IOSR Journal of Dental and Medical Sciences (IOSR-JDMS)
e-ISSN: 2279-0853, p-ISSN: 2279-0861.Volume 13, Issue 3 Ver. IV. (Mar. 2014), PP 28-31
www.iosrjournals.org

Reliability of Examiners' Evaluations of Oral Presentations in Fixed Prosthodontics

Salma A. Bahannan
Oral and Maxillofacial Prosthodontics Department, Faculty of Dentistry, King Abdulaziz University. Jeddah, Saudi Arabia.

Abstract: *This study assessed the inter-examiner variability on student scores when assessing oral presentations in Fixed Prosthodontics course. Oral presentations of fifth-year dental students were randomly selected. Students presented their topic within 15 minutes, followed by 10 minutes of detailed personal feedback and group discussion. Three senior faculty members participated in the evaluation session based on predefined criteria. Four levels of grading for rating presentation as highly acceptable, acceptable, marginally acceptable, or unacceptable were used. The scoring patterns of the evaluators were statistically analyzed using Friedman's test. The study revealed significant differences among the mean rank scores of the evaluators (P=0.001) for all criteria. Moreover, the mean rank scores of the 3 evaluators for the total value of all criteria were significantly different (P=0.001). The pairwise comparison test showed significant difference between each pair of evaluators (P=0.001) for all criteria. However, the difference was not significant for the first and the third evaluators for the second criteria (P=0.238). There was significant difference between each pair of evaluators (P=0.001) for the total value of all criteria.*

Key words*: reliability, evaluation, presentation, students, examiner.*

I. Introduction

Dental faculty must continually evaluate students to assess developing skills and clinical judgment.(1) The principal components of systems designed to evaluate student proficiency are clinical grading and practical examination performance.(2) Quantification of 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)

Although student evaluation in dental schools has received increasing attention, most investigations have concentrated on intra-rater or inter-rater liability.(6) Reliability in student evaluation presents serious problems for faculty who must render such judgments, and any lack of evaluation 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 frustration for both dental students and clinical instructors.(10)

Salvendy et al.(11) evaluated Class I amalgam preparations and found a high degree of both intra- and inter-examiner variation. The results of that investigation led the authors to suggest the development of more objective evaluation methods, such as optical scanners and electronic devices, to accurately measure cavity dimensions. In 1998, Jenkins et al.(12) evaluated the intra- and inter-examiner variability of a panel of examiners using a "glance and grade" marking system when assessing Class II preparations. The study revealed a high degree of both intra- and inter-examiner variability, with some preparations being given a passing grade on one occasion and failing on another and vice versa.

Worried by the extent of the problem of examiner consistency, Schiff et al.(13) 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 operator consistency using this equipment. Although such devices may have been beneficial as a teaching aid, presumably their use would have been limited in an examination situation where raters would also need to consider other features 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 equivocal. Some researchers found that development of an analytical approach using detailed checklists improved examiner reliability.(5)

Student evaluation thus has been the focus of much concern and discussion. In addition to the reliability and validity problems, evaluation in such courses has been plagued by a number of other issues

including a perceived lack of feedback to students. Furthermore, in some instances, preclinical evaluation has been perceived as arbitrary, and these problems and inconsistencies can undermine the learning process.(14) Another consideration is the legal aspect of student evaluation.(15) In reviewing the Horowitz case, Nash et al.(16) noted that such evaluation should include several conditions to satisfy legal requirements. Weinlander(17) suggested that more valid and reliable evaluations could be obtained if students received immediate feedback about their performance on specific tasks but did not learn the actual score assigned for individual performance. He suggested that this system reduced the faculty tendency to become too generous in assigning scores and therefore might improve the validity and reliability of instructor scoring. Biller and Kerber(18) claimed that the effects of low inter-rater reliability could be reduced by rotating instructors 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 oral presentations in Fixed Prosthodontics course. The null hypothesis was that faculty staff perform similarly in their judgment regarding the students' final grades.

II. Methodology

At the Faculty of Dentistry, King Abdulaziz University, 72 oral presentations of the fifth-year dental students were evaluated. Each student was given a subject review form in the Prosthodontic course. The objective of each topic was prescribed and readily available to students at the beginning of the academic year. The course content encompassed preliminary diagnosis, treatment planning, clinical and laboratory procedures. The student's work was graded anonymously and independently by three full-time faculty staff members who had been calibrated through training and verification procedures to attempt to standardize the evaluation. After faculty members training was completed, they were given the grading forms and asked not to discuss the grading system with one another. 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 computer numbers.

Evaluation used standard written criteria for each component of the evaluation. The four grades used for each of five aspects of the presentation were highly acceptable, acceptable, marginally acceptable, or unacceptable, using the checklist and detailed list of criteria (analytical method). These criteria gave a description for each possible grade for each component of an evaluation (Table 1). During the grading of the presentation, each instructor began on a different bench; no talking or consultation occurred until the evaluators had finished grading. During the session, each student was asked to present the topic in approximately 15 minutes, followed by 10 minutes of detailed personal feedback and group discussion. The grading sheets were reviewed to ensure their legibility and to make sure that each student had received a grade. A grade of 0 to 3 was assigned on each step, with 0 being most ideal and 3 the least favorable depending on the extent to which the specific criteria were met. Statistical Analysis was done using SAS. Friedman's test(19) was considered as a first option because the data had an ordinally scaled response variable. However, the Z-test was used to test if the variance of the distribution of block effects equaled zero. The test-setting value of significance was $\alpha = 0.001$.

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	Use of communication aids: Proper use of media and explanatory aids (Photographs, illustrations, etc).
4	Content: Accurate and complete explanation of key concepts and theories with up-to-date information.
5	Length of presentation: Within time allocated and number of slides .

III. Results

The frequency and percentage of evaluator scores for each criterion are shown in Table 2. The Friedman's F-test showed a significant difference among the mean rank scores of the three evaluators ($P = 0.001$) for all criteria (Table 3) and among their mean rank scores ($P = 0.001$) for the total value of all criteria ($df = 2$, $F = 101.74$, $P = 0.001$). The pairwise comparison test showed a significant difference between the mean rank scores of each of pairs of evaluators ($P = 0.001$) for all criteria; however, the difference was not significant for the first and the third evaluators for the second criterion ($P = 0.238$). Also, there was a significant difference between each of the two evaluators ($P = 0.001$) for the total value of all criteria. The Z-test showed no significant difference ($P = 0.136$).

Table 2. Score frequency/percentages among evaluators for all criteria 1 (n=72).

Criteria		Evaluator 1				Evaluator 2				Evaluator 3			
		0	1	2	3	0	1	2	3	0	1	2	3
1	No.	0	1	5	66	2	33	36	1	3	6	18	45
	%	0.00	1.39	6.94	91.67	2.78	45.83	50.00	1.39	4.17	8.33	25.00	62.50
2	No.	0	4	33	35	4	41	27	0	3	6	36	27
	%	0.00	5.56	45.83	48.61	5.56	56.94	37.50	0.00	4.17	8.33	50.00	37.50
3	No.	0	9	24	39	9	45	17	1	4	17	43	8
	%	0.00	12.50	33.33	54.17	12.50	62.50	23.61	1.39	5.56	23.61	59.72	11.11
4	No.	0	26	38	8	9	49	14	0	16	39	17	0
	%	0.00	36.11	52.78	11.11	12.50	68.06	19.44	0.00	22.22	54.17	23.61	0.00
5	No.	0	8	23	41	4	38	26	4	7	17	42	6
	%	0.00	11.11	31.94	56.94	5.56	52.78	36.11	5.56	9.72	32.61	58.33	8.33

No.= numbers of Students % = percentage of the students
0 = unacceptable,1= marginally acceptable, 2= acceptable, 3= highly acceptable

Table 3. Mean rank scores among evaluators for each criteria applying Friedman's test.

Criteria	df	F value	P value
1	2	118.6	0.001
2	2	79.9	0.001
3	2	80.7	0.001
4	2	34.9	0.001
5	2	90.1	0.001

IV. Discussion

The study findings support rejection of the null hypothesis that the faculty performed similarly in their judgment regarding student grades. In concordance 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.(11,12)

Several authors(14,21-23) have concurred 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, concrete examples, a collection of pre-training scores, use of a gold standard, and a limited number of points on a rating scale. Although 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.(2,4,5,17) The literature is in consensus, however, on the appropriate frequency of calibration: It should be ongoing and held at regular intervals.(4) Calibration can be difficult and time-consuming, but it is achievable through hard work, repetition, and maintenance.(7,24)

Trying to reduce variability among examiners, Geopferd and Kerber(5) 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.(24) who reported a significantly increased inter-examiner reliability with application of percentage cut-off scores.

In many teaching institutions, the glance-and-grade method is applied especially by more experienced faculty because of practicalities. Salvendy et al.(11) reported that it is important to develop a practical, reproducible, and easily applicable method, while others suggested including better faculty training and developing a more comprehensive system for evaluation or obtaining two or more assessments provided by at least two evaluators and calculating an average.(12) Other authors have recommended application of more frequent and uniform training sessions to improve evaluator reliability. Finally, examiner consistency is crucial in the teaching and learning 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 promote an efficient system of learning.

As noted, the purpose of this study was to investigate the inter-examiner variability on student scores using a checklist and criteria system when assessing oral presentations in fixed Prosthodontics course. The resulting scores are presumably an accurate reflection of student performance levels; however, a number of situational factors can also influence the score so that it may not be an accurate reflection of the student's true performance level. These limitations are that certain faculty may be particularly stringent or lenient in their ratings. To improve dental student presentation evaluation, more faculty training and calibration are needed, and the presence of an analytic rubric might increase consistency between graders by providing a clear understanding of the scoring criteria.(25)

V. Conclusion

Within the limitations of this study, it can be concluded that evaluators' scores differed significantly, indicating that the problem of inter-examiner reliability and variability persists. Further researches in this area are needed.

Acknowledgment

The author would like to express her sincere thanks to Dr Ali Al-Marshadi for his assistance in the statistical analysis.

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 srudents 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] Myers B. Beliefs of dental faculty and students about effective teaching behaviors. *J Dent Educ. 1977;41(2):68-76.*
[11] Salvendy G, Hinton WM, Ferguson GW, Cunningham PR. Pilot study on criteria in cavity preparation. *J Dent Educ. 1973;37(10):27-31.*
[12] 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.*
[13] 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.*
[14] Vann WF, May KN, Shugars DA. Acquisition of psychomotor skills in dentistry: an experimental teaching method. *J Dent Educ. 1981;45(10):567-75.*
[15] Shugars DA, May KN, Vann WF. Comprehensive evaluation in a preclinical restorative dentistry technique course. *J Dent Educ. 1981;45(12):801-3.*
[16] Nash DA, Moore RN, Andes JO. Academic dismissal for clinical reasons: implications of the Horowitz case. *J Dent Educ. 1981;45(3):150-5.*
[17] Weinlander GH. Period end clinical evaluation. J Dent Educ. 1979;43(12):633-6.
[18] Biller IR, Kerber PE. Reliability of scaling error detection. *J Dent Educ. 1980;44(4):206-10.*
[19] Friedman M. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *J Am Statist Assoc. 1937; 32(2):675-701.*
[20] Moore RN, Nash DA, Andes JO. Academic and disciplinary dismissal in dental education: the legal basis. *J Dent Educ. 1980;44(12):705-11.*
[21] Courts FJ. Standardization and calibration in the evaluation of clinical performance. *J Dent Educ. 1997;61(12):947-50.*
[22] Knight GW. Toward faculty calibration. *J Dent Educ. 1997;61(12):941-6.*
[23] Scruggs RR, Daniel SJ, Larkin A, Stoltz RF. Effects of specific criteria and calibration on examiner reliability. *J Dent Hygiene. 1989;63:125-9.*
[24] 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.*
[25] Stellmack MA. An assessment of reliability and validity of a rubric for grading APA-style introductions. *Teaching of Psychology, 2009; 36: 102–107.*