

Reproducibility of Probing Depth Measurements Using a Constant-Force Electronic Probe: Analysis of Inter- and Intraexaminer Variability

Marcelo W.B. Araujo,*† Kathleen M. Hovey,† Janice R. Benedek,† Sara G. Grossi,† Joan Dorn,† Jean Wactawski-Wende,†§ Robert J. Genco,† and Maurizio Trevisan†

Background: Probing depth (PD) is a commonly used method to determine periodontal disease severity in both treating and evaluating disease progression. Agreement among examiners collecting data in scientific investigations is necessary to establish reliable criteria for determining levels of periodontal attachment loss. The objective of our study was to evaluate inter- and intraexaminer variability of PD measurements among study examiners using a constant force periodontal probe, and to compare the variability of tooth-mean and quadrant-mean.

Methods: Three examiners, who had been previously trained and calibrated, performed measurements on 20 volunteers. Intra- and interexaminer variability of sites was determined by means of standard error of measurement (SE). Data analysis included determination of error for both quadrant mean and tooth mean.

Results: PD measurements for the quadrant-mean were used to calculate the intraexaminer variability, resulting in a mean (SD) value for an SE of 0.40 mm (± 0.02). Interexaminer variability for quadrant mean was 0.16 mm (± 0.02). For tooth-mean SE, the intraexaminer variability values were equal to 0.38 mm (± 0.07), and interexaminer variability equal to 0.24 mm (± 0.05).

Conclusions: All three examiners participating in our study were able to obtain reliable measurements for PD, using the constant force electronic probe. Reproducibility did not vary appreciably when using the whole quadrant mean compared to the tooth mean. These trained examiners were able to provide reproducible measures under 0.5 mm. *J Periodontol* 2003;74:1736-1740.

KEY WORDS

Periodontal pockets/diagnosis; periodontal probes; reproducibility of results.

* Dental Research and Graduate Studies Division, Department of Periodontics, Guarulhos University, São Paulo, SP, Brazil.

† Department of Social and Preventive Medicine, School of Public Health and Health Professions, University at Buffalo, Buffalo, NY.

‡ Department of Oral Biology, School of Dental Medicine, University at Buffalo.

§ Department of Gynecology-Obstetrics, School of Medicine and Biomedical Sciences, University at Buffalo.

Probing depth (PD) is a method frequently used to determine periodontal disease severity, and to evaluate both treatment and disease progression in scientific investigations. Models of periodontal disease progression have been developed based on PD data in different research design protocols.¹

Reliability estimates of PD have been shown to vary between individuals, between tooth types and across a series of examinations.² Therefore, in order to obtain site-specific diagnostic threshold values, it has been conventional to use a single estimate obtained by pooling recordings across sites or across teeth.² In 1983, Haffajee et al. proposed the tolerance method to determine periodontal disease measurement reproducibility and to assess disease progression.³ In this method, the difference between replicate attachment level measurements was used to calculate a standard deviation for all the measurements made in one individual.³ For reproducibility purposes, that research group considered 3 standard deviations from the mean as the threshold for attachment loss in an individual site.³ This method has been used throughout the years to determine reproducibility of measurements of periodontal disease.

Research methodology requires inter- and intraexaminer agreement in order to determine periodontal attachment loss in individuals beyond the error of the

method.⁴ Reproducibility of attachment level recordings is usually reported as the overall (pooled) standard deviation of the difference between duplicate recordings, or as the percentage of differences that agrees within a specified millimeter range.¹

Previous studies comparing the reproducibility of PD data used two units of comparison, the overall mean of the mouth (patient as the unit) or the tooth as the unit.^{1,2,4,5} Criticisms of both methods have been made. Gunsolley et al. investigated the influence of within-subject (mouth mean) and within-tooth variability of attachment level.⁵ Both units were included in the model at the same time to examine the subject correlation when building the model.⁵ However, the use of the subject as the unit may hide site-specific associations. Their results showed significant variance using both the tooth and the subject as the unit. Therefore, they concluded that both units should be evaluated when studying the reproducibility of PD or clinical attachment level measurements.⁵

Determinations of factors which influence the measurements of probing depth remain unclear, may vary by study, and are not yet well understood.⁶ Time between examinations, different examiners, and sequence of probing may be potential source(s) of error. In addition, patient characteristics such as severity of periodontal disease may influence error.⁶

The main objective of our study was to evaluate the inter- and intraexaminer variability of PD using a constant-force electronic probe and to compare the variability of tooth mean and quadrant mean, as the unit of measure.

MATERIALS AND METHODS

Three examiners (one dentist and two dental hygienists) measured probing depth of 20 healthy volunteers recruited from the staff and faculty of the Department of Social and Preventive Medicine at the University at Buffalo. The study protocol was reviewed and approved by the Health Sciences Institutional Review Board of the University at Buffalo. All participants provided signed consent.

All three examiners were trained and calibrated by an examiner considered to have the strongest expertise in the protocol, which was similar to the one used by Grossi et al.⁶

The PD measurements were made using a constant-force electronic probe, connected to a computer for instantaneous data capture and storage. Probing force was preset at 20 g. When the probe was in place, a foot switch was depressed and the computer automatically captured the measurement to the nearest 0.1 mm.⁶

Measurements were taken from one quadrant per subject, which was predetermined at the screening examination. The choice of quadrant was based on the number

of teeth present, to decrease the standard deviation from the mean.⁶ For better standardization quadrant 1 was the first choice, followed by 2, 3, or 4 respectively. Six sites around each tooth were measured (mesial-buccal, mid-buccal, distal-buccal, mesial-lingual, mid-lingual, and distal-lingual).⁶ Third molars were not included in the assessment because of their large variation in anatomy and position in the oral cavity. Care was taken to assure that measurements were made with the probe tip parallel to the tooth axis at the mid-buccal, mid-lingual and as close as possible to the mid-interproximal area from the buccal and lingual aspects, and the same angulations were carried on through all the sites.²

Subject screening was performed by one of the dentist examiners, using a conventional manual periodontal probe. Sites from one arch (upper or lower) were measured including molars and incisors. A participant was included if the mean PD at screening visit was ≤ 3.0 mm.

Volunteers had to meet the following criteria, which were established for a large epidemiologic study planned by our institution. Inclusion criteria: 1) 35 to 79 years old; 2) either gender; any ethnic background; 3) ≥ 6 natural teeth; and 4) PD mean value ≤ 3 mm. Exclusion criteria: 1) currently receiving periodontal therapy or receiving treatment in the last 6 months; 2) PD > 3 mm; 3) currently taking antibiotics; or 4) requiring prophylactic antibiotic therapy prior to dental examination, since participants would be probed six times in 14 days (twice by each examiner).

The quadrant chosen should present with six teeth. Where there were less than six teeth in the selected quadrant, the next quadrant was chosen. If no quadrant contained six teeth, the quadrant with the most teeth was measured to ensure that the highest number of possible sites was examined. Each examiner measured up to 42 sites per patient at the first visit (sequence 1) and then a different examiner repeated the measurement (sequence 2) in the same quadrant at the same appointment. At the second appointment, approximately 7 days later, the measurements were repeated by one of the examiners from visit one and the third examiner, who had not previously examined this subject. On the third visit, approximately 14 days after baseline, the examiners who had collected data only once completed the sequence of measurements. Therefore, all 20 subjects had PD data recorded twice in different visits by each of three examiners.

Statistical Analysis

Intra- and interexaminer variability of sites was determined by calculating means of standard error of measurement (SE), which was calculated as the square root of the site-specific variance, calculated by dividing the sum of the difference between two measurements, squared, divided by the number of pairs.⁶⁻⁸ The same

formula was applied for the analyses at both tooth and quadrant level assessment:

$$\text{Standard error of measurement (SE)} = \sqrt{(\Sigma D^2/2N)}$$

where D = difference between 2 mean values for each tooth per subject, across examiners and N = number of subjects

For intraexaminer variability calculations, we used values measured by the same examiner on two different occasions. Interexaminer variability was determined by the difference of values obtained by two examiners at the same visit; i.e., measurements recorded at the same appointment 5 to 10 minutes apart, by different examiners.

Data analysis was performed two ways: at quadrant-mean and tooth-mean measures, to allow us to determine if there is any influence of the type of tooth and/or its location on the variability among examiners, compared to the overall quadrant measurements.

RESULTS

Table 1 shows the values for intraexaminer variability of the probing data at the quadrant (subject) level. Quadrant mean values ranged from 2.09 to 2.24 mm and 2.10 to 2.21 mm at first and second visit, respectively, for all three examiners. Standard error of measurement for quadrant mean for examiner A was 0.38 mm; 0.39 mm for examiner B and 0.42 mm for examiner C (mean for all three examiners: 0.40 mm ± 0.02). These results indicate that all three examiners had a similar error when assessed at the quadrant level. In addition, the absolute threshold of error was less than 0.5 mm for all examiners.

Interexaminer variability at the quadrant level between examiners A and B was 0.14 mm; between examiners A and C, 0.18 mm; and for examiners B and C, 0.16 mm (mean SE: 0.16 mm ± 0.02) (Table 2).

Table 3 presents the data for intraexaminer variability at the tooth level. All three examiners had a similar SE, which was consistent across different teeth. The overall range of values was from 0.24 to 0.52 mm, depending on the tooth and the examiner.

When analyzing data for interexaminer reliability at

Table 1.
Intraexaminer Reliability of Constant-Force Electronic Probe

Examiner	Mean (SD) Visit 1	Mean (SD) Visit 2	Overall Mean (SD)	SE (mm)
A	2.24 (0.48)	2.10 (0.31)	2.17 (0.30)	0.38
B	2.28 (0.48)	2.16 (0.31)	2.22 (0.30)	0.39
C	2.09 (0.41)	2.21 (0.45)	2.15 (0.31)	0.42

Table 2.
Interexaminer Reliability of Constant-Force Electronic Probe

Examiners	Mean (SD) A	Mean (SD) B	Mean (SD) C	Mean (SD)	SE (mm)
A and B	2.08 (0.26)	2.18 (0.34)		2.13 (0.29)	0.14
A and C	2.22 (0.38)		2.21 (0.45)	2.21 (0.40)	0.18
B and C		2.2 (0.29)	2.03 (0.29)	2.11 (0.28)	0.16

Table 3.
Intraexaminer Reliability of Constant-Force Electronic Probe

Tooth	Mean Visit 1	Mean Visit 2	Overall Mean (SD)	SE (mm)
Examiner A				
2 molar*	2.28 (0.47)	2.22 (0.43)	2.25 (0.38)	0.34
1 molar*	2.32 (0.66)	2.22 (0.43)	2.16 (0.68)	0.44
2 premolar†	2.32 (0.59)	2.07 (0.41)	2.20 (0.42)	0.44
1 premolar*	2.31 (0.58)	2.12 (0.41)	2.22 (0.38)	0.47
Canine†	2.17 (0.43)	2.08 (0.37)	2.13 (0.28)	0.40
Lateral†	2.18 (0.49)	2.00 (0.33)	2.09 (0.30)	0.42
Central†	2.16 (0.49)	2.04 (0.40)	2.10 (0.30)	0.46
Examiner B				
2 molar*	2.32 (0.42)	2.45 (0.45)	2.39 (0.40)	0.25
1 molar*	2.42 (0.51)	2.38 (0.53)	2.40 (0.44)	0.40
2 premolar†	2.35 (0.48)	2.32 (0.49)	2.34 (0.41)	0.35
1 premolar*	2.21 (0.44)	2.22 (0.41)	2.22 (0.30)	0.41
Canine†	2.10 (0.39)	1.96 (0.37)	2.03 (0.28)	0.36
Lateral†	2.13 (0.30)	1.96 (0.28)	2.05 (0.24)	0.26
Central†	2.00 (0.32)	1.87 (0.28)	1.94 (0.25)	0.24
Examiner C				
2 molar*	2.29 (0.40)	2.49 (0.54)	2.39 (0.40)	0.38
1 molar*	2.34 (0.49)	2.55 (0.67)	2.44 (0.52)	0.40
2 premolar†	2.11 (0.37)	2.26 (0.58)	2.18 (0.41)	0.52
1 premolar*	2.10 (0.31)	2.27 (0.50)	2.18 (0.35)	0.35
Canine†	1.84 (0.35)	2.07 (0.37)	1.96 (0.27)	0.38
Lateral†	1.75 (0.25)	1.95 (0.44)	1.85 (0.27)	0.35
Central†	1.81 (0.33)	1.95 (0.38)	1.88 (0.26)	0.34

* N1 = 19.
† N2 = 20.

the tooth level (Table 4), we observed that the error was larger compared to the quadrant level (Table 2). At the tooth level, the overall range of values for interexaminer variability overlaps the range of values presented for intraexaminer variability. The lowest variation was observed between examiners A and C for the second premolars (0.15 mm). The highest SE was for measurements of the first molar between examiners A and B (0.35 mm).

Table 4.
Interexaminer Reliability of Constant-Force Electronic Probe

Tooth	Mean	Mean	Overall Mean (SD)	SE (mm)
Ex. A and B*	Examiner A	Examiner B		
2 molar†	2.21 (0.40)	2.44 (0.47)	2.32 (0.45)	0.22
1 molar†	2.18 (0.37)	2.40 (0.55)	2.29 (0.47)	0.35
2 premolar‡	2.14 (0.38)	2.33 (0.52)	2.35 (0.46)	0.27
1 premolar†	2.08 (0.29)	2.27 (0.45)	2.17 (0.39)	0.26
Canine‡	2.02 (0.33)	2.02 (0.41)	2.02 (0.37)	0.25
Lateral‡	1.97 (0.31)	1.98 (0.30)	1.98 (0.30)	0.27
Central‡	1.99 (0.34)	1.88 (0.29)	1.94 (0.32)	0.27
Ex. A and C†	Examiner A	Examiner C		
2 molar†	2.29 (0.50)	2.50 (0.56)	2.40 (0.53)	0.25
1 molar†	2.36 (0.69)	2.54 (0.69)	2.45 (0.69)	0.22
2 premolar‡	2.25 (0.63)	2.27 (0.58)	2.26 (0.60)	0.15
1 premolar†	2.35 (0.63)	2.29 (0.50)	2.32 (0.56)	0.21
Canine‡	2.24 (0.46)	2.07 (0.35)	2.16 (0.40)	0.26
Lateral‡	2.21 (0.48)	1.95 (0.43)	2.08 (0.47)	0.27
Central‡	2.21 (0.51)	1.95 (0.34)	2.08 (0.45)	0.27
Ex. B and C†	Examiner B	Examiner C		
2 molar†	2.34 (0.40)	2.29 (0.38)	2.31 (0.39)	0.18
1 molar†	2.40 (0.50)	2.35 (0.47)	2.37 (0.48)	0.20
2 premolar‡	2.34 (0.44)	2.10 (0.37)	2.22 (0.42)	0.20
1 premolar†	2.16 (0.39)	2.09 (0.31)	2.13 (0.35)	0.16
Canine‡	2.04 (0.36)	1.84 (0.37)	1.94 (0.38)	0.25
Lateral‡	2.12 (0.29)	1.75 (0.27)	1.93 (0.34)	0.27
Central‡	2.00 (0.31)	1.81 (0.37)	1.90 (0.35)	0.27

* Comparison of measurements from two examiners on the same visit.

† N1 = 19.

‡ N2 = 20.

DISCUSSION

Probing depth, i.e., measuring the distance from the gingival margin to the bottom of a pocket, is recognized as one of the most important tools used to determine periodontal disease status. Probing accuracy and precision are affected by factors such as probe design, probing force, position, pocket depth, and tissue inflammation.⁹ Several types of probes are available, with the most common probe used in clinical practice a manual version.¹⁰ The manual probe, however, relies on the clinician's judgment to indicate PD and usually the values are rounded up to the mm.¹¹ Electronic probes have been recently developed, and are widely used in clinical research due to their potential to reduce error because measurements are computerized using a constant probing force.¹⁰

The determination of variability and reproducibility of manual and electronic probe measurements have been widely reported. Osborn et al.¹⁰ reported that the range of error, measured by standard deviation (SD) of the difference, was similar for the probes tested. They indi-

cated that intraexaminer variability ranged from 0.46 to 0.58 mm for the electronic probe, while the manual probe was slightly higher (0.58 to 0.66 mm). In the case of interexaminer variability, the electronic probe measurements showed an SD ranging from 0.50 to 0.85 mm and the manual probe a range from 0.63 to 0.74 mm.

Quirynen et al.¹² performed a similar study to address the advantages of the electronic probe over the manual probe. The results indicated that when using a manual probe, examiners tend to report deeper recordings, probably due to the lack of standardization regarding the force applied during probing. This study indicated SD was equal to 0.60 mm for the manual probe and 0.76 mm for the electronic probe.¹²

A slightly different approach was presented by Perry et al. when, besides determining values for intra- and interexaminer variability consistent with the literature, they compared the probes in regard to time, comfort, and cost. They reported a similar variability in measurements for all probes, but noted the higher cost of electronic probes, as well as the increased time of examination compared to manual probes.¹¹

There is general agreement among investigators that electronic probes are associated with improved reproducibility of measures; this may be due to the lack of standardization of probing forces used during manual probing, manual recording of data, or level of calibration of examiners.¹³

The goal of the present study was to determine inter- and intraexaminer variability of PD using a constant force electronic probe and to compare the variability of tooth mean to the variability of quadrant mean within and between examiners. Our results are consistent with other findings of electronic probing reliability results.² Our examiners were able to show consistency of agreement, in terms of mean and standard deviation. There was an agreement between examiners when the measurements were taken at the same visit, and within examiners, using measurements taken by the same examiner in two different visits, 7 or 14 days apart. Regardless of comparison, error was always under 0.5 mm.

Our study went further than previously reported work,^{6,9-15} in that we calculated these values of reliability at both the quadrant and the tooth level. We observed similar ranges of values regardless of which measure was assessed. When the entire quadrant was used to calculate the intraexaminer variability, we obtained a mean SE value of 0.40 mm ± 0.02 (range: 0.38 to 0.42) (Table 1); at the same level (quadrant), interexaminer variability showed an SE mean value of 0.16 mm ± 0.02 (range: 0.14 to 0.18) (Table 2).

When we calculated the standard error of measurement at the tooth level, we observed a change in values, depending on the type of tooth and its location (Tables 3 and 4). We observed that anterior teeth presented a smaller degree of variation compared to

posterior, especially molars, likely due to easier access to the anterior teeth.

The differences between the values of inter- and intraexaminer variability may be due to the difference in days between the two measurements used to calculate the intravariability. Measurements for intra- and interexaminer variability at the same visit were not performed due to the discomfort caused by multiple probing actions with an electronic probe. We decided to measure the variability between two examiners at the same visit, and the intravariability with data from different visits, 7 to 14 days apart. We were aware that this time difference could affect the results, because some change in the pocket depth may occur due to probing and to the Hawthorne effect.¹⁶ However, when we stratified the calculations by tooth type this difference between intra- and interexaminer variability seemed to be diminished (Tables 3 and 4).

Replicate measurements recorded by the same examiner at the same visit are needed to clarify the role of time between two visits on determining technical error.² Magnusson et al. performed such a study and their results showed higher differences for interexaminer variability than our results, with error reaching values as high as 1.02 mm.² Grossi et al. performed a similar reliability study and proposed a similar method of examiner variation;⁶ their results showed interexaminer variability ranging from 0.4 to 0.6 mm. Our results are similar to those presented by other studies.^{6,14,15}

Our examiners received previous training and appeared to maintain the calibration protocols for probing measurement.⁶ Standardization of technique in this clinical research setting resulted in low measurement error, similar to prior studies.^{6,14,15} This level of reproducibility in other settings, such as private practice, may not be as high if strict assessment protocols are not followed. In addition, subjects in this study had limited periodontal disease (≤ 3.00 mm); therefore, these results may not apply to patients with more severe periodontal disease. In addition, private practice settings may limit assessment to manual probe techniques.

Our study suggests limited differences in the measurement of tooth and quadrant using the electronic probe. Replication of this study using manual probes needs to be undertaken before we determine if these results are consistent regardless of probe used.

Reliable measurements are necessary in periodontal research to determine new parameters of disease and their association to putative risk factors. Ongoing studies of probing measurements should strive to attain measures with limited error. Our study suggests that both inter- and intraexaminer variability measures of probing depth assessed by an electronic constant-force probe can be consistently reduced to levels under 0.5 mm, in a careful monitored setting. In addition, this study

demonstrates that reproducibility using electronic probe measures does not vary substantially whether data are evaluated at quadrant or tooth level.

REFERENCES

1. Villata L, Baelum V. Reproducibility of attachment level recordings using an electronic and a conventional probe. *J Periodontol* 1996;67:1292-1300.
2. Magnusson I, Clark WB, Marks RG, Gibbs CH, Manouchehr-Pour M, Low SB. Attachment level measurement with a constant force electronic probe. *J Clin Periodontol* 1988;15:185-188.
3. Haffajee AD, Socransky SS, Goodson JM. Comparison of different data analyses for detecting changes in attachment level. *J Clin Periodontol* 1983;10:298-310.
4. Best AM, Burmeister JA, Gunsolley JC, Brooks CN, Schenkein HA. Reliability of attachment loss measurements in a longitudinal clinical trial. *J Clin Periodontol* 1990;17:564-569.
5. Gunsolley JC, Williams DA, Schenkein HA. Variance component modeling of attachment level measurements. *J Clin Periodontol* 1994;21:289-295.
6. Grossi SG, Dunford RG, Ho A, Machtei EE, Genco RJ. Sources of error for periodontal probing measurements. *J Periodont Res* 1996;31:303-336.
7. Trevisan M, Canessa M, Laurenzi M, et al. Outward Na-K-Cl cotransport in human red blood cells: A methodological evaluation of a method for epidemiologic investigations. *Am J Epidemiol* 1987;125:860-868.
8. Trevisan M, Ostrow D, Cooper R, Liu K, Sparks S, Stamler J. Methodological assessment for red cell sodium concentration and sodium-dependent lithium efflux. *Clinica Chimica Acta* 1981;116:319-329.
9. Hefti AF. Periodontal probing. *Crit Rev Oral Biol Med* 1997;8:336-356.
10. Osborn J, Stoltenberg J, Huso B, Aeppli D, Pihlstrom B. Comparison of measurement of variability using a standard and a constant force periodontal probe. *J Periodontol* 1990;61:497-503.
11. Perry DA, Taggart EJ, Leung A, Newbrun E. Comparison of a conventional probe with electronic and manual pressure regulated probes. *J Periodontol* 1994;65:908-913.
12. Quirynen M, Callens A, van Steenberghe D, Nys M. Clinical evaluation of a constant electronic probe. *J Periodontol* 1993;64:35-39.
13. Wang S, Leknes KN, Zimmerman GJ, Sigurdsson TJ, Wikesjö ÜME, Selvig KA. Reproducibility of periodontal probing using a conventional manual and an automated force-controlled electronic probe. *J Periodontol* 1995;66:38-46.
14. Osborn JB, Stoltenberg JL, Huso BA, Aeppli DM, Pihlstrom BL. Comparison of measurement variability in subjects with moderate periodontitis using a conventional and constant force periodontal probe. *J Periodontol* 1992;63:283-289.
15. Clark WB, Yang MCK, Magnusson I. Measuring clinical attachment: Reproducibility of relative measurements with an electronic probe. *J Periodontol* 1992;63:831-838.
16. Jeffcoat MK. Principles and pitfalls of clinical trials design. *J Periodontol* 1992;63(Suppl.):1045-1051.

Correspondence: Dr. Marcelo W.B. Araujo, Universidade Guarulhos-UnG, Praça Tereza Cristina, 1 CEPPE, Guarulhos, SP 07023-070 Brazil. E-mail: mwaraujo@prof.ung.br.

Accepted for publication April 28, 2003.