AJMAS
https://journal.utripoli.edu.ly/index.php/Alqalam eISSN 2707-7179

Original article

Evaluations of Case Presentations in Fixed Prosthodontics: Reliability of Examiner Pairings

Amina Elsalhin*, Milad Eshah, Mohamed Zeglam

Department of Fixed Prosthodontics, Faculty of Dentistry and Oral Surgery, University of Tripoli, Libya

ARTICLE INFO

Corresponding Email. *ami_rajab@yahoo.com*
Received*: 14-11-2022* ***Accepted****: 12-12-2022* ***Published****: 15-12-2022*
Keywords*. Evaluation, Case Presentation, Examiner Reliability*

ABSTRACT

Aims*. This study evaluates the inter-examiner reliability on student scores when assessing case presentations in Fixed Prosthodontics course.* ***Methods****. About 50 CDs were randomly selected from total of 64 CDs and these 50 CD has been distributed to four examiners, 25 CD for each two examiners. Each examiner evaluates the CDs separately. Fourth-year dental students presented their case presentation in CD in group of 4, hence the total of student is 100. Four full time faculty members participated in the evaluation session based on predefined criteria. Grading for rating presentation was scores of 20 marks. The scoring patterns of the evaluators were statistically analyzed using Intra-class coefficient ICC.* ***Results****. The results revealed that the obtained ICC value is 0.9493 (indicating excellent reliability) and its 95% confidence interval ranges between 0.9246 and 0.9659, meaning that there is 95% chance that the true ICC value lands on any point between 0.9246 and 0.9659. The results also revealed that the consistency between examiner 3 and 4 was more strong than examiner 1 and 2.* ***Conclusion****. The study concluded that excellent consistency among examiners (inter reliability) were persist. The results also revealed that the consistency between examiner 3 and 4 was higher than examiner 1 and 2.*

Cite this article: *Elsalhin A, Eshah M, Zeglam M. Microfacies Evaluations of Case Presentations in Fixed Prosthodontics: Reliability of Examiner Pairings. Alq J Med App Sci. 2022;5(2):580-584.*
https://doi.org/10.5281/zenodo.7443381

INTRODUCTION

Dental faculty must regularly evaluate students to estimate developing skills and clinical judgment [1]. The main element of systems designed to evaluate student competency are clinical grading and practical examination performance [2]. The performance of dental students has been described in the dental literature using different evaluation systems and grading methods. These approaches include cut-off scores, checklists, functional evaluation systems that employ performance criteria, analytical grading, rater calibration, student self-evaluation, mark-sense grading, computer tabulation of clinical tests using written criteria, anonymous examination, glance-and-grade evaluation systems, and a novel logbook checklist assessment system [3-5].

In spite of the fact that student assessment in dental schools has gain increasing attention, many investigations have focused on intra-rater or inter-rater liability [6]. Reliability in student evaluation presents serious problems for faculty who must make such judgments, and any lack of assessment consistency can also be a source of confusion and stress for dental students [7,8]. This problem was recognized as early as 1930 yet received little notice in the dental literature before 1970 [9]. However, after a comprehensive review of the literature in 1977, Myers concluded that subjectivity associated with clinical evaluation of student performance remained a source of disappointment for both dental students and clinical demonstrator [10].

Concerned by the extent of the problem of examiner consistency, Schiff et al., [11] designed a device called the “pulpal floor measuring instrument” to measure the profile of preparations, including depth, smoothness, and flatness of the pulpal floor. These authors reported significant improvement in examiner consistency using this equipment. In spite of the fact that such devices may have been useful as a teaching aid, probably their use would have been limited in an examination situation where raters would also need to consider other aspects of a preparation. An investigation has concentrated on the development of marking systems centered on specific criteria and checklists as an alternative to the glance-and-grade

method to improve rater performance, but the results have been blurry. Some researchers found that development of an analytical approach using detailed checklists improved examiner reliability [5].
Weinlander [12] proposed that more valid and reliable evaluations could be achieved if students received quick feedback about their performance on specific tasks but did not learn the actual score assigned for individual performance. He reported that this system reduced the faculty tendency to become too generous in assigning scores and therefore might improve the validity and reliability of rater scoring. Biller and Kerber [13] claimed that the effects of low inter-rater reliability could be decreased by rotating evaluator among the students.
The purpose of the current study was to identify whether the evaluators made similar judgments in the assignment of student scores when assessing case presentations in fixed Prosthodontics course. The null hypothesis was that faculty staff perform similarly in their judgment regarding the students' final grades.

METHODS

This descriptive study was conducted at Department of Fixed Prosthodontics, Faculty of Dentistry, Tripoli University. The CD contains case presentation that has been submitted by the 4th year dental students as part of their assessment competency, no special instructions were given to the students or examiner that we are going to use these presentations for subjective evaluations. 50 CDs were randomly selected from total of 64 CDs and these 50 CD has been distributed to four examiners, 25 CD for each two examiners. Each examiner evaluates the CDs separately. Students were given tutorial during the fixed Prosthodontics course. The objective of case presentation was prescribed and readily available to students at the beginning of the academic year. The course content encloses initial diagnosis, treatment planning, clinical and laboratory procedures. The student's work was scored separately by four full-time faculty staff members who had been assessed through confirmation procedures to make the evaluation standardize. The faculty members were then used the checklist to score the CDs. They had no information about the research goals. To reduce potential subjective bias, the evaluators were not provided with any students' academic details except their id numbers.
Evaluation used standard written criteria for each component of the evaluation. The grades used for each of five aspects of the presentation were 20 marks, using the detailed list of criteria (analytical method). These criteria gave a description for each possible grade for each component of an evaluation (Table 1). The grading sheets were reviewed to ensure their legibility and to make sure that each student had received a score. The data collected were entered to SPSS (statistical package for social science, Ink Illinois, USA) version 26.

Table 1. Criteria used for evaluation of the presentation.

Criteria Details	
1	Objectives: Presentation contents meet objectives.
2	Organization: Presentation well prepared and well organized.
3	Proper use of descriptive aids (Photographs, x-ray, study cast)
4	Content: Accurate of concepts and theories with up-to-date information
5	Length of presentation: Within time allocated and number of slides .

RESULTS

The data collected were entered to SPSS (statistical package for social science, Ink Illinois, USA) version 26. Intra-class correlation coefficient (ICC) is a widely used reliability index in test inter-rater reliability analyses. A more desirable measure of reliability should reflect both degree of correlation and agreement between measurements. Intra-class correlation coefficient (ICC) is such as an index. Reliability value ranges between 0 and 1, with values closer to 1 representing stronger reliability. The results of the Inter-rater analysis were shown in tables 2-5. (Examiner1& 2) and (examiner 3 & 4)
Tables 2,3 show output of a reliability analysis from SPSS. the obtained ICC was computed by two raters, 2-way random-effects model with 2 raters across 25 subjects, although the obtained ICC value is 0.9059 (indicating excellent reliability), its 95% confidence interval ranges between 0.8602 and 0.9367, meaning that there is 95% chance that the true ICC value lands on any point between 0.86 and 0.94. Therefore, based on statistical inference, it would be more appropriate to conclude the level of reliability to be "good" to "excellent."

Table 4,5 show that the obtained ICC was computed by two raters, 2-way random-effects model with 2 raters across 25 subjects, the obtained ICC value is 0.9493 (indicating excellent reliability) and its 95% confidence interval ranges between 0.9246 and 0.9659, meaning that there is 95% chance that the true ICC value lands on any point between 0.9246 and 0.9659. Values less than 0.5 are indicative of poor reliability, values between 0.5 and 0.75 indicate moderate reliability, values between 0.75 and 0.9 indicate good reliability, and values greater than 0.90 indicate excellent reliability.
For instance, according to the above guideline, if the 95% confident interval of an ICC estimate is 0.9246 and 0.9659 the level of reliability can be regarded as" excellent." It is because, in this case, the true ICC value supposes to land on any point between 0.9246 and 0.9659. However, let us say that the 95% confident interval of an ICC estimate is 0.92-0.97; the level of reliability should be regarded as "excellent" because even in the worst case scenario, the true ICC is still greater than 0.9. The results also revealed that the consistency between examiner 3 and 4 was more strong than examiner 1 and 2.
Figure 1 revealed Bland and Altman plot presentation of the limits of agreement (dot line) from -1.96 to +1.96

Figure 1. Bland and Altman plot

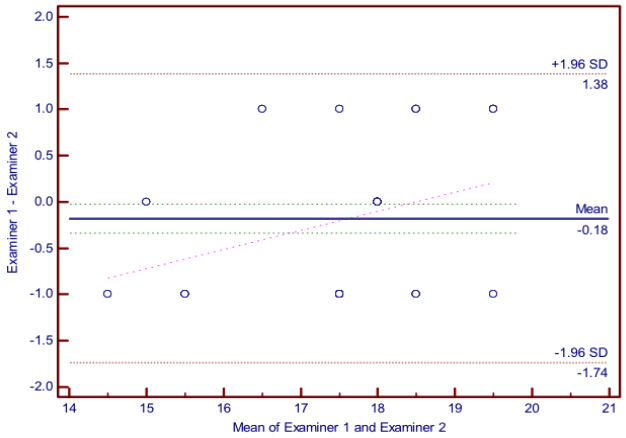

Table 2. ICC of Examiner 1and 2

Items	Number
Number of subjects (n)	100
Number of rater	2
Model	The same raters for all subjects. two-way model
Type	consistency
Measurements	Examiner1 Examiner 2

Table 3. ICC of Examiner 1and 2

Intraclass Correlation Coefficient							
	Intraclass Correlation[a]	95% Confidence Interval		F Test with True Value 0			
		Lower Bound	Upper Bound	Value	df1	df2	Sig
Single Measures[b]	0,828	0,755	0,881	10,632	99	99	0,000
Average Measures[c]	0,906	0,860	0,937	10,632	99	99	0,000

[a]*The degree of consistency among measurements*
[b]*Estimates the reliability of single ratings*
[c]*Estimates the reliability of averages of K ratings*

Table 4. ICC of Examiner 3 and 4

Items	Numbers
Number of subjects (n)	100
Number of rater	2
Model	The same raters for all subjects. two-way model
Type	consistency
Measurements	Examiner 3 Examiner 4

Table 5. ICC of *Examiner* 3 and 4

Intra-class Correlation Coefficient							
	Intraclass Correlation[a]	95% Confidence Interval		F Test with True Value 0			
		Lower Bound	Upper Bound	Value	df1	df2	Sig
Single Measures[b]	0,903	0,860	0,934	10,632	99	99	0,000
Average Measures[c]	0,949	0,925	0,966	10,632	99	99	0,000

DISCUSSION

The study findings support acceptance of the null hypothesis that the faculty performed similarly in their judgment regarding student grades. In disagreement with this result, the majority of researchers have agreed on the inconsistency among examiners in evaluating the performance of students even though instructors are calibrated annually [14,15].

In agreement with the current study some authors [16,17-19] have agreed that a calibration training program should include criteria development, a discussion of concepts, an explanation of the rating technique, practice with the rating technique, clearly defined criteria, a collection of pre-training scores, use of a gold standard, and a limited number of points on a rating scale. even though it appears that faculty members can become more consistent through calibration training, the literature contains mixed results for this training, ranging from slightly effective to not at all effective [2,4,5,12]. The literature is in agreement, however, on the appropriate frequency of calibration: It should be ongoing and held at regular intervals [4]. Calibration can be demanding and time-consuming, but it can be achieving through hard work, repetition, and maintenance [7,20].

In agreement with current study, Geopferd and Kerber [5] try to reduce variability among examiners, they used an analytical system for evaluation using specific criteria and a checklist. They reported that the technique was better than the glance-and-grade method in reducing variability among examiners. In another effort to reduce variability, researchers have used cut-off scores with percentages and a grading system; however, this approach disagrees with the work of Dahlstrom et al., [20] who reported a significantly increased inter-examiner reliability with application of percentage cut-off scores.

As noted, the purpose of this study was to evaluate the inter-examiner variability on student scores using a checklist and criteria system when assessing case presentations in fixed Prosthodontics course. The resulting scores are presumed precise reflection of student performance levels; nevertheless, a number of situational factors can also influence the score so that it may not be a precise reflection of the student's true performance level. These limitations are that certain faculty may be particularly firm or moderate in their ratings. To improve dental student presentation evaluation, more faculty training and calibration are needed, and the presence of an analytic method might improve consistency between evaluator by giving a clear understanding of the scoring criteria [21].

CONCLUSION

Within the limitations of this study, it can be concluded that there was Excellent consistency among examiners (inter reliability), indicating that the examiner reliability persists. Furthermore, the study findings reveal an increase in inter-

examiner 3 and 4 reliability comparing to examiner 1 and 3. This suggests that having more frequent calibration sessions may be valuable for maintaining an optimum level of calibration among the course faculty.

Disclaimer

The article has not been previously presented or published, and is not part of a thesis project.

Conflict of Interest

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

REFERENCES

1. Ingebrigtsen J, Røynstrand E, Berge M. An evaluation of the preclinical prosthodontic training at the Faculty of Dentistry, University of Bergen, Norway. Eur J Dent Educ. 2007;12(2):80-4.
2. Taleghani M, Solomon ES, Wathen WF. Non-graded clinical evaluation of dental students in a competency based educational program. J Dent Educ. 2004;68(6):644-55.
3. Deranleau NJ, Feiker JH, Beck M. Effect of percentage cut-off scores and scale point evaluation on preclinical project evaluation. J Dent Educ. 1983;47(10):650-5.
4. Gaines WG, Rasmussen RH, Uchello E. Increasing the objectivity of clinical grading. Dent Hyg. 1975;49(6):277-80.
5. Goepferd SJ, Kerber PE. A comparison of two methods for evaluating primary class II cavity preparations. J Dent Educ. 1980;44(9):537-42.
6. Garland JV, Newell KJ. Dental hygiene faculty calibration in the evaluation of calculus detection. J Dent Educ. 2009;73(3):383-9.
7. Haj-Ali R, Feil P. Rater reliability: short- and long-term effects of calibration training. J Dent Educ. 2006;70(4):428-33.
8. Henzi D, Davis E, Jasinevicius R, Hendricson W. North American dental students 'perspectives about their clinical education. J Dent Educ. 2006;70(4):361-77.
9. O'Connor P, Lorey RE. Improving inter-rater agreement in evaluation in dentistry by the use of comparison stimuli. J Dent Educ. 1978;42(4):174-9.
10. 10 Myers B. Beliefs of dental faculty and students about effective teaching behaviors. J Dent Educ. 1977;41(2):68-76.
11. Schiff AJ, Salvendy G, Root CM, Ferguson GW, Cunningham PR. Objective evaluation of quality in cavity preparation. J Dent Educ. 1975;39(2):92-6.
12. Weinlander GH. Period end clinical evaluation. J Dent Educ. 1979;43(12):633-6.
13. Biller IR, Kerber PE. Reliability of scaling error detection. J Dent Educ. 1980;44(4):206-10.
14. Salvendy G, Hinton WM, Ferguson GW, Cunningham PR. Pilot study on criteria in cavity preparation. J Dent Educ. 1973;37(10):27-31.
15. Jenkins SM, Drummer PM, Gilmore AS, Edmunds DH, Hicks R, Ash P. Evaluating undergraduate preclinical operative skill: use of glance and grade marking system. J Dent. 1998;26(8):679-84.
16. Vann WF, May KN, Shugars DA. Acquisition of psychomotor skills in dentistry: an experimental teaching method. J Dent Educ. 1981;45(10):567-75.
17. Courts FJ. Standardization and calibration in the evaluation of clinical performance. J Dent Educ. 1997;61(12):947-50
18. Knight GW. Toward faculty calibration. J Dent Educ. 1997;61(12):941-6.
19. Scruggs RR, Daniel SJ, Larkin A, Stoltz RF. Effects of specific criteria and calibration on examiner reliability. J Dent Hygiene. 1989;63:125-9.
20. Dahlström L, Keeling SD, Fricton JR, Galloway-Hilsenbeck S, Clark GM, Rugh JD. Evaluation of a training program intended to calibrate examiners of temporomandibular disorders. Acta Odontol Scand. 1994; 52(4):250-4.
21. Stellmack MA. An assessment of reliability and validity of a rubric for grading APA-style introductions. Teaching of Psychology, 2009; 36: 102–107.