*L*ibyan *M*edical *J*ournal
https://lmj.ly/index.php/ojs/index

Original Article

A Study of Examiner Variability in Assessment of Preclinical Class II Amalgam Preparation Using Two Evaluation Methods

Sumaya Aghila*, Eyman Elhadar, Amal Keshlaf, Farouk Ben Fadl

Citation: *Aghila S, Elhadar E, Keshlaf A, Ben Fadl F. A Study of Examiner Variability in Assessment of Preclinical Class II Amalgam Preparation Using Two Evaluation Methods. Libyan Med J. 2024;16(2):45-51.*

Received*: 16-06-2024*
Accepted*: 31-07-2024*
Published*: 04-08-2024*

Funding*: This research received no external funding.*

Conflicts *of Interest: The authors declare no conflict of interest.*

Department of Conservative Dentistry, Faculty of Dentistry and Oral Surgery, University of Tripoli, Tripoli, Libya
***Correspondence**: S.aghila@uot.edu.ly*

Abstract

Preclinical assessment is a useful strategy for promoting skill improvement in the clinical phase. It enables early intervention in the learning process and promotes effective use of training resources. The study aims to assess inter-examiner variability in class II cavity preparation performed by undergraduate dental students' evaluations using different scoring methods. The study evaluated 20 undergraduate students performing two Class II amalgam preparations performed on plastic molar teeth. The preparations were evaluated by four blinded independent examiners using two methods viz., Modified Neelakantan method and objective checklist scoring method. Statistical analysis for inter and intra examiner variability was tested using Friedman test and Wilcoxon signed rank test, respectively. The Kruskal-Wallis H test was employed to analyze scoring system variability and examiner consistency. The results showed that, scoring method (I) tends to have higher ranks than scoring method (II), the findings suggest that both Scoring method I and Scoring method II are reliable and consistent tools for evaluating Class II cavity preparations, with good inter-examiner agreement and intra-examiner reliability. Conclusion: The most important conclusion of our study is that both scoring methods are reliable for evaluating. Class II cavity preparations, with minimal inter- and intra-examiner variability. This suggests that these scoring methods can be used with confidence in pre-clinical practice, as they provide a consistent and accurate way to assess the quality of Class II cavity preparations.

Keywords: *Variability, Consistency, Undergraduate and Preparation.*

Introduction

The transition from pre-clinical training to clinical training in dentistry is a special time in the student's life because of the wide range of challenges, that the student must overcome. It is a difficult stage that is possibly one of the most important in the development of a career identity on both a technical and a personal level [1]. Preclinical laboratory instruction in the field of operative dentistry integrates exercises and tasks [2].

Second-year students at the university of dentistry in Tripoli, Libya, are examined with Class II cavity preparation for amalgam filling, although it is rarely used on Libyan patients. On the other hand, amalgam is still widely used in many developing countries [3].

This training requires assessment, which is designed to determine a student's level of knowledge, behavior, or skill development. It can be used not just to formally recognize the acquisition of knowledge or abilities but also to support learning and give students feedback on how they performed [4]. Since most students prefer to focus more on assessments and their results than any other aspect of the educational process, assessments are an essential component of the learning process [5]. The optimal assessment concept should have outstanding characteristics such as reliability, validity, accountability, flexibility, comprehensiveness, feasibility, timeliness, and reliance [6,7].

Problems with examiner consistency may lead students to recognize that evaluation methods are somewhat uninformed [5]. This concept can determine the learning process and produce a negative effect on undergraduate confidence and performance [7]. A method of assessment of both objectivity and reliability is essential [8], therefore, to endorse an effective system

of learning and to reduce friction between students and faculty over the issue of grading, objective, reliable, and practical methods need to be applied [9]. Examiner consistency is critical in the teaching and training process because it can affect the confidence and performance of the students. Therefore, new evaluation techniques and methods of standardizing assessments need to be further studied to encourage an efficient system of learning [10].
It is essential to highlight that inter-examiner reliability or agreement estimates the degree of consistency or agreement among examiners when assessing the results of the same group of students on the same task [11], whereas intra-examiner reliability or agreement describes the consistency of a single examiner in grading the same sample on two different occasions [12].
This study aimed to assess the impact of inter- and intra-examiner variability, both within and between examiners, on second-year dental students' Class II tooth preparation scores using two methods of evaluation at the school of dentistry at university of Tripoli.

Methods

Study Design

This study involved 20 second-year undergraduate dental students. The students attended a 2-hour didactic lecture on Class II cavity preparation for amalgam, followed by a 1-hour demonstration session, and they exercised one hour per week for three weeks. After the lecture and demonstration, the students performed Class II cavity preparations on teeth under controlled conditions.

Data Collection

The tooth models were collected after a 30-minute preparation period, adhering to the university examination time schedule. Each preparation was assigned a number code to ensure blinding. Four independent examiners evaluated the preparations in a double-blind manner using two different scoring systems. The evaluations were conducted without magnification, using an explorer and a mouth mirror under illumination. The assessments were performed in two stages, with a two-week interval between evaluations.

Study residents

Four independent faculty members from the school of dentistry at university of Tripoli, each with over ten years of clinical and teaching experience, served as examiners. These examiners were not involved in designing the checklist or scoring criteria and did not undergo specific calibration procedures. They were only briefed on the scoring distribution and criteria.

Methods of assessment

In the first stage, the examiners evaluated the preparations using the modified Neelakantan method (subjective study = scoring method I) [13]. After a two-week interval, the same examiners re-evaluated the same preparations using the objective checklist criteria scoring method (objective study = scoring method II) [14]. This was followed by a second evaluation two weeks later using the other scoring method two times with a two-week interval. E. Khalaf et al. [13] employed the Neelakantan method for student self-assessment, which we adapted to fit professor and examiner evaluation. It was chosen because it most closely resembles the college's evaluation procedure. It has only four subjective evaluation points, making it easier than Scoring Method II, which is deemed analytical.

Data analysis

The data were analyzed using non-parametric tests due to their non-normal distribution. The Wilcoxon signed-rank test was used to compare evaluation scores between the two scoring systems. Intra-examiner variability in evaluating Class II cavity preparations using each scoring system was assessed using the Friedman test. Inter-examiner variability in evaluating Class II cavity preparations using each scoring system was also analyzed using the Friedman test. The Kruskal-Wallis H test was employed to analyze scoring system variability and examiner consistency.

Results

Compare the evaluation scores between the two scoring systems.

Table 1 presents a comparative analysis revealing a statistically significant difference between the two scoring systems using the Wilcoxon Signed-Rank Test. A closer examination of the test results shows that both negative Z-values were statistically significant: Z = -4.545

for the first evaluation and Z = -5.833 for the second evaluation. These values indicate that Scoring System II tends to have lower ranks than scoring system I in both comparisons. Furthermore, the asymptotic significance values (0.001) confirm that these results are highly significant.

Table 1. Wilcoxon Signed-Rank Test: Comparative analysis of system I and system II

Comparison	Mean rank	Z-value	Asymp. sig.
Scoring systemII - Scoring System I (First evaluation)	40.88	-4.545b	.001
Scoring systemII - Scoring system I (second evaluation)	42.02	-5.833b	.001

Z = Z Value, Asymp. Sig. = Asymptomatic Significance

Inter-examiner variability in evaluating class II cavity preparations using each scoring system

Assessment of inter-examiner variability using scoring system I:

In table 2: the Friedman test was employed to assess inter-examiner variability in evaluating Class II cavity preparations using Scoring System I. The results revealed no statistically significant differences between the first and second evaluations across all four examiners, with p-values exceeding 0.05 (Table 3). Specifically, the p-values for each examiner were as follows: Examiner 1 (p = 0.251), Examiner 2 (p = 0.796), Examiner 3 (p = 0.637), and Examiner 4 (p = 0.796). These findings indicate that there is no significant inter-examiner variability in evaluating Class II cavity preparations using Scoring System I.

Table 2. Friedman test results for assessing inter- examiner variability in evaluating class II cavity preparations (first evaluation)

Examiner	Mean rank (First evaluations)	Mean rank (Second evaluations)	Chi-square	Asymp. sig.
Examiner 1	1.38	1.63	1.316	0.251
Examiner2	1.53	1.48	.067	0.796
Examiner3	1.55	1.45	.222	0.637
Examiner4	1.53	1.48	.067	0.796

Asymp. sig. = Asymptomatic significance

Assessment of inter-examiner variability using scoring system II

Table 3 presents the results of the Friedman test assessing inter-examiner variability using scoring system II. No statistically significant differences were found between the first and second evaluations across all four examiners, with p-values exceeding the conventional significance threshold ($\alpha = 0.05$). Specifically, the p-values for each examiner were as follows: examiner 1 (p = 0.346), examiner 2 (p = 0.166), examiner 3 (p = 0.251), and examiner 4 (p = 0.109). these findings indicate that there is no statistically significant inter-examiner variability in evaluating class ii cavity preparations using scoring system II**.**

Table 3: Friedman test results for assessing inter-examiner variability in evaluating class II cavity preparations (scoring system II)

Examiner	Mean rank (First evaluations)	Mean rank (Sec evaluations)	Chi-square	Asymp. sig.
Examiner 1	1.60	1.40	0.889	.346
Examiner 2	1.38	1.63	1.923	.166
Examiner 3	1.63	1.38	1.316	.251
Examiner 4	1.65	1.35	2.571	.109

Asymp. Sig. = Asymptomatic significance

Intra-examiner variability in evaluating class II cavity preparations using each scoring system

Assessment of intra-examiner variability using scoring system I

The Wilcoxon Signed Ranks Test was conducted to assess intra-examiner variability in evaluating Class II cavity preparations using Scoring System I (Table 4). The test revealed that there were no significant differences between the two evaluations for all examiners, as indicated by asymptotic significance (p-value) greater than 0.05.

These results suggest that all examiners demonstrated good intra-examiner reliability when evaluating Class II cavity preparations using Scoring System I, as there were no significant differences between their repeated evaluations. In other words, each examiner was consistent in their scoring and did not show any significant variation in their evaluation of Class II cavity preparations using Scoring System I over time.

Table 4. Wilcoxon Signed Ranks Test results for assessing intra-examiner variability in evaluating Class II cavity preparations using Scoring System I

Examiner	Z	Asymp. sig.
Examiner 1	-1.608	0.108
Examiner 2	-0.233	0.816
Examiner 3	-0.418	0.676
Examiner 4	-0.631	0.528

Asymp. Sig. = Asymptomatic significance

Assessment of intra-examiner variability using the scoring system, II

The Wilcoxon signed ranks test revealed that there were no significant differences between the two evaluations for all examiners, as indicated by asymptotic significance (p-value) greater than 0.05. Specifically, the results showed that Examiner 1 had a Z value of -0.066 and an asymptotic significance of 0.194, Examiner 2 had a Z value of -1.686 and an asymptotic significance of 0.083, Examiner 3 had a Z value of -1.733 and an asymptotic significance of 0.676, and Examiner 4 had a Z value of -0.381 and an asymptotic significance of 0.703 (table 5).

Table 5. Wilcoxon signed ranks test results for assessing intra-examiner variability in evaluating class II cavity preparations scoring system II

Examiner	Z	Asymp. sig.
Examiner 1	-.066	0.194
Examiner 2	-1.686	0.083
Examiner 3	-1.733	0.676
Examiner 4	-0.381	0.703

Asymp. Sig. = Asymptomatic significance

Scoring System Variability: An Analysis of Examiner Consistency

An analysis of four examiners' scoring patterns across four different systems revealed some interesting findings. While there were no significant differences in mean scores between examiners for Scoring System I's first evaluation, with a Kruskal-Wallis H value of 0.283 and an asymptotic significance of 0.963, there was a marginally significant difference for its second evaluation, with a Kruskal-Wallis H value of 6.056 and an asymptotic significance of 0.109, suggesting some variation in scoring approaches. In contrast, Scoring System II showed no significant differences in mean scores between examiners for both evaluations, with Kruskal-Wallis H values of 1.441 and 2.219, and asymptotic significances of 0.696 and 0.528, respectively (table 6).

Table 6. Mean ranks by examiner and scoring system

Examiner	Scoring system, I (First)	Scoring system, I (Second)	Scoring system II (First)	Scoring system II (Second)
Examiner 1	41.85	49.33	43.20	45.08
Examiner 2	41.50	43.68	35.20	40.03
Examiner 3	38.90	34.50	41.80	34.58
Examiner 4	39.40	34.50	41.80	42.33
Kruskal-Wallis H	0.283	6.056	1.441	2.219
Asymp. Sig.	0.963	0.109	0.696	0.528

Discussion

Pre-clinical exercises are an important way for students to improve the manual skills required to reach high competency levels in restorative dentistry [15]. To support an effective learning system and eliminate grading friction between learners and teachers, an objective and reliable assessment approach is required. In the current study, it was found that the students' scores were higher when using the scoring method I than when using the scoring method II, whether among the same assessor both times or among the results of all. The high scores of students in the scoring method I are most likely due to the fact that it is simple, as it consists of only four evaluation points for each toothand does not cover the exact details of the work and therefore may overlook some errors, unlike the scoring method II, which is analytical, consisting of fourteen points, is more accurate, and gives time for the examiner to investigate the preparation details [16].The scoring method II used in this study was based on earlier studies byHaj-Ali *et al.,* and Park *et al.,* as these scoring distributions more closely met the criteria of allowing feedback and reflecting at what stage or stages of preparation, students were performing poorly [17, 18].

In this study, we find that there was minimum variability between the examiners in both methods. A comparable result was achieved in studies by Bazan and Seale [19] and Schmitt et al [20] where a similarly conceived examiner's checklist for evaluation led to a similar reliability value. This agreement between the raters in our results may be due to the educational and evaluation experience of the four raters, despite their lack of knowledge of the two evaluation methods at the time of the research. However, experience has been shown to enhance inter-examiner agreement, with statistically significant variations amongst examiners [21].

The results obtained in this study were dissimilar to those seen in studies by Sherwood A. & Douglas V [14], Mhanni [16], Sharaff *et al.* [8], Vann et al. [22], Lilley et al. [23], and Satterthwaite & Grey [24]. This variation might be related to the preparation design (only one surface of class II) that examined .in the current study unlike eight different preparation designs and four preparation designs as in the studies by [16] and [14], respectively, which could have made the examiners better adjusted to our results for assessing students' preparations, leading to an improvement in their reliability in scoring. According to the results of [20], the reliability value can be increased by a higher number of examiners. This is relatively consistent with our study, where the teeth were evaluated by four examiners instead of three or two, as in other studies [8,14, 23,24], This interpretation is contrary to study of Vann et al. [22], which did not find agreement despite dealing with six examiners.

Moreover, the study revealed that, among the four examiners, the intra-examiner variability was non-significant. These results were similar to a study by Sharaf et al., [8] by showing that there was a reduction in intra-examiner variation when an objective check-list criteria scoring system was used. Our results were similar to those of the third examiner in the study of [14], he used in his studies five different evaluation methods, from descriptive to nearly subjective to a more analytical method, and his results had significant intra-examiner variability, except for the third examiner, and the validity of his evaluation was confirmed [16].

On the other hand, a marginally significant difference in mean scores was found between examiners for Scoring System I (second occasion), This is somewhat similar to studies by [8], [14], and [16] when they used subjective or nearly subjective scoring methods. The similarity in results may be due to the fact the fact that both evaluations are not analytical as an objective checklist criterion, which gives the opportunity to precisely define the evaluation, unlike descriptive or nearly subjective methods, which give opportunity to the examiner's personal opinion and sometimes his mood, which's led to limiting the variability

among examiners. Geopferd and Kerber used an analytical system for evaluation, using specific criteria; they reported that the strategy reduced variability among examiners more effectively in objective checklist criteria than the glance and grade method [25].
However, with the current study's scoring method I had the lowest intra-rater reliability compared to inter-raters, particularly on the second occasion. similar to the results of Lilley et al. [23] and Fuller [26], although the relationship in this study is not significant compared to their studies. While the present results differ from the studies of Houpt and Kress [27] and Deranleau et al. [26], they suggest that fewer points of evaluation are more likely to result in higher agreement. While in the current study, the method with lower evaluation points had relatively less agreement between the same rater both times, that may be attributable to the experience of the four examiners in our study in accurately evaluating students they were familiar with, even if they collaborated with it for the first time in this study and used the analytical method in an excellent manner.
It is important to mention that the scoring method I takes less time to evaluate than the scoring method II, as the average time taken to evaluate all teeth does not exceed 25 minutes compared to the scoring method II (50 minutes). According to Caro et al. (1979) [29], if the examiners spent more than 30 minutes assessing the tooth preparations, they might become tired and impact the result of the assessment also in Mhanni's A study indicated that examiners felt somewhat exhausted following the twelfth examinations [16].
Finally, this clear agreement between inter and intra examiners, especially in the objective checklist criteria, this indicates the reliability of the two methods and the extent of the examiners' compatibility and experience in dealing with the different methods.

Conclusions

The results suggest that there is no statistically significant inter-examiner or intra-examiner variability in evaluating Class II cavity preparations using either Scoring System I or Scoring System II. However, a marginally significant difference in mean scores was found between examiners for Scoring System I (Second occasion), suggesting possible differences in scoring patterns.

Limitations

Scoring system I may require further refinement to reduce potential variation in scoring approaches among examiners. In addition, the small sample size used in this study may negatively affect the results. The use of a Wilcoxon signed-rank test and a Friedman test may have limitations in terms of assumptions and interpretation of results.

Recommendations

Based on the conclusion, the recommendations could be: Both Scoring Systems I and II are reliable and consistent methods for evaluating Class II cavity preparations. The study suggests that both scoring systems can be used interchangeably, as they demonstrate minimal inter- and intra-examiner variability. Using the magnification, students prepare the teeth, and at the examiner's evaluation, they compare the results to the current study. The use of computerized dental assessment was proposed to overcome the limitations of using typodont in preclinical dental teaching compared to the visual method. Entering the students' self-assessment and evaluating its reliability and credibility, to improve self-directed learning.

Acknowledgments
would like to thank all examiners for their time.

Conflicts of Interest
There are no financial, personal, or professional conflicts of interest to declare.

References

1. Rodrigues P, Nicolau F, Norte M, Zorzal E, Botelho J, Machado V, et al. Preclinical dental students' self-assessment of an improved operative dentistry virtual reality simulator with haptic feedback. Scientific Reports. 2023;13(1):2823.
2. Obrez A, Briggs C, Buckman J, Goldstein L, Lamb C, Knight WG. Teaching clinically relevant dental anatomy in the dental curriculum: description and assessment of an innovative module. J Dent Educ 2011;75(6):797-804.
3. Scott BJ, Evans DJ, Drummond JR, Mossey PA, and Stirrups DR. An investigation into the use of a structured clinical operative test for the assessment of a clinical skill. Eur J Dent Educ 2001;5(1):31–7.
4. Taylor C, Grey N, Satterthwaite J. Assessing the clinical skills of dental students: a review of the literature. Journal of Education and Learning 2013;2(1):20–31.
5. El-Kiwashi M, Khalaf K, Al-Najjar D, Seraj Z, Al-Kawas S. Rethinking assessment concepts in dental education. Int J Dent 2020; ID8672303.

6. Mays KA, Levine E. Dental students's self-assessment of operative preparation using CAD/CAM: a preliminary analysis. J Dent Educ 2014;78:1673–80.
7. Turnbull J, Gray J, Improving in-traning evaluation programs. J Gen Intern Med 1998;13: 317–23.
8. Sharaf A, AbdelAziz A, El Meligy O. Intra- and inter-examiner variability in evaluating preclinical pediatric dentistry operative procedures. J Dent Educ. 71(4):540–44.
9. Lammari M, Alkhiary Y, Nawar A. Intra- and Inter-examiner Variability in Evaluating Impression Procedures at the Undergraduate Level. Kamla-Raj J Life Sci. 2013;5(1): 5–10.
10. Elsalhin A. Inter- and Intra-Examiner Variability in Evaluating Extra Coronal Full Coverage All Ceramic Crown Preparation. Aljabal J App Sci Hum. 2020;1(5):24–35.
11. Brown G, Manogue M, and Martin, M.: The validity and reliability of an OSCE in dentistry. European Journal of Dental Education, 1999;3(3): 117–125.
12. Dhuru VB, Rypel TS, and Johnston WM. Criterion-oriented grading system for the preclinical operative dentistry laboratory course. J Dent Educ. 1978;42(9): 528–31.
13. Khalaf M, Alkhubaizi Q, Alomari Q. Layered Base Plate Blocks and Operative Dentistry Skills. The Journal of Contemporary Dental Practice, 2018;19(5):1-6.
14. Sherwood A, Douglas A. A study of examiner variability in assessment of preclinical class II amalgam preparation. Journal of Education and Ethics in Dentistry 2014;1(4):12–17.
15. Huth KC, Baumann M, Kollmuss M, Hickel R, Fischer MR, and Paschos E. Assessment of practical tasks in the Phantom course of Conservative Dentistry by pre-defined criteria: a comparison between self-assessment by students and assessment by instructors. Eur J Dent Educ, 2017;21:37–45.
16. Mhanni AA. Evaluating and Improving the Assessment and Consistency of Feedback within the Clinical Skills Laboratory at Dundee Dental School. Dundee, 2018 Jan.
17. Haj-Ali R, Feil P. Rater reliability: short- and long-term effects of calibration training. J Dent Educ 2006;70(4):428–33.
18. Park RD, Susarla SM, Howell TH, Karimbux NY. Differences in clinical grading are associated with instructor status. Eur J Dent Educ 2009;13:318.
19. Bazan MT, Seale NS. A technique for immediate evaluation of preclinical exercises. J Dent Educ. 1982;46(12):726-28.
20. Schmitt L, Möltner A, Rüttermann S, and Gerhardt-Szép Study on the Interrater Reliability of an OSPE (Objective Structured Practical Examination)—Subject to the Evaluation Mode in the Phantom Course of Operative Dentistry. GMS Journal for Medical Education, 2016;33(4):1–19.
21. Mast, T. A., and Bethart, H. Evaluation of clinical dental procedures by senior dental students. Journal of Dental Education, 1978;42(4):196–97.
22. Vann W, Machen J, Hounshell P. Effects of criteria and checklists on reliability in preclinical evaluation. Journal of Dental Education. 1983;47(10):671–75.
23. Lilley JD, Bruggen Cate HJ, Holloway PJ, Holt JK, and Start KB. Reliability of practical tests in operative dentistry. Br Dent J 1968;125(5):194–97.
24. Satterthwaite J, Grey N. Peer-group assessment of pre-clinical operative skills in restorative dentistry and comparison with experienced assessors. European Journal of Dental Education, 2008;12(2): 99–102.
25. Goepferd S, Kerber P. Comparison of two methods for evaluating Class II cavity preparations. Journal of Dental Education. 1980;44(9):537–41.
26. Fuller J. The effects of training and criterion models on inter-judge reliability. Journal of Dental. 1972;36(4):19–22.
27. Houpt M, Kress G. Accuracy of measurement of clinical performance in dentistry. Journal of Dental Education. 1973;37(7):34–46.
28. Deranleau N, Feiker J, Beck M. Effect of percentage cut-off scores and scale point variation on preclinical project evaluation. Journal of Dental Education, 1983;47(10):650–55.
29. Caro T, Roper R, Young M, Dank G. Inter-observer reliability. Behaviour. 1979;69(3):303-15.