and temporomandibular disorders. *Advances in Neurology*, **49,** 329.

Seligman, D.A. & Pullinger, A.G. (1995) The degree to which dental attrition in modern society is a function of age and of canine contact. *Journal of Orofacial Pain*, **9,** 266.

Seligman, D.A., Pullinger, A.G. & Solberg, W.K. (1988) The prevalence of dental attrition and its association with factors of age, gender, occlusion, and TMJ symptomatology. *Journal of Dental Research*, **67,** 1323.

Shrout, P.E. & Fleiss, J.L. (1979) Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin*, **86,** 420.

Smith, B.G. & Knight, J.K. (1984) An index for measuring the wear of teeth. *British Dental Journal*, **156,** 435.

Valachovic, R.W., Douglass, C.W., Berkey, C.S., McNeil, B.J. & Chauncey, H.H. (1986) Examiner reliability in dental radiography. *Journal of Dental Research*, **65,** 432.

Yap, A.U. (1998) Effects of stabilization appliances on nocturnal parafunctional activities in patients with and without signs of temporomandibular disorders. *Journal of Oral Rehabilitation*, **25,** 64.

Correspondence: Dr Karen G. Raphael, UMDNJ, BHSB 183 South Orange Ave., Rm F1512, Newark, NJ 07103, USA.
E-mail: karen.raphael@umdnj.edu