Electronic Physician (ISSN: 2008-5842) http://www.ephysician.ir
September 2018, Volume: 10, Issue: 9, Pages: 7287-7292, DOI: http://dx.doi.org/10.19082/7287

Inter-rater and intra-raters' variability in evaluating complete dentures insertion procedure in senior undergraduates' prosthodontics clinics

Manal Alammari[1], El-Sayed Nawar[2]

[1] Ph.D., Associate Professor in Prosthodontic, Oral and Maxillofacial Rehabilitation Department, Faculty of Dentistry, King Abdulaziz University, Jeddah, Saudi Arabia
[2] Ph.D., Professor in Prosthodontic, Oral and Maxillofacial Rehabilitation Department, Faculty of Dentistry, King Abdulaziz University, Jeddah, Saudi Arabia

Type of article: Original

Abstract

Background: Direct clinical assessment is the mainstay of evaluation in dentistry education. An effective evaluation method in prosthodontics should be equally valid and consistent; however, this is not attained frequently. A limited number of studies have applied an analytic evaluation in prosthodontics.

Objective: To compare the intra- and inter-raters' variability in two evaluation methods: glance and grade (global), and checklist and criteria (analytical). Moreover, to identify the components of the analytical evaluation system and its applicability.

Methods: This cross-sectional study was carried out on outpatients attending removable prosthodontics clinics affiliated with King Abdulaziz University (Jeddah, Saudi Arabia) from December 2017 to April 2018. Two prosthodontist examiners evaluated a sample of 35 complete denture cases (20 male, 15 female) twice over a period of five months. Inter-rater and intra-rater agreement were computed using reliability test (interclass correlation coefficient ICC). Data were analyzed in IBM SPSS version 23, using paired-samples t-test, weighted kappa, and Wilcoxon signed-rank test. The level of significance was set at $p \leq 0.05$.

Results: The intra-rater agreement for the first and second exposures under global and analytical evaluation methods with Examiner A was outstanding with 90.7% and 99.8% agreement respectively. While with Examiner B, global was lower but still in the acceptable range with about 78.1%, and 96.1% for the analytical evaluation. Inter-rater reliability analysis showed high agreement between the two raters in the first exposure of the analytical evaluation with 97.3%, while it was 87.5% in the global evaluation. This trend goes the same with analytical in the second exposure with 93.2%; however, the second exposure under global evaluation failed the cut off, which is only 56.6% agreement. In evaluation of inter-raters agreement, the second exposure of the global method demonstrated inconsistency between the two examiners ($p=0.002$), while the analytical second exposure demonstrated more homogeneity ($p=0.050$). Intra-rater variability between first and second exposure in analytical evaluation was (0.711 for the first rater and 0.677 for the second rater). Intra-rater variability between first and second exposure in global evaluation was (<0.001 for the first rater and 0.075 for the second rater).

Conclusion: A simple objective and detailed method to evaluate complete denture insertion procedure was developed, and it showed that both intra-rater and inter-rater agreement were excellent for the analytical method that might overcome errors and subjectivity in evaluation that result from the limitations of global method. Results recommend suitability of using the analytical evaluation to improve reliability between raters.

Keywords: Complete denture, Delivery, Intra-rater agreement, Inter-rater agreement, Variability

1. Introduction

Consistency in preclinical or clinical evaluation displays disagreeable matters to prosthodontics staff. Any lack of evaluation reliability can similarly be a source of misunderstanding and pressure for dental students. Removable

Corresponding author:
Associate Professor Dr. Manal Alammari, Oral and Maxillofacial Rehabilitation Department, Faculty of Dentistry, King Abdulaziz University, P.O.Box 80209, Jeddah, 21589, Kingdom of Saudi Arabia.
Tel: +966(2)6403443 Ext: 23273, and +966536111149, Email: malammari@kau.edu.sa
Received: July 10, 2018, Accepted: July 29, 2018, Published: September 2018
iThenticate screening: July 19, 2018, English editing: August 12, 2018, Quality control: August 15, 2018
This article has been reviewed / commented by three experts
Ethics approval: REC 091-10-17 (King Abdulaziz University, Saudi Arabia)

complete prostheses are increasingly being used in everyday prosthodontics clinics in undergraduate dental education. Clinical prosthodontics is often concerned not only with knowledge gaining, but also with attainment of skills and their application (1). To achieve this, a reliable objective evaluation tool and criteria are required. The nature of clinical supervision, academic teaching and dental skills development in undergraduate dentistry require competent faculty calibration and agreement. Issues in agreement and calibration have been recognized to be a cause of misperception and misunderstanding among dental students, which might affect their knowledge and acceptance of information. (2, 3). Lanning and colleagues reported that dental students thought that variations and discrepancies in clinical evaluation between their supervisors and evaluators affected their clinical development (4). Moreover, it was found that inconsistency between evaluators and clinical faculty staff disturbs the students' confidence in their evaluators' feedback, which in turn affects their tendency to progress and to attain better marks (5). Furthermore, these variations in clinical evaluation on a daily basis, might affect the dental students' understanding of what standards of clinical skills and capability they need to achieve, to recognize and to amend (2). Schönwetter and coworkers (6) determined that bias associated with clinical assessment of dental student performance remains a cause of hindrance for both dental students as well as clinical evaluators.

Investigators in more recent years have concentrated on the development of marking systems centered on specific criteria and checklists as an alternative to the glance and grade method commonly used in clinical sessions, in order to improve rater performance, but the results have been vague (7). Some researchers found that development of an analytical approach using detailed checklists improved examiner reliability (8, 9). However, other investigators reported no difference between glance and grade and checklist methods of assessment (10). It has been resolved that improved staff calibration and a more inclusive system of evaluation are required to reduce the tricky issue of discrepancy during evaluation of dental students' work (11). Consequently, to encourage a well-informed system of teaching, an objective and reliable method of evaluation is vital (12).

On reviewing the literature, there is a lack of data concerning evaluation and grading of clinical procedures for construction of removable prostheses. A clinical study was done lastly to compare intra and inter-rater variability by two evaluation methods for final impressions of completely edentulous patients (13). Another study was done in 2014 to compare intra- and inter-rater variability by two evaluation methods: glance ahead grade (global), and checklist and criteria (analytical) for recording jaw relations for complete denture prosthodontics (14). Many authors have reported attempts to develop reliable laboratory and clinical evaluation systems. Reliability in clinical evaluation presents a serious problem to faculty, who must provide such judgments especially in an educational environment for undergraduates. Any deficit in consistency in evaluation will lead to confusion and it will affect the educational operation for dental students. Learning results from direct exposure to events, which forms and subsequently strengthens cognitive associations (15). Therefore, the aims of this study were to determine intra-raters and inter-raters reliability and agreement by two evaluation methods: glance and grade (global) and checklist criteria (analytical), for complete denture insertion procedures in a removable prosthodontics course in 6th year undergraduate dental students. The study also aimed to determine the most effective methods conducive to practicality, time management and equality. The null hypothesis of this study was that there would be no difference in the intra- and inter-raters reliabilities of both evaluation methods.

2. Material and Methods

2.1. Research design and participants

Using a cross-sectional study, two prosthodontist examiners were enrolled, who evaluated a sample comprised of 35 complete denture cases (20 male, 15 female) selected from outpatients attending removable prosthodontics clinics at the Faculty of Dentistry, King Abdulaziz University Dental Hospital, Jeddah, Saudi Arabia. The study was conducted from December 2017 until April 2018. Regarding the eligibility criteria, any patients who fitted the following criteria and agreed to participate in the study were included: healthy oral mucosa, normally formed ridges, Class I jaw relation and no history of temporomandibular joint dysfunction.

2.2. Examiners and the evaluations

Two evaluators were faculty members, both with master's degrees and PhDs in prosthodontics and had been practicing and teaching for more than fifteen years. Examiners evaluated the prosthodontic procedures per¬formed on completely edentulous dental patients by thirty-five dental students in prosthodontics clinical sessions at the delivery (insertion) visit. Two prosthodontists evaluated the work separately, and each procedure was given a score on a 1 to 10 scale. Two sessions of calibration by both evaluators included a real complete denture patient in 4th year prosthodontics clinic in order to simulate the environment and to standardize the process. Upper and lower

complete dentures were constructed for each patient following standardized clinical and laboratory procedures by final year undergraduate dental students. Each delivered denture was given a number code and graded independently by faculty staff members (A and B) with a ten-point scale using the eye balling (glance and grade) method. After completion of the first evaluation, the examiners used the analytical method of grading. They agreed upon certain standards for denture insertion (Table 1) and sufficient training of its (analytical) method after researchers formulated their own criteria based on the requirements and the required factors needed to be present in complete denture at the delivery and insertion visit. After denture use for three days, each patient was recalled for evaluation of the denture by the same faculty staff using the same two evaluation methods. The second exposure evaluation was performed and graded using the same ten-point scale.

Table 1. The form used in analytical evaluation with its components criteria and scores for each criteria and step of evaluation.

Criteria		Grade given by examiner
Maxillary complete denture	Extensions (out of 2)	Check Labial extension (0.5)
		Check Buccal extension (0.5)
		Check Posterior extension (0.5)
		Check palatal extension (0.5)
	Retention (out of 0.5)	Apply a firm downward vertical pull if it is satisfactory (0.5)
	Stability (out of 0.5)	Apply pressure in tissue ward direction in premolar and molar region if it is satisfactory (0.5)
Mandibular complete denture	Extension (out of 2)	Labial (0.5)
		Buccal (0.5)
		Posterior (0.5)
		Lingual (0.5)
	Retention (out of 0.5)	Apply pull on lower anterior teeth straight upwards (0.25)
		Tilt denture outward from canine region to test the retention of opposite retro molar pad if it is satisfactory (0.25)
	Stability (out of 0.5)	Apply pressure in tissue ward direction in premolar and molar region if it is satisfactory (0.5)
For Both maxillary and Mandibular dentures	Centric Occluding relations (out of 2)	Vertical (0.5)
		Horizontal (0.5)
		Evenness of occlusion (1)
	Artificial teeth (out of 2)	Shape (0.5)
		Shade (0.5)
		Size (0.5)
		Arrangement (0.5)
Total	Grade out of 10	

2.3. Statistical methods

This study was analyzed using IBM© SPSS© Statistics version 23 (IBM© Corp., Armonk, NY, USA). A simple descriptive statistic was used to define the characteristics of the study; continuous variables are presented by mean and standard deviations. To compare the means of two variables, a paired-samples t-test was used. These tests were done with the assumption of normal distribution. Otherwise, to compute the differences between the two continuous variables for all cases and classify the differences, a Wilcoxon signed-rank test was used. Lastly, agreement was measured by weighted kappa.

2.4. Ethics of research

To conduct the study, an ethical approval was obtained from the Research Ethics Committee of Faculty of Dentistry, King Abdulaziz University (Ref: REC 091-10-17). After explaining the objective of the study verbally to the patients, a verbal patient's consent was obtained from each of them, as they were elderly and unable to read the information sheet or write.

3. Results

Characteristics of the study sample are presented in table 2, which shows that no full marks were given. In addition, the global method produced higher grades. Table 3 shows that the first and second exposures agreement under global and analytical with Examiner A is excellent with 90.7% and 99.8% agreement respectively. While under

Examiner B, global was lower with about 78.1% and 96.1% in analytical. There was high agreement between the two raters in the first exposure in analytical (97.3%). This trend continues with analytical in the second exposure with 93.2%, however, in the second exposure under global it failed the cut off, which is only 56.6% agreement. Regarding the evaluation of inter-rater agreement, the first observations demonstrated homogeneity among the examiners in global evaluation (p=0.0581). However, for the second observations, inconsistency was observed (p=0.002). Moreover, in the analytical method, it was more homogenous as shown in Table 4. The intra-rater variability for the glance and grade or the criteria and checklist methods of evaluation were measured using the paired sample t-test. The study showed sig¬nificant difference in examiner A in the global method (p<0.001) (Table 5). The inter-rater variability tests were mea¬sured using Wilcoxon signed rank test. As displayed in Table 6, for most of the measurements, there was a non-significant difference among the evaluators except for the global method for evaluator A (p=0.001).

Table 2. Characteristics of the 35 Study Samples

Examiner	Method of evaluation	Exposure	n	Min	Max	Mean	SD
Examiner A	Global Method	First exposure	35	5.00	8.50	6.99	.9
		Second Exposure	35	6.00	9.50	7.64	.9
	Analytical Method	First exposure	35	4.25	8.25	6.60	1.2
		Second Exposure	35	4.50	8.00	6.61	1.2
Examiner B	Global Method	First exposure	35	6.00	9.00	6.93	.9
		Second Exposure	35	6.50	9.00	7.14	.7
	Analytical Method	First exposure	35	5.00	8.25	6.76	1.0
		Second Exposure	35	5.00	8.25	6.79	1.0

Table 3. Inter and intra-raters agreement.

Agreement	Examiner and exposure	Global Evaluation (Glance & Grade)		Analytical Evaluation (criteria &checklist)	
		Correlation	p-value *	Correlation	p-value *
Intra-rater agreement	Examiner A	0.907	<0.001	0.998	<0.001
	Examiner B	0.781	<0.001	0.961	<0.001
Inter-rater (Between Examiner A and Examiner B)	First Exposure	0.875	<0.001	0.973	<0.001
	Second Exposure	0.566	0.009	0.932	<0.001

* Paired-samples t-test was used

Table 4. Difference between first and second exposure in both evaluation methods.

Variables	Global Method (Mean ± SD)		Analytical Method (Mean ± SD)	
	First exposure	Second Exposure	First exposure	Second Exposure
Examiner A	6.99±0.9	7.64±0.9	6.60±1.2	6.61±1.2
Examiner B	6.93±0.9	7.14±0.7	6.76±1.0	6.79±1.0
p-value *	0.581	0.002	0.010	0.050

* Paired-samples t-test was used

Table 5. Intra-rater's variability.

Examiners	Methods	Exposures	Mean ± SD	p-value *
Examiner A	Global Method	First exposure	6.99±0.9	<0.001
		Second Exposure	7.64±0.9	
	Analytical Method	First exposure	6.60±1.2	0.711
		Second Exposure	6.61±1.2	
Examiner B	Global Method	First exposure	6.93±0.9	0.075
		Second Exposure	7.14±0.7	
	Analytical Method	First exposure	6.76±1.0	0.677
		Second Exposure	6.79±1.0	

* Paired-samples t-test was used

Table 6. Inter-rater's Variability.

Comparison of exposures	p-value	
	Examiner A	Examiner B
First exposure Global Method vs. Second Exposure Global Method	<0.001 *	0.056
First exposure Analytical Method vs. Second Exposure Analytical Method	0.705	0.751

* Wilcoxon signed-rank test was used.

4. Discussion

In this study, reliability refers to an indication of the consistency of scores between raters and for the same rater at different times. It has been found that the glance and grade method failed the agreement. As Chambers and coworkers revealed, assessors do not consider the glance and grade method, which is commonly used for evaluation in dental clinics, as a perfect system, nevertheless, it is widely used (16). Paskins et al., assessed the use of a criterion based checklist in which two assessors used the tool and showed inter-rater agreement of >0.9 (17), which is similar to the result of this study with analytical evaluation agreement >0.93. In this study, the variability was from 3 to 4 marks and full marks were still not awarded (18). Moreover, using the analytical method had less in variability among evaluators. However, the global method in the current study findings indicated that the problem in discrepancy of evaluating undergraduate dental students for completely edentulous patients in removable prosthodontics exists. For the two examiners, this agreement may be considered highly acceptable, because it exceeded 90%. Similarly, Goepferd and Kerber (8) used the analytical system for evaluation using specific standards. They reported that the method was superior to the glance and grade method in reducing the variability among the two examiners. Our results agreed with their findings and other recent study findings (13, 14). Therefore, the analytical method was assumed to be reliable since one examiner performing multiple evaluation could obtain high agreement, and different evaluators acquired comparable evaluation values.

It was found that global reading was higher than analytical reading for both examiners. Our results did not agree with the work of Vann and coworkers (10) who reported that no method of evaluation resulted in more reliability between examiners. In many dental schools and due to practical conditions, the glance and grade method is still functional (12). It is vital to develop a practical reproducible, easily applicable method to evaluate the denture insertion procedure, as this step has a major impact on the patients' satisfaction. On the other hand, it is anticipated that by using the analytical method, it might help the student to learn the important aspects of a finished complete denture and what it should fulfill. Nevertheless, the analytical method needs staff training, as suggested by Satterthwaite and Grey (18). Lastly, examiners' reliability is essential in the teaching and learning development of the undergraduate dental students. Consequently, new evaluation methods and techniques of unvarying evaluation need to be executed to endorse a well-organized scheme of evaluation. Regarding the study limitations, we acknowledge that further research including validation of the developed analytical method for evaluation of the delivery step in complete denture construction process is needed.

5. Conclusions

Results showed that both intra-rater and inter-rater agreement were excellent for the analytical method that might overcome errors and subjectivity in evaluation that result from the limitations of the global method. Results of this study indicated that the analytical method of evaluation might provide more accurate, reliable and unbiased evaluation. This study also provided an example of how analytical evaluation in denture delivery procedure can be applied to prosthodontic dentistry. Finding a method like the analytical explained in this paper is highly needed, as it showed that it is an agreeable method that satisfies the practicality. It is important to develop a reproducible, and easily applicable evaluation method to correctly evaluate complete denture constructions steps in the clinic. Consequently, improved staff training and standardization are required for better evaluation.

Acknowledgments:

This study was an original research idea. The researchers would like to thank the authorities, the students and the patients who agreed to participate and had enough patience in this research, and all those who had helped in the various stages of this research at King Abdulaziz University Dental Hospital, Jeddah, Saudi Arabia. The authors received no financial support for the research, authorship, and/or publication of this article.

Conflict of Interest:

There is no conflict of interest to be declared.

Authors' contributions:
Both authors contributed to this project and article equally. Both authors read and approved the final manuscript.

References:

1) Douglass CW, Watson AJ. Future needs for fixed and removable partial dentures in the United States. J Prosthet Dent. 2002; 87: 9-14. doi: 10.1067/mpr.2002.121204. PMID: 11807477.
2) Jacks ME, Blue Ch, Murphy D. Short-and long-term effects of training on dental hygiene faculty members' capacity to write SOAP notes. J Dent Educ, 2008; 72(6): 719-24. PMID: 18519602.
3) Quinn F, Keogh P, McDonald A, Hussey D. A study comparing the effectiveness of conventional training and virtual reality simulation in the skills acquisition of junior dental students. Eur J Dent Educ. 2003; 7(4): 164-9. doi: 10.1034/j.1600-0579.2003.00309.x. PMID: 14753762.
4) Lanning SK, Pelok SD, Williams BC, Richards PS, Sarment DP, Oh TJ, et al. Variation in periodontal diagnosis and treatment planning among clinic instructors. J Dent Educ. 2005; 69(3): 325-7. PMID: 15749943.
5) Haj-Ali R, Feil R. Rater reliability: Short-and long-term effects of calibration training. J Dent Educ. 2006; 70 (4): 428-33. PMID: 16595535.
6) Schönwetter DJ, Lavigne S, Mazurat R, Nazarko O. Students' perceptions of effective classroom and clinical teaching in dental and dental hygiene education. J Dent Educ. 2006; 70(6): 624-35. PMID: 16741130.
7) Zawawi KH, Afify AR, Yousef MK, Othman HI, Al-Dharrab AA. Reliability of didactic grades to predict practical skills in an undergraduate dental college in Saudi Arabia. Adv Med Educ Pract. 2015; 6: 259–63. doi: 10.2147/AMEP.S72648. PMID: 25878519, PMCID: PMC4386792.
8) Goepferd SJ, Kerber PE. A comparison of two methods for evaluating primary Class II cavity preparations. J Dent Educ. 1980; 44(9): 537-42. PMID: 6931147.
9) Schmitt L, Möltner A, Rüttermann S, Gerhardt-Szép S. 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 J Med Educ. 2016; 33(4): 1-19. doi: 10.3205/zma001060. PMID: 27579361, PMCID: PMC5003144.
10) Philips Z, Ginnelly L, Sculpher M, Claxton K, Golder S, Riemsma R, et al. Review of guidelines for good practice in decision-analytic modelling in health technology assessment. Health Technol Assess. 2004; 8(36): iii-iv, ix-xi, 1-158. PMID: 15361314.
11) Sherwood IA, Douglas GV. A study of examiner variability in assessment of preclinical class II amalgam preparation. Journal of education and ethics in dentistry; 2014; 4(1); 12-7.
12) Sharaf AA, Abdel-Aziz MM, EI-Meligy OA. Intra-and inter-examiner variability in evaluating preclinical pediatric dentistry operative procedures. J Dent Educ. 2007; 7(4): 540-4.
13) Alammari MR, Alkhiary YM, Nawar AA. Intra-and inter-examiner variability in evaluating impression procedures at the undergraduate level. J Life Sci. 2013; 5(1): 5-10. doi: 10.1080/09751270.2013.11885208.
14) Al-Dharrab AA, Alammari MR, AlKhiary YM, Nawar EA. Intra-and inter-examiner variability in evaluating jaw relation records for complete denture prosthodontics. Wulfenia J. 2014; 21(7): 311-7.
15) Walstead BK. Faculty Perceptions Regarding Best Practices in Clinical Dental Hygiene Assessment. Walden dissertation and doctoral studies collection. Walden University; 2015.
16) Chambers WA, Geissberger M, Leknius C. Association amongst factors thought to be important by instructors in dental education and perceived effectiveness of these instructors by students. Eur J Dent Educ. 2004; 8(4): 147-51. doi: 10.1034/j.1600-0579.2003.00324.x-i1. PMID: 15469440.
17) Paskins Z, Kircaldy J, Allen M, Macdougall C, Fraser I, Peile E. Design, validation and dissemination of an undergraduate assessment tool using Sim Man in simulated medical emergencies. Med Teach. 2010; 32(1): e12-7. doi: 10.3109/01421590903199643. PMID: 20095761.
18) Satterthwaite JD, Grey NJ. Peer‐group assessment of pre‐clinical operative skills in restorative dentistry and comparison with experienced assessors. Eur J Dent Educ. 2008; 12(2): 99-102. doi: 10.1111/j.1600-0579.2008.00509.x. PMID: 18412738.