Intra- and Inter-Examiner Variability in Evaluating Preclinical Pediatric Dentistry Operative Procedures

Aly A. Sharaf, B.D.S., M.Sc., Ph.D.; Amr M. AbdelAziz, B.D.S., M.Sc., Ph.D.; Omar A.S. El Meligy, B.D.S., M.Sc., Ph.D.

Abstract: Many investigators have reported attempts to develop reliable laboratory and clinic evaluation systems. However, few studies, regardless of level of success, have used an analytic procedure to identify those components of the evaluation system that, if refined further, could improve reliability. The purpose of this study was to compare intra- and inter-examiner variability in two evaluation methods: glance and grade (global), and checklist and criteria (analytical). Three faculty staff members with more than ten years of clinical and teaching experience evaluated operative procedures performed on plastic teeth representing the primary teeth by thirty dental students in pediatric dentistry preclinical laboratory sessions. The preparations were graded blindly by each of the three evaluators (A, B, and C) three times without magnification. The values were statistically analyzed using Wilcoxon signed rank test and Friedman test setting value of significance at 5 percent. The study revealed that, among the three examiners, the intra-examiner variability was nonsignificant in most situations. On the other hand, there was statistically significant variability between evaluators (i.e., inter-examiner) for almost all preparations. Neither cutting off the scores nor using either evaluation method (glance and grade or criteria and checklist) caused an improvement in variability. The problem of inter-examiner reliability and variability still existed.

Dr. Sharaf is Professor, Pediatric Dentistry, King AbdulAziz University, and Professor, Pediatric Dentistry, Alexandria University; Dr. AbdelAziz is Associate Professor, Pediatric Dentistry, King AbdulAziz University, and Assistant Professor, Pediatric Dentistry, AinShams University; and Dr. El Meligy is Associate Professor, Pediatric Dentistry, King AbdulAziz University, and Assistant Professor, Pediatric Dentistry, Alexandria University. Direct correspondence and requests for reprints to Dr. Omar A.S. El Meligy, Preventive Dental Sciences Department, Pediatric Dentistry Division, Faculty of Dentistry, King AbdulAziz University, P.O. Box 80209, Jeddah, Saudi Arabia 21589; 00966-2-6402000 phone; 00966-2-6403316 fax; oelmeligy@dataxprs.com.eg.

Key words: variability, preclinical operative procedures, evaluation methods

Submitted for publication 9/11/06; accepted 11/29/06

Reliability in preclinical or clinical evaluation presents serious problems to faculty who must render such judgments, and any lack of evaluation consistency can also be a source of confusion and stress for dental students. This problem was recognized as early as 1930 by Brown,[1] but the subject received little attention in the dental literature before 1970. Mackenzie's 1973 recommendations for clinical dental education stimulated interest and research on this subject,[2] and since that time, the subject has received increased visibility in the dental literature. However, after a comprehensive review of the literature in 1977, Myers[3] concluded that subjectivity associated with clinical evaluation of student performance remains a source of frustration for both dental students and clinical instructors. In short, the problem still existed.

Lilley et al.[4] and Fuller[5] found not only significant disagreement between examiners, but also wide intra-examiner variation when the same rater evaluated the same operative procedure on a second occasion. Similarly, Salvendy et al.[6] evaluated Class I amalgam preparations and found a high degree of both intra- and inter-examiner variation. The results of this investigation led the authors to encourage the development of more objective evaluation methods, such as optical scanners and electronic devices, to accurately measure cavity dimensions. Jenkins et al.[7] 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 pass on one occasion and a fail on another and vice versa. Worried by the extent of the problem of examiner consistency, Schiff et al.[8] 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 of benefit 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.

Investigations 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 in order 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.[9,10] However, other investigators reported no difference between glance and grade and checklist methods of assessment.[11] Patridge and Mast[12] emphasized the paucity of controlled research in dental clinical evaluation and pointed out that very few studies provide comparative research data on different evaluation methods. These authors and other investigators recommended using more frequent and uniform training sessions to improve evaluator reliability.[12-15] Feil[16] analyzed the reliability of a laboratory evaluation system. An analytic technique for identifying components contributing to the reliability of the evaluation system was also described. This study demonstrated that reliability can be increased through the use of two raters as opposed to the traditional use of only one. Deranleau et al.[17] introduced several modifications in a criterion-referenced evaluation system to determine which modification would result in optimal evaluator reliability and performance differentiation on two representative wax-up projects. Variables investigated included the use of percentages as cut-off scores and the number of scoring options. The use of percentage cut-off scores to define minimal competency significantly increased intra-judge reliability for one of the two projects. Two-option scoring resulted in significantly higher rater agreement than the three-option scoring in half of the comparisons. Jenkins et al.[7] concluded that better staff training and a more comprehensive system of assessing preclinical skills are needed to address the problem of inconsistency among dental faculty who evaluate students' work.

Problems with examiner consistency may lead students to perceive that evaluation methods are somewhat arbitrary. This concept can undermine the learning process and produce a negative effect on undergraduate confidence and performance. A method of assessment that is both objective and reliable is essential, therefore, to promote an efficient system of learning and to reduce friction between students and faculty over the issue of grading. This study was conducted to increase our knowledge of factors that contribute to evaluation in preclinical laboratory courses when using two evaluation strategies. Intra- and inter-examiner variability was compared in two evaluation methods: glance and grade (global), and checklist and criteria (analytical).

Materials and Methods

We evaluated the operative procedures performed on plastic primary teeth by thirty dental students in pediatric dentistry preclinical laboratory sessions. Three authors evaluated the work separately, and each procedure was given a score on a 1 to 10 scale. The three evaluators were faculty members with both master's degrees and Ph.D.s in pediatric dentistry and have been practicing and teaching pediatric dentistry for more than ten years.

The dental phantom head laboratory was a simulator resembling the real clinical situation. The plastic teeth were fixed in a typodont and mounted in a phantom head simulator. The students used bur #330 on a high-speed handpiece or small round burs on a slow-speed handpiece to perform the procedures. Students were sitting during the procedure using a unit light source similar to the clinical situation, water cooling, and suction; whenever necessary as in upper teeth, indirect vision was applied through a dental mirror.

The procedures that were evaluated were:

- Class I on a lower primary molar (ILD)
- Class I on an upper primary molar (IUD)
- Modified Class I with a palatal extension (finger preparation) on an upper second primary molar (IUE)
- Class II occluso distal on a lower first primary molar (IILD)
- Class II occluso mesial on a lower second primary molar (IILE)
- Class III mesial slot on an upper primary central incisor (IIIS)
- Modified distal Class III with a lingual dove tail on a lower primary canine (IIIM)
- Class V on an upper primary central incisor (V)

The work of the students was collected after each session and was given a number code. The preparations were graded blindly by each of the three investigators (A, B, and C) three times without magnification. For the first evaluation, each author graded the preparation with the ten-point scale using the eyeballing (glance and grade) method. After three days, the work was reevaluated again using the same method to measure intra-examiner variability.

After completion of the first two evaluations, we agreed upon certain criteria for each preparation. The criteria for each cavity preparation were developed from the students' clinical manual to determine cavity dimensions as length, width, and depth and the cavity shape as extension, roundation (absence of sharp line angles), centralization, cavity margins, and undercuts. Using the criteria and checklist, together with an explorer to verify cavity form and dimensions, the third evaluation was performed blindly and graded using the same ten-point scale. The values were tabulated, and statistical analysis was performed using SPSS package version 10.0 to test the intra- and inter-examiner variability among the three examiners. Statistical analysis was done using Wilcoxon signed rank test and Friedman test setting value of significance at 5 percent. Z value was the calculated statistic that was compared with the tabulated Zα, and the P value was used to indicate statistical significance. After completing the analysis, the data were rearranged in a percentage pattern to cut off scores, and the tests were run again. The new values represented only five grades—A, B, C, D, and E—instead of ten and represented excellent, very good, good, acceptable, and failed scores, respectively.

Results

The intra-examiner variability tests were measured using Wilcoxon signed rank test at 0.05 level of significance. As displayed in Table 1, for most of the measurements, there was a nonsignificant difference among the evaluators except for IUD for evaluator A (P=0.008), IIIS, IIIM, and V for evaluator B (P=0.00, 0.00, and 0.001, respectively), and finally IILD for evaluator C (P=0.047). When the scores were cut off into grades, the values remained significant (Table 2).

The inter-examiner variability for the glance and grade or the criteria and checklist methods of evaluation were measured using the Friedman test at 0.05 level of significance. Table 3 shows a significant difference among examiners in all preparations except for IUE (P=0.091) and IIIM (P=0.076) for the glance and grade and criteria and checklist,

Table 1. Intra-examiner variability shown by Z and P values of Wilcoxon signed rank test using glance and grade method of evaluation and grading in a scale from 1 to 10

Preparation	Examiner A		Examiner B		Examiner C	
	Z value	P value	Z value	P value	Z value	P value
ILD	1.084	0.278	0.637	0.524	1.526	0.127
IUD	2.64	0.008*	0.267	0.79	0.767	0.443
IUE	0.836	0.403	0.082	0.935	1.938	0.053
IILD	0.041	0.967	1.095	0.273	1.987	0.047*
IILE	0.942	0.346	0.957	0.339	1.025	0.305
IIIS	1.853	0.064	3.886	0.000*	0.174	0.862
IIIM	0.161	0.872	4.037	0.000*	1.08	0.28
V	0.099	0.921	3.345	0.001*	1.382	0.167

*=significant variability

Table 2. Intra-examiner variability shown by Z and P values of Wilcoxon signed rank test using glance and grade method of evaluation and cutting off scores into five grades

Preparation	Examiner A		Examiner B		Examiner C	
	Z value	P value	Z value	P value	Z value	P value
ILD	0.735	0.462	0.44	0.66	1.405	0.16
IUD	2.242	0.025*	0.198	0.843	0.55	0.583
IUE	1.091	0.275	0.3	0.765	1.822	0.068
IILD	0.034	0.973	1.165	0.244	1.968	0.049*
IILE	1.232	0.218	0.894	0.371	0.619	0.536
IIIS	1.602	0.109	3.8	0.000*	0.915	0.36
IIIM	0.44	0.66	3.923	0.000*	1.41	0.159
V	1.069	0.285	3.419	0.001*	1.502	0.133

*=significant variability

respectively. When the five-point (A, B, C, D, E) grading system was applied, the same preparations (IUE and IIIM) were the only preparations where the grades assigned by evaluators were not significantly different (P=0.116 and 0.067, respectively) (Table 4). Table 5 shows results obtained by Wilcoxon signed rank test to measure the inter-examiner variability between each of two examiners separately. In almost all preparations, there was a significant disagreement between at least two examiners in both evaluation methods.

Table 3. Inter-examiner variability comparing glance and grade method to criteria and checklist method applying Friedman test and using a grading scale from 1 to 10

Preparation	Glance and Grade		Criteria with Checklist	
	X^2	P	X^2	P
ILD	20.109	0.000*	11.176	0.004*
IUD	16.065	0.000*	7.788	0.02*
IUE	4.795	0.091NS	7.4	0.025*
IILD	13.609	0.001*	7.271	0.026*
IILE	16.289	0.000*	11.732	0.003*
IIIS	33.771	0.000*	22.107	0.000*
IIIM	19.898	0.000*	5.154	0.076NS
V	11.791	0.003*	6.5	0.039*

*=significant variability
NS=nonsignificant variability

Table 4. Inter-examiner variability comparing glance and grade method to criteria and checklist method applying Friedman test and cutting off scores into five grades

Preparation	Glance and Grade		Criteria with Checklist	
	X^2	P	X^2	P
ILD	18.381	0.000*	11.247	0.004*
IUD	14.296	0.001*	7.525	0.023*
IUE	4.314	0.116NS	6.442	0.04*
IILD	16.88	0.000*	7.487	0.024*
IILE	19.923	0.000*	9.811	0.007*
IIIS	39.057	0.000*	21.189	0.000*
IIIM	24.352	0.000*	5.4	0.067NS
V	15.462	0.000*	9.528	0.009*

*=significant variability
NS=nonsignificant variability

Discussion

The results of this study support the notion of inconsistency among examiners in evaluating the preclinical performance of students. There was a significant inter-examiner variability in our work as seen in Table 3. This supports the findings of Lilley et al.,[4] Fuller,[5] Salvendy et al.,[6] and Jenkins et al.[7] But again our results are different from their results as our work found nonsignificant intra-examiner variations in most preparations opposed to their conclusions that found significant intra-examiner variability.

In an attempt to reduce variability among examiners, Geopferd and Kerber[9] 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. Our results, however, did not agree

Table 5. Inter-examiner variability between each of two examiners using either glance and grade method or criteria and checklist method as shown by Z and P values of Wilcoxon signed rank test

Preparation		Glance and Grade			Criteria with Checklist		
		A vs B	A vs C	B vs C	A vs B	A vs C	B vs C
ILD	Z	3.166	3.742	0.645	0.314	2.409	2.545
	P	0.002*	0.000*	0.519NS	0.754NS	0.016*	0.011*
IUD	Z	2.971	3.335	0.998	2.326	0.678	2.487
	P	0.003*	0.001*	0.318NS	0.02*	0.498NS	0.013*
IUE	Z	0.489	2.117	1.385	2.033	1.041	2.7
	P	0.625NS	0.034*	0.166NS	0.042*	0.298NS	0.007*
IILD	Z	3.659	0.175	3.214	2.443	0.232	2.334
	P	0.000*	0.861NS	0.001*	0.015*	0.817NS	0.02*
IILE	Z	3.187	1.489	3.612	3.403	0.698	2.916
	P	0.001*	0.136NS	0.000*	0.001*	0.485NS	0.004*
IIIS	Z	4.004	0.996	4.494	3.743	3.636	1.498
	P	0.000*	0.319NS	0.000*	0.000*	0.000*	0.134NS
IIIM	Z	3.742	0.818	4.084	2.061	0.049	2.687
	P	0.000*	0.413NS	0.000*	0.039*	0.961NS	0.007*
V	Z	2.354	1.716	3.251	1.217	0.687	2.162
	P	0.019*	0.086NS	0.001*	0.223NS	0.492NS	0.031*

*=significant variability
NS=nonsignificant variability

with their report, as seen in Tables 3 and 4, that show a similar pattern of disagreement among examiners in both evaluation methods. Our results agree with the work of Vann et al.,[11] who reported that no method resulted in superior inter-examiner reliability.

In another attempt to reduce variability, we experimented on cutting off scores using percentages and a grading system of only five grades instead of a scale from 1 to 10, as represented in Tables 2 and 4. When Tables 2 and 4 were compared to Tables 1 and 3 where the 1 to 10 scale was used, the variability remained the same. This disagrees with the work of Deranleau et al.,[17] who reported a significantly increased inter-examiner reliability by applying the percentage cut-off scores. Our results reinforce the problem that was first reported by Brown[1] in 1930 and supports the conclusion of Myers[3] in 1977 on the subjectivity of evaluation methods.

In many teaching institutions and due to practical situations, the glance and grade method is still applied especially with more experienced faculty staff. Some authors have proposed different evaluation techniques; for example, Schiff et al.[8] used a pulpal floor measuring instrument. Although Schiff et al.'s technique may provide some advantages, it is not suitable for evaluating all aspects of a cavity, and it is not convenient to implement. It is important to develop a practical, reproducible, and easily applicable method to accurately measure cavity margins as suggested by Salvendy et al.[6]

Other alternative methods recommended include better staff training and developing a more comprehensive system for evaluation as suggested by Jenkins et al.[7] or obtaining two or more assessments provided by at least two evaluators and calculating an average as recommended by Feil.[16] The application of more frequent and uniform training sessions to improve evaluator reliability was recommended by other authors.[12-15] Finally, the examiner consistency is crucial in the teaching and learning process as 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.

Conclusions

Among the three examiners, the level of intra-examiner variability was not statistically significant for most of the preparations. On the other hand, there were statistically significant differences among evaluators for almost all preparations. Neither cutting off the scores nor using either evaluation methods (glance and grade or criteria and checklist) caused an improvement in variability. These findings indicate that the problem of inter-examiner reliability and variability still exists. Further research and improvement in this area are highly needed.

REFERENCES

1. Brown RK. Research in the use of a rating scale as a means of evaluating the personalities of senior dental students. J Dent Res 1930;10(3):271-9.
2. Mackenzie RS. Defining clinical competence in terms of quality, quantity, and need for performance criteria. J Dent Educ 1973;37(9):37-44.
3. Myers B. Beliefs of dental faculty and students about effective teaching behaviors. J Dent Educ 1977;41(2): 68-76.
4. Lilley JD, Bruggen Cate HJ, Holloway PJ, Holt JK, Start KB. Reliability of practical tests in operative dentistry. Br Dent J 1968;125(5):194-7.
5. Fuller JL. The effects of training and criterion models on inter-judge reliability. J Dent Educ 1972;36(4):19-22.
6. Salvendy G, Hinton WM, Ferguson GW, Cunningham PR. Pilot study on criteria in cavity preparation. J Dent Educ 1973;37(10):27-31.
7. 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.
8. 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.
9. Goepferd SJ, Kerber PE. A comparison of two methods for evaluating primary Class II cavity preparations. J Dent Educ 1980;44(9):537-42.
10. Dhuru VB, Rypel TS, Johnston WM. Criterion-oriented grading system for preclinical operative dentistry laboratory course. J Dent Educ 1978;42(9):528-31.
11. Vann WF, Machen JB, Hounshell PB. Effects of criteria and checklists on reliability in preclinical evaluation. J Dent Educ 1983;47(10):671-5.
12. Patridge MI, Mast TA. Dental clinical evaluation: a review of the research. J Dent Educ 1978;42(6):300-5.
13. Houpt MI, Kress G. Accuracy of measurement of clinical performance in dentistry. J Dent Educ 1973;37(7): 34-46.
14. Hinkelman KW, Long NK. Method for decreasing subjective evaluation in preclinical restorative dentistry. J Dent Educ 1973;37(9):13-8.
15. Abou-Rass M. A clinical evaluation instrument in endodontics. J Dent Educ 1973;37(9):22-36.
16. Feil PH. An analysis of the reliability of a laboratory evaluation system. J Dent Educ 1982;46(8):489-94.
17. Deranleau NJ, Feiker JH, Beck M. Effect of percentage cut-off scores and scale point variation on preclinical project evaluation. J Dent Educ 1983;47(10):650-5.