The Open Public Health Journal

Content list available at: https://openpublichealthjournal.com

REVIEW ARTICLE

Kappa Statistics: A Method of Measuring Agreement in Dental Examinations

Farzan Madadizadeh[1], Hesam Ghafari[2] and Sajjad Bahariniya[3,*]

[1]*Departments of Biostatistics and Epidemiology, Center for Healthcare Data modeling, School of Public Health, Shahid Sadoughi University of Medical Sciences, Yazd, Iran*
[2]*School of Dentistry, Shahid Sadoughi University of Medical Sciences, Yazd, Iran*
[3]*Department of Health Services Management, School of Public Health, Shahid Sadoughi, Iran*

Abstract:

Statistical methods have always been the solution to medical problems. Due to the problem of inconsistency in the diagnosis of dentists, statistical science has been provided for measuring the compatibility of diagnosis and reliability of dentists. One of the most important statistical methods for examining the agreement between the two experiments or diagnoses is Kappa statistics which can be used in dental sciences. The present study reviewed different types of Kappa statistics for assessing agreement, including Cohen's kappa, Fleiss' kappa, and Cohen's weighted kappa.

Keywords: Kappa statistics, Dental examinations, Compatibility, Cohen's kappa, Fleiss' kappa, Cohen's weighted kappa.

Article History	Received: May 09, 2023	Revised: July 04, 2023	Accepted: July 13, 2023

1. INTRODUCTION

Dentistry is one of the sciences in that direct observation of the doctor and his personal opinion during the examination has a direct effect on the final diagnoses [1]. Due to the multiplicity of dentists, the Decayed Missing Filled (DMF) Index was introduced to integrate diagnoses and standardize diagnostic criteria in the field of caries [2]. However, later it became clear that the accuracy of the index is not a guarantee for the consistency of dentists' diagnosis. Even a dentist can make a heterogeneous diagnosis on two examinations. Factors such as prior knowledge, experience, and aptitude directly affect dentists' diagnoses.

Statistical methods can always be used to solve medical problems. Considering the fact that there is inconsistency in dental diagnosis, which both the World Health Organization and the International Dental Federation are concerned about, statistics is an indicator for measuring the consistency of diagnoses in dental examinations [3]. One of the most widely used statistical methods to check the agreement between two diagnoses used in dental sciences is the Kappa statistic. To the best of our knowledge, there were no tutorial studies that mention all types of Kappa statistics and applications in dentistry. So this tutorial study aimed to introduce all types of Kappa statistics and their application in dental science in a simple way for dummies. Different types of Kappa and their applications for measuring agreement in dentistry are presented with examples in the following sections.

* Address correspondence to this author at the Department of Health Services Management, School of Public Health, Shahid Sadoughi, Iran;
E-mail: sajjadbahari98@gmail.com

2. DIFFERENT TYPES OF KAPPA STATISTICS

2.1. Cohen's Kappa

Cohen's kappa (CK) was introduced by Jacob Cohen in 1960 and is often used to assess concordance between two raters [4]. CK is a statistic that is used to measure inter-rater reliability for qualitative items. Also, it is generally thought to be a more robust measure than a simple percent agreement calculation [5, 6]. In dentistry, this index can be defined as follows:

Suppose two dentists examine the same tooth, the rate of concordance in their diagnoses can be measured in the form of a statistical index called kappa. It requires the final diagnosis of each dentist to be a dichotomous variable, such as a healthy or decayed tooth. Accordingly, a table with two rows for the first diagnosis and two columns for the second diagnosis is made as a 2 × 2 table such as the one represented in Table **1**.

After creating a 2 x 2 table, Cohen's kappa (CK) uses the numbers in the table to evaluate the agreement of two dentists in examining patients and diagnosing decayed teeth [7]. To define this statistic, we first need to introduce the observed proportion of agreement and expected probability (Table **1**).

2.1.1. Raw Agreement (P0)

In Table **1**, agreement between the two dentists is

represented in cells A and D, so calculating the raw agreement can be done as follows: (Eq 1).

$$P0 = \frac{A+D}{N} \quad [8], \qquad (1)$$

2.1.2. The Expected Agreement ($\hat{P}$)

To calculate the expected agreement ($\hat{P}$), in each row and column, the sum of rows and columns is calculated, then the sum of the first row is multiplied by the sum of the first column and the sum of the second column is also multiplied by the sum of the second row. The resulting products are added together and divided by the square of the sample size. Below is the formula for calculating the expected agreement and Cohen's kappa: (Eq 2)

$$\text{Expected Agreement}: \hat{P} = \frac{[(A+B)\times(A+C)]+[(C+D)\times(B+D)]}{N^2} \quad [9] \qquad (2)$$

N= total sample size

Now, one can define kappa statistics using the formulae presented in sections I and II as follows: (Eq 3).

$$\text{Kappa statistics} = \frac{\text{Raw agreement (P0)} - \text{expected agreement}\left(\hat{P}\right)}{1-\text{expected agreement}\left(\hat{P}\right)} = \frac{P0-\hat{P}}{1-\hat{P}} \qquad (3)$$

[7, 9 - 11]

Table 1. Summary of diagnosis of two dentists

-	-	The Diagnosis of Dentist 2	
-	-	**Healthy Tooth**	**Decayed Tooth**
The diagnosis of Dentist 1	Healthy Tooth	**A**	B
	Decayed Tooth	C	**D**

The range of values for Kappa statistics is between -1 to 1 [12, 13]. If the values are less than zero, it indicates no agreement, values between 0.6 and 0.8 indicate moderate agreement and values greater than 0.8 mean nearly complete agreement [9]. A generalized form of Cohen's kappa statistic for more than two raters was introduced by Fleiss in 1970, known as Fleiss's kappa [7, 13].

Example 1: Suppose two dentists examine 100 teeth (N=100) at the end of a working day so both of them diagnose 40 decayed and 30 healthy teeth. Then, the first dentist will diagnose 20 cases of them as rotten, while the second dentist will diagnose them as healthy. Also if dentist 1 diagnoses 10 cases as healthy while the second dentist has diagnosed them as decayed at the same time, the data are given in the 2 × 2 table as follows (Table **2**).

Table 2. Results of diagnosis of two dentists

-	-	The Diagnosis of Dentist 2	
-	-	**Healthy Tooth**	**Decayed Tooth**
The diagnosis of Dentist 1	Healthy Tooth	**40**	10
	Decayed Tooth	20	**30**

For our example in Table **2**, the raw agreement in diagnosis between the two dentists is 0.7.

$$\left(P0 = \frac{40+30}{100} = 0.7\right)$$

So, for our example in Table **2**, its value is 0.46.

$$\left(\hat{P} = \frac{[50 \times 60] + [40 \times 40]}{100^2}\right) = 0.46$$

Therefore, in our example in Table **2**, the kappa statistic is equal to 0.44.

($Kappa\ statistics = \frac{0.7-0.46}{1-0.46} = 0.44$), it shows the agreement between to dentist is low.

2.2. Fleiss' Kappa

Fleiss' kappa was introduced by Fleiss *et al*. between 1970 and 2003 [14]. This index is used to check the agreement of the results diagnosed by two or more evaluators. It is necessary for the diagnosis result to be a qualitative variable (good or bad, sick or healthy, normal or moderate or severe) [15, 16].

In calculating Fleiss' kappa, one must first create a table where each row of the table corresponds to a patient and the columns represent the possible categorical scores so that there is one column for each score. The values in the cells of the table reflect the number of raters (dentists in the previous example) who chose that score (n_{ij}). Then a probability is calculated for the row and column according to the following relations. Row probabilities for each patient are calculated from (Eq. **4**) and column probabilities are calculated from equation (Eq. **6**). Then, the raw agreement probability is calculated based mean of raw probabilities (Pi) as (Eq. **5**). The amount of expected agreement is also calculated through relation. Finally, like the previous formula that we had in Cohen's kappa, the calculations are obtained (Table **3**).

Fleiss kappa is not a multi-rater extension of Cohen's kappa [17].

Table 3. Formulation of fleiss' kappa.

-	Range of Categorical Score (K)					-
Patients (N)	**1**	**2**	**3**	**4**	**…**	$\mathbf{P}_i$
Patient 1	n11	n12	n13	n14	-	-
Patient 2	n21	n22	n23	n24	-	-
….	-	-	-	-	nij	-
$\mathbf{P}_j$	-	-	-	-	-	-

Raw Probability $P_i = \frac{1}{n(n-1)}[(\sum_{j=1}^{K} n_{ij}^2) - n]$ **(4)**

Raw Agreement Probability $P_{Raw} = mean(Pi)$ **(5)**

Column probability $P_j = \frac{1}{Nn}\sum_{i=1}^{N} n_{ij}$ **(6)**

The expected agreement $\hat{P} = \sum_{j=1}^{K} P_j^2$ **(7)**

$$\textbf{Fleiss' kappa} = \frac{\text{Raw Agreement Probability - Expected agreement Probability}}{\text{1- Expected agreement Probability}} = \frac{P_{Raw}-\hat{P}}{1-\hat{P}} \qquad (8)$$

Where n= number of raters; j= range of scores (j=1,..,k); i=number of patients (i=1,..,N).

Example 2: fourteen dentists give grades from 1 to 5 about the severity of a tooth's damage. If 5 patients are examined, the statistical value of Fleiss' kappa is calculated below (Table **4**).

For our example in Table **3**, P1 is 0.157. $P1 = \frac{0+0+2+3+6}{14 \times 5} = \mathbf{0.157}$

And taking the second row,

$$P2 = \frac{1}{14\,(14-1)}\,(0^2 + 0^2 + 3^2 + 5^2 + 6^2 - 14) = \mathbf{0.302}$$

In order to calculate $\mathcal{P}$, we need to know the sum of P_i.

$$\sum_{i=1}^{N} Pi = 1.000 + 0.302 + 0.324 + 0.237 + 0.280 = \mathbf{2.143}$$

Over the whole sheet,

$$\mathcal{P} = \frac{1}{5} \times 2.143 = \mathbf{0.428}$$

$$\hat{P}\,0.157^2 + 0.128^2 + 0.271^2 + 0.142^2 + 0.300^2 = \mathbf{0.223}$$

$$\textbf{Fleiss' kappa} = \frac{0.428-0.223}{1-0.223} = \mathbf{0.263}$$ It shows a weak agreement.

2.3. Cohen's Weighted Kappa

The weighted kappa penalizes disagreements in terms of their seriousness whereas the unweighted kappa treats all differences equally. Consequently, Cohen's weighted kappa ought to be employed when data is in the form of a graded ordinal scale. In this situation, three matrices such as observed score matrix, the expected score matrix based on agreement, and the weight matrix are involved. Weight matrix cells located on the diagonal indicate agreement and off-diagonal cells indicate disagreement. The equation for weighted kappa is: (Eq 9).

$$\kappa = 1 - \frac{\sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij}x_{ij}}{\sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij}m_{ij}} \qquad \textbf{(9)}$$

where k = number of codes and W_{ij}, X_{ij} and M_{ij} are elements in the weight, observed, and expected matrices, respectively. When diagonal cells contain weights of 0 and all off-diagonal cells weights of 1, this formula produces the same value of kappa as the calculation given above (Figs. **1** and **2**).

Table 4. Values for computing of the Fleiss' kappa

-	Categorical Scoring					-
-	**1**	**2**	**3**	**4**	**5**	$\mathbf{P_i}$
Patient 1	0	0	0	0	14	**1.000**
Patient 2	0	0	3	5	6	**0.302**
Patient 3	2	2	8	1	1	**0.324**
Patient 4	3	2	6	3	0	**0.237**
Patient 5	6	5	2	1	0	**0.280**
Total	11	9	19	10	21	-

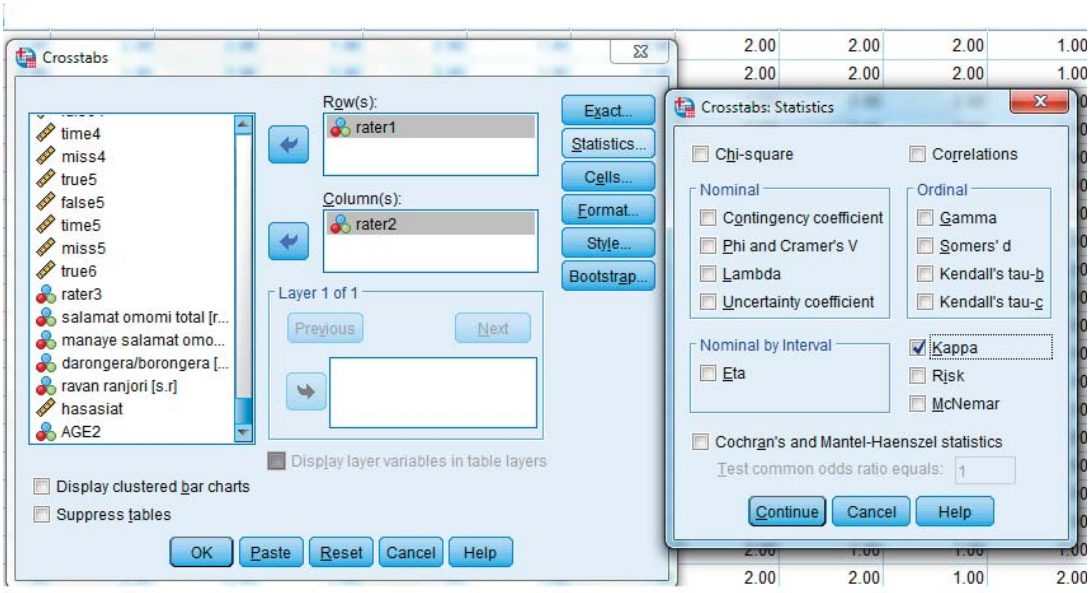

Fig. (1). Cohen's kappa in SPSS.

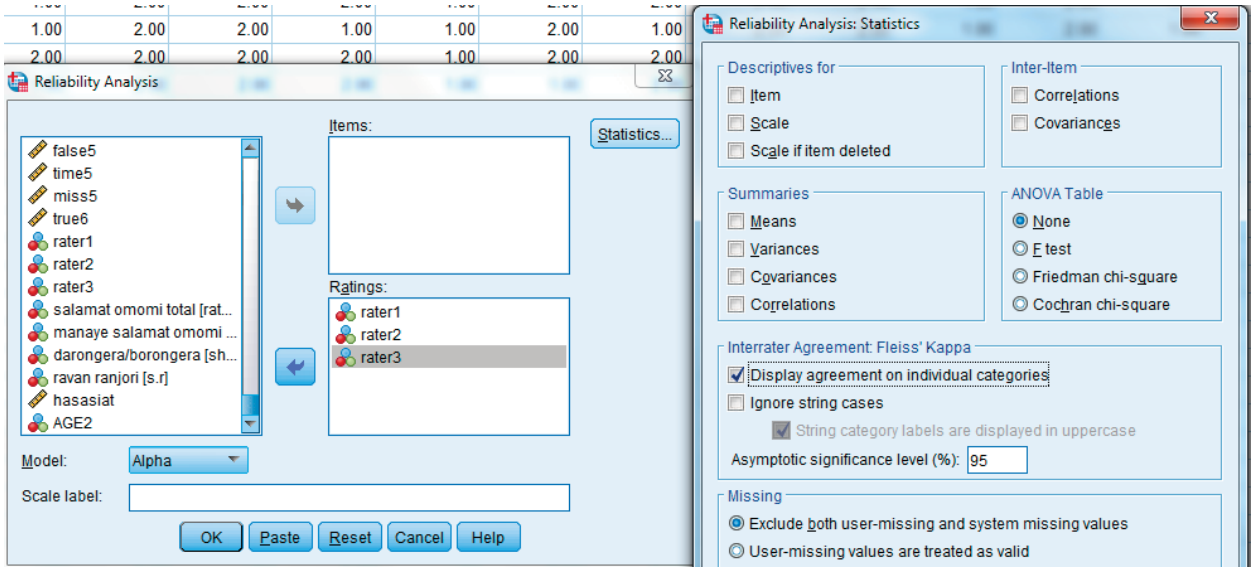

Fig. (2). Fleiss' kappa in SPSS.

Table 5. Interpretation of different values of all types of kappa statistics.

Kappa Value	**Interpretation**
< 0	NO agreement
0.01 – 0.20	Slight agreement
0.21 – 0.40	Fair agreement
0.41 – 0.60	Moderate agreement
0.61 – 0.80	Substantial agreement
0.81 – 1.00	Almost perfect agreement

Table 6. Summary of different Types of Kappa statistics.

Types of Kappa	Calculation Formulas	Usage
Cohen's kappa	$k = \frac{\text{Raw agreement (P0)} - \text{expected agreement}\left(\hat{P}\right)}{1 - \text{expected agreement}\left(\hat{P}\right)} = \frac{P0 - \hat{P}}{1 - \hat{P}}$	Determine the level of agreement among two independent raters - ratings are based solely on nominal scales -lacks the ability to measure inter-rater reliability with more than two independent raters. In such cases, a suitable alternative is to utilize Krippendorff's alpha.
Fleiss' kappa	$k = \frac{P_{Raw} - \hat{P}}{1 - \hat{P}}$	Determine the level of agreement among multiple independent raters, it can be used with binary or nominal-scale or ordinal or ranked data - it does not consider or account for the inherent order or hierarchy of categories within an ordinal variable. -it requires the random selection of patients and raters.
Cohen's weighted kappa	$\kappa = 1 - \frac{\sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij}x_{ij}}{\sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij}m_{ij}}$	Determine the level of agreement among two independent raters - ratings are based on ordinal scales

Landis and Koch (1977) gave the following table for interpreting ***Kappa*** values [18 - 20] (Table **5**).

In short, the types of Kappa calculation methods are presented in (Table **6**). Using this table, researchers will be able to choose the type of kappa suitable for their research (Table **6**).

Statistical software guideline:

SPSS:

Cohen's kappa: Analysis ⟶ Descriptive statistics ⟶ Cross tabs

Cohen's weighted kappa:

The way to perform weighted kappa is the same as Cohen's kappa, only before going to the above analysis, weighting should be done in the following way:

Data ⟶ weight cases

Fleiss' kappa: Analyze ⟶ scales ⟶ reliability analysis

R:

• Cohen's / Cohen's weighted kappa: the function 'kappa2' from the package 'irr',

• Other functions:

Cohen's unweight kappa: kappa Cohen (data, weight="unweighted")

Cohen's weighted kappa: kappa Cohen (data, weight="weighted")

Fleiss' kappa: function 'kappa m. fleiss' from the package 'irr

SATA:

Cohen's unweight kappa: kap rater1 rater2, tab

Cohen's weighted kappa: kap rater1 rater2 [fweight=wvar], tab

Fleiss' kappa: kap rater1 rater2 rater3

CONCLUSION

Cohen's kappa coefficient (κ) is a statistical measure utilized to assess the reliability between raters for qualitative items, both in terms of inter-rater and intra-rater reliability. In cross-classification, Cohen's weighted kappa is commonly employed as an agreement measure among raters(dentists). It serves as a suitable indicator for agreement when ratings are based solely on nominal scales without any inherent order structure. The original method of calculating Cohen's kappa lacks the ability to measure inter-rater reliability with more than two raters. In such cases, a suitable alternative is to utilize Krippendorff's alpha.

On the other hand, Cohen's weighted kappa is preferable when dealing with categorical data that possess an ordinal structure, such as rating systems with categories like high, medium, or low presence of a specific attribute. This weighted kappa places more emphasis on the seriousness of disagreements, while the unweighted kappa treats all disagreements as equally important. Therefore, the unweighted kappa is not suitable for ordinal scales.

Fleiss' kappa (κ), is another measure of inter-rater agreement, specifically employed to determine the level of agreement among multiple raters, it can be used with binary or nominal-scale data and can also be applied to ordinal or ranked data. However, in cases involving ordinal ratings, such as defect severity ratings on a scale of 1–5, Kendall's coefficients, which account for the ordering of categories, are usually more appropriate for determining association than kappa alone. Fleiss' kappa does not consider or account for the inherent order or hierarchy of categories within an ordinal variable.

CONSENT FOR PUBLICATION

Not applicable.

FUNDING

None.

CONFLICT OF INTEREST

The authors declare no conflict of interest financial or otherwise.

ACKNOWLEDGEMENTS

Declared none.

REFERENCES

[1] Omar H, Tai YT. Perception of smile esthetics among dental and nondental students. In: J Edu and Ethics Dentistry. 2014; 4: p. (2)54. https://www.jeed.in/text.asp?2014/4/2/54/148986

[2] Moradi G, Mohamadi Bolbanabad A, Moinafshar A, Adabi H, Sharafi M, Zareie B. Evaluation of Oral health status based on the decayed, missing and filled teeth (DMFT) index. Iran J Public Health 2020; 48(11): 2050-7. [http://dx.doi.org/10.18502/ijph.v48i11.3524] [PMID: 31970104]

[3] Carpenter CR, Detsky AS. Kappa statistic. CMAJ 2005; 173(1): 15-6. [http://dx.doi.org/10.1503/cmaj.1041742] [PMID: 15997024]

[4] Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas 1960; 20(1): 37-46. [http://dx.doi.org/10.1177/001316446002000104]

[5] Rau G, Shih YS. Evaluation of Cohen's kappa and other measures of inter-rater agreement for genre analysis and other nominal data. J Engl Acad Purposes 2021; 53: 101026. [http://dx.doi.org/10.1016/j.jeap.2021.101026]

[6] Daly CH, Neupane B, Beyene J, Thabane L, Straus SE, Hamid JS. Empirical evaluation of SUCRA-based treatment ranks in network meta-analysis: Quantifying robustness using Cohen's kappa. BMJ Open 2019; 9(9): e024625. [http://dx.doi.org/10.1136/bmjopen-2018-024625] [PMID: 31492773]

[7] Falotico R, Quatto P. Fleiss' kappa statistic without paradoxes. Qual Quant 2015; 49(2): 463-70. [http://dx.doi.org/10.1007/s11135-014-0003-1]

[8] Viera AJ, Garrett JM. Understanding interobserver agreement: The kappa statistic. Fam Med 2005; 37(5): 360-3. [PMID: 15883903]

[9] Nichols TR, Wisner PM, Cripe G, Gulabchand L. Putting the kappa statistic to use. Qual Assur J 2010; 13(3-4): 57-61. [http://dx.doi.org/10.1002/qaj.481]

[10] McHugh ML. Interrater reliability: The kappa statistic. Biochem Med 2012; 22(3): 276-82. [http://dx.doi.org/10.11613/BM.2012.031] [PMID: 23092060]

[11] Sim J, Wright CC. The kappa statistic in reliability studies: Use, interpretation, and sample size requirements. Phys Ther 2005; 85(3): 257-68. [http://dx.doi.org/10.1093/ptj/85.3.257] [PMID: 15733050]

[12] Reichenheim ME. Confidence intervals for the kappa statistic. Stata J 2004; 4(4): 421-8. [http://dx.doi.org/10.1177/1536867X0400400404]

[13] Robinson G, O'Donoghue P. A weighted kappa statistic for reliability testing in performance analysis of sport. Int J Perform Anal Sport 2007; 7(1): 12-9. [http://dx.doi.org/10.1080/24748668.2007.11868383]

[14] Fleiss JL, Nee JC, Landis JR. Large sample variance of kappa in the case of different sets of raters. Psychol Bull 1979; 86(5): 974-7. [http://dx.doi.org/10.1037/0033-2909.86.5.974]

[15] Marchevsky AM, Walts AE, Lissenberg-Witte BI, Thunnissen E. Pathologists should probably forget about kappa. Percent agreement, diagnostic specificity and related metrics provide more clinically applicable measures of interobserver variability. Ann Diagn Pathol 2020; 47: 151561. [http://dx.doi.org/10.1016/j.anndiagpath.2020.151561] [PMID: 32623312]

[16] Chicco D, Warrens MJ, Jurman G. The Matthews correlation coefficient (MCC) is more informative than Cohen's Kappa and Brier

score in binary classification assessment. IEEE Access 2021; 9: 78368-81.
[http://dx.doi.org/10.1109/ACCESS.2021.3084050]

[17] Scott WA. Reliability of content analysis: The case of nominal scale coding. Public Opin Q 1955; 19(3): 321-5.
[http://dx.doi.org/10.1086/266577]

[18] Zandbergen EGJ, Hijdra A, de Haan RJ, *et al.* Interobserver variation in the interpretation of SSEPs in anoxic–ischaemic coma. Clin Neurophysiol 2006; 117(7): 1529-35.
[http://dx.doi.org/10.1016/j.clinph.2006.03.018] [PMID: 16697253]

[19] Johnson EW, Ross J. Quantifying error in aerial survey data. Aust For 2008; 71(3): 216-22.
[http://dx.doi.org/10.1080/00049158.2008.10675038]

[20] Jayaraj G, Ramani P, Sherlin HJ, Premkumar P, Anuja N. Inter-observer agreement in grading oral epithelial dysplasia–A systematic review. J Oral Maxillofacial Surgery, Medicine Pathology 2015; 27(1): 112-6.
[http://dx.doi.org/10.1016/j.ajoms.2014.01.006]