Faculty Standardization of Dental Hygiene Instrumentation Skill Evaluation: A Calibration Exercise

Diane-Marie Smela

Old Dominion University

Follow this and additional works at: https://digitalcommons.odu.edu/dentalhygiene_etds

Part of the Dental Hygiene Commons, Higher Education Commons, and the Medical Education Commons

Recommended Citation
https://digitalcommons.odu.edu/dentalhygiene_etds/98

This Thesis is brought to you for free and open access by the Dental Hygiene at ODU Digital Commons. It has been accepted for inclusion in Dental Hygiene Theses & Dissertations by an authorized administrator of ODU Digital Commons. For more information, please contact digitalcommons@odu.edu.
FACULTY STANDARDIZATION OF DENTAL HYGIENE

INSTRUMENTATION SKILL EVALUATION:

A CALIBRATION EXERCISE

by

Diane-Marie Smela
B.S.D.H., May 1989, Old Dominion University
A.A.S., May 1986, Middlesex County College

A Thesis Submitted to the Faculty of
Old Dominion University in Partial Fulfillment of the
Requirement for the Degree of

MASTER OF SCIENCE

DENTAL HYGIENE

OLD DOMINION UNIVERSITY
December, 1991

Approved by:

Deborah Blythe Bauman (Director)

Deanne Shuman

Lynn Pelle-Watts
ABSTRACT
FACULTY STANDARDIZATION OF DENTAL HYGIENE INSTRUMENTATION SKILL EVALUATION:
A CALIBRATION EXERCISE
Diane-Marie Smela
Old Dominion University
Director: Deborah Blythe Bauman

Faculty must be standardized so that student skills are assessed in an unbiased, equal fashion. The purpose of this study was to determine the impact of calibration on dental hygiene faculty standardization. Examples of correct and incorrect instrumentation skill performance were demonstrated and discussed during the calibration exercise. Groups simultaneously viewed instrumentation skills while assessing correct or incorrect performance. Group standardization was determined through analysis of answer key agreement, intra-rater reliability, and inter-rater reliability. Data were statistically analyzed through the use of a z-test for both answer key agreement and intra-rater reliability, and chi-square ($\chi^2$) for inter-rater reliability ($\alpha=0.05$). Calibrating dental hygiene evaluators did not have statistically significant effects on their ability to achieve answer key agreement or intra-rater reliability. Significant results of $\chi^2$ for effects of calibration on inter-rater reliability revealed the non-
calibrated group more consistent in assessing skills
($\chi^2 = 2.80; p = .0475$). Standardization as measured by answer
key agreement, intra-rater, and inter-rater reliability was
not influenced by the calibration exercise. Further
exploration of methods to improve faculty standardization
are needed to ensure that dental hygiene students receive
fair and objective skill performance evaluation.
ACKNOWLEDGEMENTS

The principal author would like to acknowledge the following individuals who gave their time and effort to assist in completing this document.

Deborah Blythe Bauman, BSDH, MS, Associate Professor, Thesis Director, for her encouragement, effort, perseverance, and guidance.

Deanne Shuman, BSDH, MS, Professor, Thesis Committee Member, for her professional expertise, patience, and guidance.

Lynn Tolle-Watts, BSDH, MS, Associate Professor, Thesis Committee Member, for her professional expertise and valuable insight.

Kristin Hamman, RDH, BS, Master’s Degree Candidate, Research Assistant, for all of her help in the organization of the mock clinical evaluation exercise, her friendship, and moral support.

Alka Drakar and David Ross, for their assistance in the statistical analysis of the results.

Andrea Evans, Academic Television Services, for her assistance in the development of the videotape and in the editing process.
All of the present and past clinical faculty of Old Dominion University's School of Dental Hygiene and Dental Assisting who participated as subjects in this investigation. This investigation would not have been possible if these individuals had not unselfishly given their time to participate.

Michele Darby, BSDH, MS, Eminent Scholar and Graduate Program Director, for providing me with opportunities, for her generosity, assistance, and for suggesting that I return to Old Dominion for graduate school.

Pamela Brangan, MSDH, MPH, Eds., Associate Professor, and her daughter Mallory for being my family away from home.

My parents, Theodore and Valentina Smela, for their emotional support and for their pride and faith in me.
TABLE OF CONTENTS

ACKNOWLEDGEMENTS ................................................................. ii
LIST OF TABLES ................................................................. vi
LIST OF FIGURES ................................................................. vii

Chapter

1. INTRODUCTION ................................................................. 1

Statement of the Problem .................................................. 3
Significance of the Problem .............................................. 3
Definitions of Terms .......................................................... 6
Assumptions ................................................................. 10
Limitations ................................................................. 10
Hypotheses ................................................................. 11
Methodology ................................................................. 12

2. REVIEW OF THE LITERATURE .............................................. 14

Faculty Standardization and Calibration ............................... 14
Models of Instructor Calibration ........................................ 25
Rater Reliability .......................................................... 36
Competency Based Evaluation .......................................... 44
Summary ................................................................. 55

3. METHODS AND MATERIALS ............................................... 57

Sample Description .......................................................... 57
Research Design .......................................................... 58

iv
<table>
<thead>
<tr>
<th>TABLE OF CONTENTS (continued)</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Methodology..................................................................................................................................................</td>
<td>59</td>
</tr>
<tr>
<td>Protection of Human Subjects..................................................................................................................</td>
<td>63</td>
</tr>
<tr>
<td>Instrumentation..........................................................................................................................................</td>
<td>64</td>
</tr>
<tr>
<td>Statistical Analysis..................................................................................................................................</td>
<td>66</td>
</tr>
<tr>
<td>4. RESULTS AND DISCUSSION.....................................................................................................................</td>
<td>69</td>
</tr>
<tr>
<td>Results.......................................................................................................................................................</td>
<td>72</td>
</tr>
<tr>
<td>Discussion...................................................................................................................................................</td>
<td>82</td>
</tr>
<tr>
<td>5. SUMMARY AND CONCLUSIONS....................................................................................................................</td>
<td>104</td>
</tr>
<tr>
<td>BIBLIOGRAPHY.............................................................................................................................................</td>
<td>110</td>
</tr>
<tr>
<td>APPENDICES................................................................................................................................................</td>
<td>114</td>
</tr>
<tr>
<td>A. Letter of Introduction..........................................................................................................................</td>
<td>115</td>
</tr>
<tr>
<td>B. Informed Consent Materials................................................................................................................</td>
<td>118</td>
</tr>
<tr>
<td>C. Instrumentation Skill Evaluation Form................................................................................................</td>
<td>121</td>
</tr>
<tr>
<td>D. Calibration Session Lesson Plan..........................................................................................................</td>
<td>123</td>
</tr>
<tr>
<td>E. Answer Keys and Descriptions of Vignettes..........................................................................................</td>
<td>129</td>
</tr>
<tr>
<td>F. Raw Data Organizational Charts..........................................................................................................</td>
<td>134</td>
</tr>
</tbody>
</table>
LIST OF TABLES

<table>
<thead>
<tr>
<th>TABLE</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Factors to Be Included in Calibration Exercises to Increase Rater Reliability</td>
<td>18</td>
</tr>
<tr>
<td>2. Clinical Teaching Experience of Subjects by Group</td>
<td>70</td>
</tr>
<tr>
<td>3. Number of Calibrated and Non-calibrated Subjects Achieving Answer Key Agreement for Individual Criteria</td>
<td>72</td>
</tr>
<tr>
<td>4. Total Answer Key Agreement Scores for the Calibrated and Non-calibrated Groups</td>
<td>74</td>
</tr>
<tr>
<td>5. Number of Calibrated and Non-calibrated Subjects Achieving Intra-rater Reliability for Individual Criteria</td>
<td>76</td>
</tr>
<tr>
<td>6. Total Intra-rater Reliability Scores for the Calibrated and Non-Calibrated Groups</td>
<td>78</td>
</tr>
<tr>
<td>7. Calibrated and Non-calibrated Subjects’ Responses to Criteria Items</td>
<td>80</td>
</tr>
<tr>
<td>8. Chi-square Analysis for Inter-rater Reliability</td>
<td>83</td>
</tr>
</tbody>
</table>
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>FIGURE</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Calibrated and Non-calibrated Subjects</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Achieving Standardization Based on 80-100 Percent Answer Key Agreement</td>
<td>88</td>
</tr>
<tr>
<td>2.</td>
<td>Calibrated and Non-calibrated Subjects</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Achieving Standardization Based on 80-100 Percent Intra-rater Reliability</td>
<td>91</td>
</tr>
<tr>
<td>3.</td>
<td>Calibrated and Non-calibrated Subjects</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Achieving Standardization Based on 80-100 Percent Inter-rater Reliability</td>
<td>99</td>
</tr>
</tbody>
</table>
CHAPTER 1
Introduction

Reliable methods of clinical evaluation must be substantiated if dental hygiene students are to be graded objectively. If dental hygiene educators take measures to internalize superior evaluation techniques, the instructional process can only improve. The purpose of the following investigation was to determine the impact of calibration on dental hygiene evaluators' standardization in instrumentation skill assessment. According to Mackenzie, author of numerous articles concerning objective assessment of dental student skills,

"Just as any fine measuring instrument must be calibrated, so the examiner should be calibrated for clinical evaluation" (1974, 217).

Debated in the literature is whether rater calibration (training) improves the process of grading students in an objective, unbiased, non-judgmental way. Researchers support Mackenzie's theory that calibrating evaluators prior to assessment of a task improves the standardization of evaluation (Biller and Kerber 210; Reed, Feil, and Greer 557; Morgan and Irby 284). Others, however, did not find improvement of evaluator standardization following training (Fuller 22; Mesher and Kerber 210; Weaver and Saeger 256).
All authors agree that a problem of non-standard assessment of student skills by faculty exists in dental education programs.

Important sources of error in clinical measurement stem from the use of human evaluators (Morgan and Irby 39). Human judgement may lead to subjectivity in evaluation due to lack of documented evaluation guidelines and absence of standardized objectives (Abou-Rass 271). Dental hygiene educators may be inclined to include mandatory calibration exercises in faculty training programs if a preponderance of evidence supports the calibration of examiners as a reliable method of achieving standardization.

The topic of improving standardization via calibrating evaluators experienced a surge of research and discussion from the late 1960's through the early 1980's. The following investigation might create renewed interest and commitment to improving clinical teaching and evaluation through calibration of evaluators. Faculty bring to the evaluation process personal biases and varying degrees of critical discrimination, creating a perception of hard and easy graders (Chambers 723). Standardizing evaluators can provide a method of assuring that all students' skills are assessed accurately, efficiently, and fairly. The following investigation examined the effects of evaluator calibration on standardized assessment of instrumentation skills.
Statement of the Problem

The research questions investigated were:

1. Is there a difference in faculty standardization between calibrated and non-calibrated evaluator groups when assessing student instrumentation skills as measured by skill performance criteria answer key agreement during a mock clinical evaluation exercise?

2. Is there a difference in faculty standardization between calibrated and non-calibrated evaluator groups as measured by intra-rater reliability on skill performance criteria during two identical mock clinical evaluation exercises?

3. Is there a difference in faculty standardization between calibrated and non-calibrated evaluator groups as measured by inter-rater reliability on skill performance criteria during a mock clinical evaluation exercise?

Significance of the Problem

Valid and reliable clinical evaluation methods are essential in dental and dental hygiene educational programs to assure that students' grades are accurate reflections of their clinical abilities (Hinkleman and Long 15; O’Connor and Lorey 174). Without accurate assessment of competence, students could not be targeted correctly for remediation or advancement. Although valid and reliable clinical evaluation methods are difficult and time consuming to develop, their benefits of verifying student performance and
assuring the safety of patients while providing accountability to the educational program far outweigh any disadvantages (Ford 149; Madus, Scriven, and Stufflebeam 118).

Unreliable, inconsistent evaluators pose a significant threat to dental hygiene education. Errors made during student skill assessment can result in rewarding students for unsatisfactory performance. Because human error is unavoidable when environment, behavior, and emotion impact upon the evaluation setting, discrepancies will exist within and among evaluators (Abou-Rass 71; Fuller 19; Hinkleman and Long 13; Mackenzie 1974, 214; Mackenzie et al. 284; Mesher and Kerber 83; Natkin and Guild 152; Olson and Comet-Epstein 688; Weaver and Saeger 251).

The importance of consistency in measuring student clinical performance is discussed by Mackenzie. He states that these measurements will certify competence in learners, add to the instructional process, and assure that necessary feedback will be communicated to the learner. Further, he concludes that evaluating clinical skills should be a mainstay in quality health care delivery (Mackenzie 1974, 215). Mackenzie contends that students must be evaluated to form judgments pertaining to their competency and progress (1973, 37). Evaluator standardization must exist, especially in pre-clinical instruction where principles of instrumentation are initially established, to ensure
clinician competence. Consistent grading among evaluators can only exist after standardization has been documented through established intra- and inter-rater reliability (Biller and Kerber 206; Reed, Feil, and Greer 554; Weaver and Saeger 251). The urgency of correcting flaws in student evaluation must be perceived by dental hygiene faculty to assure adequate achievement of goals.

Policy standards monitoring the education of student dental hygienists are described by various organizational review boards. The American Dental Association’s Commission on Dental Accreditation stipulates examiner standardization for student evaluation. For example, "Evaluation criteria (for dental hygiene educational programs) must be predetermined, standardized, and communicated to students (11)."

The problem of unstandard evaluators assessing student ability might be resolved if dental hygiene faculty could find an effective method of ensuring that all evaluators assess student skills in a like manner. Because personal judgement impacts on student learning outcomes, a need exists for professional dental and dental hygiene educational institutions to calibrate evaluators for standardized assessment of skills (Hinkleman and Long 13; Morgan and Irby 274; Olson and Comet-Epstein 688; Walsh and Phelps-Sandall 1). Student learning will be more efficient if appropriate feedback is provided and methods of teaching
are organized (Natkin and Guild 153). Calibration of faculty prior to instrumentation skill evaluation may sensitize dental hygiene educators to the complexity of their task and serve as a vehicle for improving rater reliability. When dental hygiene programs include calibration in faculty development endeavors, students may benefit from more accurate assessment of their skills. Furthermore, standardization will contribute to optimal clinical teaching and may result in the presence of more qualified dental health professionals in the workforce.

Few research studies are available which specifically address standardized educational evaluation techniques for the profession of dental hygiene (Biller and Kerber 209; Daniel et al. 410; McCann, Maddock and Schneiderman 332; McNabb and Zarkowski 19). In addition, not one study evaluated a process, as opposed to product evaluation, and none could be found which assessed standardization of evaluation in the arena of dental hygiene instrumentation skills. The following investigation contributed to the body of knowledge determining whether calibration is a viable addition to faculty training programs for standardization of instrumentation skill evaluation. A significant impact on the practices of clinical evaluation may occur if calibration is shown to have an effect on standardization.

**Definition of Terms**

The following terms were defined for the purpose of
this investigation:

1. **Skill Performance Criteria:** individual performance standards describing acceptable behavior for performing instrumentation skills (Nield and Houseman v). Students are evaluated against skill performance criteria as a means of achieving objective evaluation (Ford 70; Morgan and Irby 90). Skill performance criteria also have been called subtasks (Reid, Feil and Greer 554), checkpoints (Mackenzie et al. 284), and steps (Abou-Rass 273).

2. **Instrumentation Skill Evaluation Form:** a task analysis of instrumentation skills necessary for the use of a universal curet outlined in the text, *Fundamentals of Dental Hygiene Instrumentation* (Nield and Houseman 311). The instrumentation skill evaluation form is composed of 14 skill performance criteria; only 12 were used for the purposes of the investigation. Instrumentation skill evaluation forms also have been called checklists (Mackenzie et al. 284; Morgan and Irby 90).

3. **Clinical Skill Evaluation:** a method of assessing clinical skills by observation for conformance to skill performance criteria. The clinical skill evaluated in this investigation was instrumentation using a Columbia 13/14 universal curet in the mandibular right posterior facial quadrant.

4. **Evaluators:** clinical dental hygiene educators experienced in the techniques, theories, and philosophies of
dental hygiene practice who hold a minimum of a 
baccalaureate degree and can assume the role of assessing 
student instrumentation skills against skill performance 
criteria. Evaluators also have been called raters (Reid, 
Feil, and Greer 554). Evaluators were the subjects in this 
investigation.

5. Pair Matching: a method for controlling for a small 
sample size and lack of a pre-test. Subjects were matched 
based on years of clinical teaching experience and randomly 
assigned to the control group or experimental group.

6. Faculty Standardization: achieved when evaluators' 
assess instrumentation skills consistently and similarly 
when using skill performance criteria. Standardization 
exists when subject assessment of student skills agree with 
an answer key, subjects demonstrate intra- and inter-rater 
reliability, and subjects agree on the assessment within 
their group 80-100 percent of the time (Mackenzie et al. 
287). Faculty standardization was the dependant variable in 
this investigation.

7. Calibration: a method of assuring that evaluators 
participating in clinical skill evaluation are standardized. 
For the purpose of this investigation, calibration consisted 
of one hour of group discussion, demonstration, and criteria 
analysis based on research by Mackenzie (1974, 127), Fleiss 
and Chilton (601), and Mescher and Kerber (86). Calibration 
has been referred to as a training session (Bazan and Seale
726; Fleiss et al. 608; Hinkleman and Long 15; Mackenzie 1973, 43; Mackenzie et al. 287; Natkin and Guild 158) and training program (Reid, Feil and Greer 555). Calibration prior to evaluation was the independent variable in this investigation.

8. **Mock Clinical Evaluation Exercise**: a series of four videotape vignettes depicting a dental hygiene student performing instrumentation skills using a Columbia 13/14 universal curet in the mandibular right posterior facial quadrant. Subjects were expected to assess student skills as if participating in an actual clinical skill evaluation in a pre-clinic setting. Each vignette was approximately one minute in length.

9. **Answer Key Agreement**: agreement of subjects' evaluation of the skills depicted on the vignettes with the predetermined assessment. Answer key agreement was a dependant variable measure in this investigation.

10. **Intra-rater Reliability**: the degree to which subjects used skill performance criteria consistently within themselves as measured by their assessment of one vignette, shown twice during the mock clinical evaluation exercise. Intra-rater reliability was a dependant variable measure in this investigation.

11. **Inter-rater Reliability**: the degree to which subjects used skill performance criteria consistently among each other while viewing one vignette from the mock clinical
exercise. Inter-rater reliability was a dependant variable measure in this investigation.

Assumptions

The following assumptions regarding this investigation were made:

1. The instrumentation skill evaluation form was a valid and reliable data collection instrument.
2. Subjects were experienced clinical instrumentation skill evaluators.
3. Subjects had the ability to distinguish between skill deficiencies and abilities according to the skill performance criteria.
4. Subjects were familiar with the use of the Columbia 13/14 universal curet.
5. Subjects read and were familiar with the material included in the information packets.
6. Evaluator bias toward the clinician was decreased because subjects viewed the same clinician throughout the mock clinical evaluation exercise (Walsh and Phelps-Sandall 13).
7. The mock clinical exercise was a reasonably accurate portrayal of actual clinical skill performance abilities and deficiencies.

Limitations

The internal and external validity of this investigation may have been limited by these factors:
1. Convenience sampling resulted in limiting inferences to the general population.

2. Subjects may have analyzed skill performance more critically than in a normal clinical evaluation situation due to their intrinsic motivation to evaluate correctly (Mackenzie 1974, 217).

3. The mock clinical evaluation exercises’ environment created for the purposes of this study may have made subjects feel uncomfortable.

4. The intact group serving as subjects may have had certain characteristics in common that affected the results of this investigation (Darby and Bowen 75).

5. Novelty effect, or newness of the calibration exercise, may have affected the subjects more than the actual calibration exercise (Darby and Bowen 58).

6. Only one method of faculty calibration was investigated.

7. No measurement of learning, to ensure that the experimental group was standardized, took place following the calibration exercise.

**Hypotheses**

The following three, null hypotheses were tested:

1. There is no statistically significant difference at the 0.05 level of significance in faculty standardization between calibrated and non-calibrated evaluator groups as measured by skill performance criteria answer key agreement
during a mock clinical evaluation exercise.

2. There is no statistically significant difference at the 0.05 level of significance in faculty standardization between calibrated and non-calibrated evaluator groups as measured by skill performance criteria intra-rater reliability during two identical mock clinical evaluation exercises.

3. There is no statistically significant difference at the 0.05 level of significance in faculty standardization between calibrated and non-calibrated evaluator groups as measured by skill performance criteria inter-rater reliability during a mock clinical evaluation exercise.

Methodology

Twelve subjects with instrumentation skill evaluation experience were pair-matched based upon years of clinical teaching experience and randomly assigned to the calibrated or non-calibrated group. An information packet was distributed to allow subjects to begin the investigation with similar baseline knowledge of instrumentation skills. Informed consent materials also were included. Subjects were expected to review the information and sign the required forms prior to the mock clinical evaluation exercise.

The experimental group (calibrated) arrived at the mock clinical evaluation exercise one hour prior to the control group (non-calibrated) so that they could participate in the
calibration exercise consisting of analysis of skill performance criteria, demonstration of acceptable and unacceptable skills, and discussion of ambiguous terms.

Subsequently, calibrated and non-calibrated groups simultaneously viewed a videotape of four vignettes depicting acceptable and unacceptable instrumentation skills using the Columbia 13/14 curet. All subjects completed an instrumentation skill evaluation form for each of the four vignettes.

Data were analyzed for answer key agreement, intra-rater reliability, and inter-rater reliability. Criteria responses obtained from one vignette were analyzed for answer key agreement. Intra-rater reliability was determined by comparing the percentage of calibrated and non-calibrated subjects who assessed skill performance criteria consistently within themselves on two viewings of an identical vignette. Inter-rater reliability was determined by comparing the percentage of calibrated and non-calibrated subjects who assessed instrumentation skill performance similarly among each other. Intra- and inter-rater reliability measured at 80-100 percent agreement established standardization (Mackenzie et al. 287). Calibrated and non-calibrated group results were analyzed for statistical similarities and differences using the z-test for answer key agreement and intra-rater reliability, and chi-square for inter-rater reliability.
CHAPTER 2

Review of the Literature

Literature relevant to the study was reviewed and is presented in the following sections: faculty standardization and calibration, methods of calibration, rater reliability, and competency based evaluation.

Faculty Standardization and Calibration

Student grades are affected by faculty evaluators' interpretation and utilization of evaluation criteria. Dissimilar interpretation of skill performance criteria among evaluators results in inaccurate assessment of students' abilities. When evaluators are calibrated prior to skill assessment, standard definitions of criteria may result; thus, student progress can be monitored in a more unbiased, nonjudgemental way. For example, the literature states that following examiner standardization, all students will be evaluated as similarly as the evaluation criteria will allow (Mackenzie 1974, 217; Mackenzie et al. 287; Reed, Feil, and Greer 554).

The following relates to faculty standardization and calibration investigations conducted in various dental education programs. Discussion will focus on successes and failures in standardization following calibration and
supporting theory.

Reed, Feil, and Greer, questioned the agreement of instructors' scores when evaluating student performance (554). The goal of the study was to assess the effectiveness of a calibration exercise designed to improve inter-rater reliability and grade agreement on individual skill performance criteria, called subtasks. To establish a baseline measurement, nine subjects were asked to analyze 10 amalgam preparations utilizing 17 subtasks. One week later, subjects participated in an extensive training program where case studies were offered describing amalgam preparations. Each case was independently reviewed by each evaluator then discussed among the group. If an evaluator's score did not match the consensus, rationale for grading was required from the outlying evaluator. Subjects then participated in a discussion to eliminate bias and subjectivity while most importantly, striving to change their unstandardized grading behaviors.

The subjects then participated in a second training exercise where an amalgam preparation was evaluated and scores compared and discussed among the group. One letter grade disagreement was allowed for inter-grader agreement, but a failed product required an unanimous vote. Subjects stated that the most common reason for grade inconsistencies lies in difficulty in recalling criteria specification. Upon conclusion of the two training sessions, the subjects
were given new projects to evaluate so that they could use the same subtasks to compare their initial baseline measurement. This procedure provided a mechanism for evaluating reliability and agreement of examiners. Statistical analysis included analysis of variance at the 0.05 level to determine reliability for different numbers of graders. Inter-rater agreement, the frequency with which instructors agreed within one letter grade of each other, was shown to improve after training. Results showed a statistically significant improvement in subjects' grading behaviors at the time of the post-test (p <0.05).

The authors noted that more product failures were assigned at the post-test possibly due to increased ability to distinguish between skill criteria following calibration. Score reliability also improved following the standardization exercise. Baseline rater reliability scores showed a correlation coefficient range of $r = 0.63$ to $r = 0.78$, whereas following training, a closer range of $r = 0.81$ to $r = 0.89$ occurred indicating improved correlation. A correlation coefficient is placed on a continuum from -1.00 to 1.00. The closer the correlation coefficient is to 1.00, the better the rater reliability (Popham and Sirotnik 66). The authors concluded that following training sessions, evaluation was more reliable and agreement among raters increased.

Mackenzie, et al., explored consistency in evaluators'
when observing the same student project (285). The authors equated the subjectiveness of rating scales to the errors resulting from the halo effect, leniency in which evaluators consistently grade high, severity in which instructors consistently grade below the mean, and central tendency when evaluators assign average grades (284). Four faculty members, with three or more years of clinical teaching experience, participated as subjects. All were asked to identify common errors in tasks that would reduce the quality of the product. Also required from the subjects was a rationale for including the error as a critical factor (285). A checklist for each skill stressing observation by sight and touch was used to determine grades. Ten representations of procedures containing errors were developed coinciding with the checklists. Subjects met weekly for one and one half hours over a period of four months for calibration. At the meetings, procedure representations were discussed, consensus grades were determined, and checklists were closely evaluated for conclusiveness. Each subject was required to record at least one evaluation on an audio-cassette, including thoughts and rationale. The audio-tape procedure was repeated one week later and compared for intra-rater reliability (285). Sixteen suggestions for improving rater reliability during calibration sessions based on these authors' findings are offered in Table 1. Data was not
Table 1
Factors to be Included in Calibration Exercises to Increase Rater Reliability

1. Avoid overlap of criteria definition.
2. Commit critical dimensions of criteria to memory.
3. Cover critical dimensions of all items.
4. Indicate all exceptions to definitions.
5. Train evaluators in size estimation if necessary.
6. Use matching evaluation aids (i.e. the same brand and types of probes and explorers).
7. Specify method of evaluation to be utilized (i.e. sight, touch).
8. Operationalize definitions.
9. Evaluate each dimension systematically.
10. Consider evaluators' visual acuity.
11. Discuss when student should be given benefit of doubt.
12. Correlate verbal definitions with visual examples.
13. Be sure evaluators understand verbal/visual examples.
14. Clarify misinterpreted criteria and revise.
15. Encourage evaluators to standardize regardless of educational background.
16. Discuss differences in evaluators' mental processing.

statistically analyzed in the manuscript.

The authors stated that the element of chance must not be overlooked when evaluating rater agreement. Further, they recommended the use of two statistical tests to determine rater reliability (288). The authors stated that the element of chance must not be overlooked when evaluating rater agreement. Finally, evaluation should include a method of assessing the quality of the service (288). For example, did the product meet minimal competency or did it exemplify an outstanding demonstration of ability. The authors concluded that evaluators should be involved only in determining mastery of each checkpoint; a computer then should be used to determine the final grade (288).

An earlier article by Mackenzie addressed factors such as validity and measurement of ability which are essential to the evaluation of clinical performance (1974, 214). Validity of evaluation ensures that the assessment is an accurate reflection of the student's performance (1974, 215). Validity of evaluation can be improved by including both process and product evaluation of all major services performed by a practitioner, incorporating criteria which will effect the health and well-being of patients, selecting criteria which can overlap to other procedures, and providing a mechanism for self-evaluation (1974, 215).

Objectivity is essential in perfecting evaluation of clinical performance; when objectivity is attempted, there
is less chance of committing an error in decision making (1974, 217). Mackenzie notes that judgement will be affected by personal expectations and motivations, and thus, evaluators must not let personal reactions to students unconsciously alter objectiveness (1974, 217). Rater reliability will decrease the number of measurement errors. For example, if a crown is measured at seven millimeters with a Boley gauge, all Boley gauges should measure the crown as seven millimeters. Mackenzie stressed the complexity of training examiners to maintain constant decision making criteria.

To assist evaluators with a standardized method of evaluation, Mackenzie proposed the following: priority should be placed on goals and purposes of the procedure, criteria should be consistently defined, room for decision making about marginally acceptable skills should be available, as few criteria as possible should be used to test competence, and examiners should be allowed to view correct and incorrect procedures in addition to verbal discussion (1974, 217).

Methods of addressing potential errors in evaluation to be included in a calibration exercise are pair matching, divergent matched pairs, and varying the degree of difficulty in detecting crucial difference (Mackenzie 1974, 218). Pair matching involves comparing a good example with a bad example so that evaluators can see ranges on the
quality continuum; this method is used for product evaluation. Divergent matched pairs requires selection of a pair different from the matched pair in irrelevant factors such as size, shape, and color so that evaluators can discriminate differences. Varying the difficulty of detecting the crucial difference allows evaluators to develop consistent judgement of acceptability by challenging their abilities. To assist in detection of crucial differences, evaluators should be allowed to become familiar with any diagnostic aid that will be used throughout training.

Motivating faculty to evaluate accurately is another important facet impacting upon standardization and calibration procedures. In order for students’ to receive grades that are true reflections of work ability, motivation to adhere to criteria guidelines must exist in clinical faculty. Mackenzie suggests the use of a faculty reward system so that evaluators will be intrinsically motivated to provide objective, consistent evaluation. Examiners should be certified to evaluate prior to assessment of student skills to ensure that the assessment will be exercised in a non-biased, standardized manner. When an examination requires significant time to implement, faculty may become less motivated and score students improperly. Attempts should be made to make assessment of student skills quickly and concisely so that significant time is not spent on the

Weaver and Saeger studied the effects of training on rater accuracy and reliability (252). The study was divided into three phases: pre-training, rater training, and post-training. During the pre-training phase, four faculty evaluators serving as subjects rated nine complete dentures which best demonstrated a variety of students' abilities and errors in pre-clinical prosthodontics. Prior to subject evaluation, the denture sets were assessed and graded by board certified prosthodontists to establish an expert basis for comparison. Subjects were asked to rate the dentures on a three point scale: R (cannot be improved), S (clinically acceptable), or T (clinically unacceptable) (252). If a grade of T was given for any one criteria, the subject was instructed to explain the grade as well as suggest ways of improving the skill in writing.

Six months later, the same subjects participated in the rater training phase. For four hours, subjects were lectured on definitions and descriptions of rater terminology, rater training efforts used previously, the rationale for measuring clinical performance, specific rater terminology, and examples of rater errors relating to the process being evaluated. The subjects also participated in a group discussion of the observer's role, methods of
improving rater accuracy, and areas of the evaluation not being graded, such as patient management. The pre-clinical grading criteria for dentures was then reviewed and discussed. The post-training phase began upon completion of the written evaluation. Subjects evaluated the same group of nine dentures as was used in the pre-training phase using two grading scales: R, S, and T, as well as the traditional scale of A, B, C, D, and F. Accuracy of raters' assessments of projects within and among each other was determined by the differential accuracy correlation for the pre- and post-training phases.

Intraclass correlation using analysis of variance measured inter-rater reliability. Mean differential accuracy correlations were determined to be $r=0.76$ for the pre-training phase and $r=0.78$ for post-training. These correlations tested statistically significant at the 0.001 alpha level. Authors stated that each accuracy correlation was significant, however, there was no improvement following training. Higher reliability was noted when using a five-point rating system; an 84 percent rater reliability estimate was determined in the post-training phase when using the three point scale while the five point scale yielded 93 percent rater reliability (256).

Possible explanations given by the authors for the unchanged rater reliability after training included subjects' pre-existing ability to evaluate leaving no room
for improvement. Comments made about the five point scale assuring better reliability stemmed from raters being more comfortable with this scale. The three point scale was a temporary experiment, lowering the degree of confidence which the evaluators had when using it. To support their findings, the authors quoted a study which concluded that as the number of points on a scale increases, so does the average inter-class correlation of ratings (Ebel 423).

Fuller studied necessary alterations in a biased, "glance and grade" evaluation system (19). The purpose of the investigation was to determine if another evaluation method should be implemented to replace the method in use at the time. Specific purposes of this study were to determine if: 1) agreement among raters in traditional grading methods was significant, 2) consistency among raters exists in their own judgments, 3) improvement in rating could occur by training and/or other aids, and 4) sequencing of training methods affect rater reliability (20). A sample of operative dentistry products were used as the projects to be evaluated by eight, independent faculty raters who were paired for statistical purposes. The examination procedure occurred three times and calibration methods were introduced throughout the experiment within each evaluator pair except for the first pair which consistently used the glance and grade method. Calibration methods introduced included models with checklists and training. Training sessions were
two hours in length and included topics such as teaching/learning concepts specific to these students; human factors which affect grading, such as the halo effect and the generosity factor; an overview of the specific technique raters were to use in the experiment; and lastly, the opportunity to practice evaluating pre-rated restorations (21).

Spearman's rank correlation test was used to measure intra-rater reliability. Results revealed no statistically significant correlations between rater scores. Reliability coefficients for each examiner pair for all three examinations were procured through use of analysis of variance. A reliability pattern was not noticed among scores. Fuller suggested the following conclusions based on the results of the study: no significant agreement between raters existed when using traditional evaluation methods, variable reliability for raters existed in repeated ratings of the same product, rating cannot be improved by training or other aids, and the sequence of training and use of aids is unimportant in improving reliability (22).

Models of Instructor Calibration

Suggestions and recommendations made by authors using various calibration techniques were reviewed to enhance understanding of calibration theories. Review of other calibration methods used throughout clinical education assisted in designing a method of standardizing the subjects
for this study.

O'Connor and Lorey tested the use of photographic slides, called comparison stimuli, to improve inter-rater agreement during product evaluation (175). Inter-rater agreement was assessed in this investigation through analysis of accord in "level of standards", consistency in evaluating the relative quality of products, and agreement in rating assignment to a given product on a specific criteria (175). During phase one of the two phase study, ten subjects participated in two rating sessions where they used six criteria and a five point rating scale to assess 20 restorations. At the first rating session, subjects were instructed to score 3, 4, or 5 as evidence that the student met the criteria at increasing degrees of mastery while scores of 1 or 2 meant that the criteria was unacceptable (175). During the second rating session, subjects viewed comparison stimuli representing what an average representation of each criteria should look like to be assigned a score of 3. Comparison stimuli were used as references for grading the products. Subjects were to assign 1 or 2 if the criteria was performed below average and a 4 or 5 if the criteria was executed better than shown in the slide (175).

Again, subjects evaluated restorations using six criteria and a five point scale for the second phase of the study; however, a different restoration was evaluated and 18
subjects participated. Subjects were then divided into experimental and control groups for the second rating session by statistically equalizing group ratings from the first session. This procedure created groups with inter-rater agreement. The control group did not use comparison stimuli. The two groups were created to determine whether observed improvements in inter-rater agreement resulted from the use of the comparison stimuli or were a result of participating in the first rating session (176).

Analysis for the first phase and the experimental group of the second phase compared inter-rater agreement in the rating sessions with and without comparison stimuli. The control group was analyzed by comparing measures of agreement in the first and second rating sessions of phase two. Statistical testing included a one-tailed F test to assess "levels of standards", analysis of variance to assess reliability of measurements, and kappa to assess agreement among judges in assigning ratings. For each criterion, raters were categorized per phase into one group if their kappa value was higher in the first session and another if kappa was higher in the second session. A sign test was then applied to assess whether the number of raters showing improvement in the second session was statistically significant.

Results of phase one indicated that "levels of standards" improved; variability among raters statistically
decreased when using comparison stimuli. When assessing relative quality of products, agreement between groups of raters was high and improved when comparison stimuli were utilized. Agreement was established at $r=0.30$ at the first rating session and $r=0.66$ at the second rating session. Comparison stimuli also increased agreement in ratings of individual criteria.

Unanticipated by the investigators at the end of phase two, none of the significant outcomes of phase one reoccurred. The authors cited an existing high level of inter-rater reliability leaving little range for statistical improvement. The authors also suggested that a single comparison stimuli may not have been enough for one criteria; subjects may need to see different interpretations of acceptability. They stated that variable photographic quality may have affected the results. Overall, the authors confirmed the use of comparison stimuli to improve inter-rater reliability when assessing a product.

Walsh and Phelps-Sandall, in their discussion of minimizing bias during research experiments, recommended key elements which increase the internal validity of an investigation by decreasing rater bias during skill evaluations (13). Recommendations included decreasing the amount of researcher involvement during data collection and increasing the degree of encounter. When the researcher has minimal involvement in the data collection process, the
authors hypothesized that the research design will be more objective. For example, the authors suggest that the principal investigator not be heavily involved in data collection. This suggestion lends credence to the practice of not having the instructor who taught a group of students evaluate that group due to personal bias regarding skill improvement over time. The use of the "double-blind" method in which neither the person administering the experimental variable nor the subject receiving it is aware of the investigation's purpose, is the best method of data collection to ensure objectivity (14).

The authors also stated that increasing the degree of encounter, defined as the number of times the variables are measured, insures that objectivity will be increased when a large number of subjects or variables are measured all at once or a small sample is measured over a long period of time (14). The selection of the sample size should be representative of the population while meeting the criteria established for subjects. Pair-matching subjects (pairing examiners whose average scores represent a high and a low score so that combined scores will reflect an acceptable average) will also increase objectivity. Time series designs with multiple, repeated measures are excellent for studies with small groups. Lastly, the degree of operation, i.e., the preparation and adhesion to defined terms within the study and how strictly they are followed, has been shown
to decrease evaluator bias. Degree of operation can be compared with evaluation criteria being clearly outlined so that the evaluators' own interpretation of the task does not bias the outcome of results. Each one of these suggestions can be applied to skill assessment; thus, the amount of evaluator bias may decrease, thereby, increasing objectivity. No statistical interpretation of the suggestions was offered by the authors.

Olson and Comet-Epstein described a method of increasing instructor objectivity of student performance and feedback during a pre-clinical techniques course (688). They stressed that effective teaching and student learning requires total objectivity of the instructor when evaluating student performance. One hundred and three students were divided into three groups which rotated between faculty evaluators five times in a semester. Projects were graded on a daily basis with a specially designed form using feedback scales of "excellent," "satisfactory," "satisfactory if modified," and "unsatisfactory." Students were given random identification numbers every eight weeks to place on their projects. Each instructor graded every project anonymously using previously determined criteria which were assigned only to them; for example, instructor A graded criteria 1, 2, and 3 on each project, instructor B graded 4, 5, and 6 on each project and so on. Scores were then tabulated by a computer to see how the students in each
group rated in comparison with the rest of the class. Students needing further skill improvement were identified and required to meet with their course instructor to analyze the project. The researcher cited the following three advantages of a conference for student skill improvement: 1) self-evaluation skills are sharpened, 2) course instructor can have a quality check of project evaluators, and 3) all students, regardless of grades, may discuss their concerns about the project and final grades. This paper was designed to share knowledge of a teaching technique; no statistical analysis is offered.

Bazan and Seale conducted a study to determine if negatively worded performance criteria would effect feedback and student performance reliability when using 140 students and a five point grading scale. They define the purpose of pre-clinical evaluation as "a method of improving learning through feedback, and determine progress and competence by quality assessment" (726). By using immediate feedback, students develop perceptual skills for self education. Training sessions used to standardize evaluators were implemented in the study. Evaluation forms were completed each time the student attempted the skill and students were given feedback. When the student felt confident with the skill, an identical form was completed for a grade. Rotation of evaluators compensated for any subjectivity. Immediate written and verbal feedback was available; the
form was completed in an average of 60 seconds. A inter-rater co-efficient of $r=0.70$ was determined. The authors concluded that the use of a three point grading scale may further improve inter-rater agreement.

Edwards, Morse, and Mitchell incorporated skill evaluations analyzed by computer in a study utilizing extensive instructor calibration for evaluation of restorative dentistry projects (294). All projects within a class were screened by the course director to collect a random sample of six representative dentoforms. As the calibration exercise, six raters were required to evaluate all samples using a scale of 1-10 points in five categories not mentioned in the study (694). A discrepancy was determined to exist if a particular evaluator was two points away from the rest of the group. Overall results revealed that all problems were identified by all evaluators; however, degree of error severity was the usual problem in agreement. Many of the instructors became aware of their tendencies toward strict grading or lenient grading as a result of the training sessions. When all areas were discussed in detail and raters felt confident with assignment of grades, six more projects were selected and graded. Inter-rater reliability became stronger on the second evaluation. Upon assessment of individual instructor's scores, subjects were paired depending on variations in scoring: a high scorer was paired with a low
scorer. The subjects then evaluated projects of an entire class with names covered to assist in decreasing personal bias. Incorporated into this method was an interesting addition; students were required to self-evaluate their projects. If a close agreement was determined between their own score and that of the instructor's, the student was awarded five bonus points. A computer analyzed the entire evaluation including determination of criteria marked deficient by raters and a reliability agreement for paired evaluators. Calibration of examiners increased inter-rater reliability, although it was expressed by the authors that an excessive amount of time was involved to perform faculty calibration.

Natkin and Guild, in their pioneering dental calibration study, were concerned with inconsistent evaluating and grading practices in the University of Washington Dental School (153). Inconsistent evaluation techniques had built a reputation for negatively effecting student grades. A major assumption of the clinical grading policy was that students would receive an "A" unless evaluators determined an error in the project. Previously outlined criteria, such as a skill evaluation, was not available to guide clinical evaluation; instructors would have to subjectively evaluate and record errors which they determined to be worthy of point reduction. High inconsistencies were present during student evaluation
making student self-evaluation impossible and compromising grade continuity.

Severity of errors were determined by grades of slight, moderate, serious or critical and based on the effect of the error on the prognosis of case success (Natkin and Guild 156). Five evaluators participated in the three phases of the study. Evaluators consisted of two, first-year graduate students with no evaluation experience; two second-year, graduate students with one year of evaluating experience; and one faculty member with four years of evaluating experience (158). During the first phase, subjects were required to give letter grades to 50 projects and to justify the grade given. In the subsequent two phases, 25 projects were evaluated by the subjects. In addition to a letter grade and comments, the degree of error seriousness was determined by the subjects. An example of error seriousness occurred if the patient's health was compromised as a result of the treatment. At the conclusion of phase two, eight evaluator training sessions were scheduled in which ten lab projects were evaluated by each subject. Each session represented a different type of project students would be required to fabricate that semester. An informal discussion of evaluation methods, criteria to use when determining errors and delegation of error severity commenced after the ten examples were individually evaluated.

Statistical inferences were made in relation to range
of scores, the average deviation of grades assigned, and the mean average deviation; a small number signified a closer agreement between evaluators. The percent of cases with grade ranges within two points of each other also was analyzed by the researchers. The highest degree of overall rater-reliability occurred in the phase in which subjects only read information pertaining to the evaluation method. Training of subjects did not make a statistically significant difference in reliability of grading although there was an increase in rater reliability. The increase in rater reliability resulted from the definition of a baseline grade, development of an exhaustive list of possible errors, four categories of seriousness, and a limited number of seriousness categories (Natkin and Guild 160).

Problems noted by Natkin and Guild in using this system were lack of significant decrease in error serious agreements, training not addressing evaluator ability of error detection, and lack of definitions of error seriousness and definition of errors which should be placed in specific error seriousness categories (160). The authors proposed that reducing the number of error severity categories from four to three would increase rater reliability. Benefits of a formal and reliable evaluation method cited by the authors include: a decrease in the amount of student confusion and an increase in student motivation, a way of communicating to students on what
exactly they will be evaluated so that they may critically analyze their own work, development of a method of introducing the evaluation to new faculty members, and a record of faculty grading so that inconsistencies can be monitored in an attempt to improve judgement in evaluation.

**Rater Reliability**

The literature reveals methods for increasing the validity and reliability in clinical evaluation through various methods of increasing rater reliability. These methods resemble those reviewed in the previous two sections; however, the suggestions presented here are mathematical. Objectivity in evaluation mandates that all evaluators be reliable within themselves (intra-rater reliability), and among each other (inter-rater reliability). Rater reliability can be thought of as insurance that subjectivity did not interfere with skill evaluation from student to student, from evaluator to student, from evaluator to evaluator, or from student to evaluator.

A 1987 study by Daniel, et al. assessed the ability of examiners to evaluate dental sealants similarly in relation to their experience in clinical evaluation (410). Subjects were divided into two groups, "experienced" dental instructors (n=2) and "inexperienced" dental hygiene instructors (n=2). Evaluation criteria included area coverage, porosity, and retention. Extracted molars
prepared with a sealant and mounted in plaster were evaluated twice on separate occasions by each subject. Statistical analysis of intra-rater reliability occurred via Pearson’s Product Moment Correlation as well as two by four split pilot analysis of variance. After analyzing the experience of the evaluator and characteristics of evaluation criteria, both inter- and intra-rater reliability were directly related to the amount of experience in sealant evaluation of the evaluator. The authors concluded that more experience yielded better reliability within and among examiners when evaluating sealants. The authors' interpretation of these findings suggest that increased familiarity with evaluation criteria, previous calibration experience, and the ability to separate critical areas from non-critical areas within evaluation might be related to experience.

Degrees of intra- and inter-rater reliability between two examiners were evaluated in a two stage study by Fleiss and Chilton to assess agreement between scorers in measuring periodontal disease (601). Examiners were calibrated in the assessment of patients and definitions of each type of periodontal disease. Twenty-six adult volunteers were evaluated by each examiner for the following aspects of the disease: gingival recession, tooth mobility, calculus accumulation, pocket depth, gingival inflammation, and bacterial plaque accumulation. Examiners were required to
decide if signs of periodontal disease were present in the patients' mouths. Prior to examining the second set of 24 patients, the examiners were calibrated again.

Kappa values assessing degrees of inter- and intra-rater reliability served to analyze results. Following the second calibration, kappa values increased for all aspects of periodontal disease rating except for the criteria, "presence of plaque"; examiners thought that all patients showed evidence of plaque in every examination. Inferences from this data suggest that inter- and intra-rater reliability increases following subsequent calibration exercises.

Biller and Kerber conducted a study to detect differences in scaling error detection among groupings of clinical dental hygiene faculty. They believed that factors of variance and reliability within evaluators should be assessed prior to student evaluation so that standardization of examiners could be determined (206). Participating faculty members were divided into three groups: 1) "experienced faculty" having five or more years of clinical teaching background, 2) "experienced graduate students" including graduate teaching assistants with nine or more months of clinical teaching experience, and 3) "inexperienced graduate students" who were just entering the program without any experience as clinical evaluators.

Patient classifications as determined by calculus and
degree of disease were defined operationally for the purposes of the study. The Reliability Theory expressed in a mathematical equation upon which the study was based, explains that, "part of any scaling score is due to errors made during the evaluation process and the remainder is due to the quality which the instructor is attempting to measure (203)." Scaling scores were divided into three areas: true score (the number of surfaces on which calculus is actually remaining); error score (a mistake in either detecting too few or too many calculus errors), and recorded score (the algebraic sum of the true score and error score). Error scores in relation to true scores were analyzed to evaluate reliability in the procedure. Groups of instructors were assessed separately and compared with a mean scaling score determined by each individual within the group. Variations in instructors' mean scores were evident only within inexperienced graduate students.

An analysis of variance design accommodating for random error determination, including a reliability procedure which allows for errors evolving from patient-student interaction and instructor differences, was used for statistical analysis. When estimating evaluation reliability, correlation coefficient scores varied from $r=0.06$ to $r=0.74$. Once various levels were determined, the preponderance of consistency in instructor groups was found primarily in the experienced graduate students, where only 26 percent of
differences were due to differences among instructors. Experienced faculty members accounted for 64 percent of the differences. Inexperienced graduate students showed the least consistency at 97 percent. Differences among groups were not statistically significant. The authors cite individual educational background and personal approaches to clinical teaching situations as two possible explanations for differences among faculty groups. Suggestions made for increasing instructor reliability included increasing instructor calibration in calculus detection, reducing the numbers of evaluators, and rotating faculty during student encounters.

A study by Mescher and Kerber investigated three areas: variance among instructors in assigning student scaling performance scores, the effect of any one instructor's evaluation on students' final scaling proficiency scores, and the effect of weighing scaling performance at 50 percent in students' final grades while weighing other performance criteria at 20 and 10 percent (83). Twenty-three dental hygiene students yielded 1,125 evaluations. Five faculty evaluators determined scores for each student on two occasions. Relationships between instructors and students and numbers of evaluations were determined by chi-square tests. Scaling errors were defined as one area of calculus remaining after completion of scaling per tooth surface. Variance among instructors was determined by analyzing mean
scaling error scores of individual instructors to total mean error scaling scores within the subject sample. Performance scores were based on number of scaling errors, number of polishing errors, patient assessment, and tissue integrity. Raw scores were converted into standard scores containing a mean of 50 and a standard deviation of ten. Total standard scores and scaling standard scores were then recalculated. Individual instructor variance on grades was determined by scores assigned by each instructor which were compared to the actual final grade. Analysis of variance determined whether variance of final scores was the result of variance of individual student performance compared to instructor scoring.

Significant intra-instructor and inter-instructor variance was discovered upon data analysis. If certain scores were eliminated, 13 out of 23 of the students' scores would change. Most total performance scores were located less than one standard deviation from the mean; therefore, student ability affected the variance in scores. The authors concluded that discrepancies do exist when multiple evaluators are used in clinical evaluation of students.

Fleiss, et al., were concerned with standardization of examiners in relation to inter- and intra-rater reliability of the assessment of a populations' dental needs (604). Methods of statistical interpretation of results included the degree to which an examiner may depart from usual
scoring criteria when confronted with an unusual patient. Two examiners were trained in evaluation procedures for two days. Training included a verbal explanation of the scoring method, a clinical examination of patients, and a discussion of findings. Training and examination of patients occurred once a year for three years; 27 patients were examined in the first year, 19 in the second year, and 18 in the third year. A number of results included a negative number noting the presence of an appreciable interaction between examiner and patient which could have been remedied by further examiner calibration. The authors believed that two days of required examiner calibration were the key to the outstanding results of the study. According to the authors, when choosing the proper statistical test to assess rater reliability of examiners, the investigator should determine if the examiner results will be fixed or random. The authors stated rater reliability is important for proper assessment of populations so that their true dental needs could be accurately reflected in public health care decisions.

Hunt compared the statistical accuracy of accepted methods of determining rater reliability: percent agreement, Pearson's correlation and kappa values. He argued that measuring reliability among examiners using these methods frequently leads to deceptive results. In addition, Hunt believed that a single method should be
established as the most accurate and valid method of determining inter- and intra-rater reliability so that results of experiments can be accurately calculated to include exact rater reliability data.

The percent agreement test is widely used to assess rater reproducibility, but will not determine values occurring through chance. When a researcher uses only percent agreement without checking it against other tests, results may demonstrate rater reliability as excellent when in reality chance played a large role in the evaluation of the results. Pearson’s correlation coefficient also has limitations in the validity of rater reliability. Hunt’s reasoning is that systematic bias of raters is not controlled when using this test, causing overestimation of actual observations.

Perhaps the most accurate reflection of rater reliability can be determined when using the kappa test. The ratio formulated when designing the kappa statistic contains a numerator which shows observation agreement beyond chance and a denominator reflecting marginal distribution of the variable being measured. In comparing all three tests, kappa has been shown to provide the most accurate results. When using the kappa statistic, a score of zero represents random agreement among raters, while a score of one means perfect agreement; when the scores are placed in ranges, 0.40 and below represents poor agreement,
0.40 to 0.75 notes fair to good agreement, and 0.75 to 1.00 can be determined as excellent agreement. Hunt’s study validates the use of any one of these statistical methods to assess rater reliability.

**Competency-Based Evaluation**

Competency-based evaluation (CBE), involves assessment of student ability through comparison with minimum standards of effective performance (Popham and Husek 2). Both product and process evaluation can be incorporated in this evaluation method. Grading associated with CBE, is always pass/fail, mastery/non-mastery, acceptable/non-acceptable, correct/improvement needed, or right/wrong. The criteria for determining if the student performs the skill correctly or incorrectly is determined by clearly defined criteria and correct behavior. Students are not truly competent in a skill until it is repeatedly performed correctly. Competency-based evaluation allows for student error so that the individual is not expected to master the skill he/she first learns.

Criterion-referenced, or competency-based evaluation methods ascertain an individual’s status in respect to some performance standard (Popham and Husek 2). The score is most accurately interpreted when it is considered alone, without an analysis of other scores. Criterion-referenced measurement is designed to determine if the learner has mastered a pre-requisite skill so that they may attempt a
new skill. CBE also can reflect a set of instructional objectives which need to be introduced and learned in sequence and will determine who has the ability to perform a certain skill (Popham and Husek 2). Variability is irrelevant due to the necessary independence of scores. When constructing this type of test, the goal is to ensure that the criterion item is an accurate reflection of a necessary behavior. In terms of reliability, criterion-referenced tests must be internally consistent; each criteria must be similar in terms of what it is measuring. A perfect score does not mean the recipient was the best in the class, only that the individual met all of the criteria. Validity in criterion-referenced testing is assured as long as the desired criteria is represented in the evaluation. When analyzing a criterion-referenced test, items which do not always discriminate should not necessarily be discarded; the criteria may possibly assist in determining if the skill is mastered by the student (Popham and Husek 6). The item should be analyzed for construction deficiencies before being discarded as inappropriate criteria.

Interpreting a criterion-referenced test requires determining whether the examinee performed or did not perform the behavior correctly. Before evaluation commences, the examiner should define a range of acceptable behavior, for example, the student will be considered competent if 90 percent of the answers are correct. This
method allows errors and still deems the learner competent in the skill (Popham and Husek 7). Interpretation should depend exclusively on the immediate interpretation of the available data. When reporting the results of a criterion-referenced test, the most helpful information is the number of individuals who meet the criteria, the mean and standard deviation of scores, and the percent of criteria which examinees performed correctly. The following studies relate to competency based evaluation in clinical skill evaluation.

McNabb and Zarkowski define competency as the, "ability to fulfill the total performance responsibilities of a given profession; competencies are acceptable levels of proficiency (20)." They argue that unless a competency based system is used in clinical evaluation, student skill assessment will not be accurately measured. Further, they contend that in order for students to be properly evaluated, calibration of examiners is a necessity.

Their study, analyzing competency-based clinical evaluation, took place in a dental hygiene program over the span of one and one half years. Objectives of concern to the principal investigators were: 1) to decide if the system was a practical method of assessing dental hygiene students and 2) to compare a competency-based system to a requirement-based one. The criteria grading scale used in competency-based evaluation was determined to be "acceptable" or "unacceptable" in order to decrease
extraneous variables. The study was broken into three phases; phase three is unavailable, as its results are unpublished. In Phase I, six students volunteered to receive competency based evaluation. Two faculty members were calibrated and versed in philosophies of competency-based grading. Every two weeks, students were required to meet with their faculty advisor where they had an opportunity to ask questions and could receive feedback relating to clinical performance. Oral and written comments were required from students at the conclusion of Phase I.

In Phase II, half of the class was randomly assigned to the competency-based grading method; all other students used the traditional requirement-based method. An additional faculty member was added for student advising. To calibrate examiners, a videotape was viewed, examiners performed practice evaluations, and all examiners discussed their findings stating decision rationale. At the completion of Phase II, all students and faculty, whether assigned to the requirement or competency-based system, were interviewed and administered a questionnaire to measure their opinions on the objectivity of the system and specifics on methods used in the evaluation system, such as meetings and feedback. Upon evaluation of Phase II, faculty expressed a desire to continue the system and determined it to be an accurate measurement of clinical competence and successful in identifying students in need of remediation.
The authors concluded that a competency-based evaluation system was in fact a practical method of assessing student dental hygienists and that this type of evaluation system was superior to requirement-based. From analysis of the responses obtained from the questionnaire, the authors were able to conclude that there was less stress in the evaluation environment due to self-pacing, non-graded procedures, and open communication between students and instructors. They also concluded that immediate feedback was available, self-assessment was encouraged, students were challenged to achieve post-competency, and those students requiring remediation were accurately identified. All students decided that they would prefer to keep the evaluation system in which they participated whether it was competency or requirement-based.

Hinkleman and Long applied the theories of competency-based evaluation using three points of grading: "ideal", "clinically acceptable" and "unacceptable" (14). Values assigned to each point were 3, 2, and zero, respectively. The literature suggests that competency-based evaluation is usually graded as pass or fail. The authors hypothesized that students discontinue learning a skill once the skill definition is met and never again try to improve the quality of the product or process (14). Using three points of grading controls this phenomenon.

Laboratory course instructors at the University of
Pittsburgh were polled regarding criteria to be included in each specific product evaluation. This information was combined with criteria in the textbook from which the students learned the assigned skill. Sixty inlay procedures, with students’ identification covered, were randomly chosen as the evaluation product. Half of the preparations were evaluated prior to training sessions by the four laboratory evaluators participating in the study. One week later, a training session was scheduled to review evaluation criteria; thirty more inlays were evaluated at that time. Inter-rater reliability was mathematically computed using percent agreement rather than a reliability coefficient due to the three point grading scale. Percent agreement also was used to analyze the number of times paired instructors agreed on the assessment of scores.

Authors concluded that agreement among evaluators was somewhat better using the competency-based system in comparison with grading students by ranking clinical ability; however, results were not statistically significant. The training session yielded only a slight improvement in inter-rater reliability; statistical analysis was not included in the study. Benefits of using this system were: immediate feedback was available to facilitate student learning, skills were learned using the exact same criteria and identical sequencing, a component for quality assurance was added, and students accepted this method as an agreeable
method of evaluation (McNabb and Zarkowski 18).

In yet another article, Mackenzie contends that traditional "glance and grade" evaluation methods are restrictive due to their emphasis on the final product and not the process, ill-conceived due to their non-systematic approach, trivial due to their simplistic nature, and biased due to their subjectivity (1973, 37). Competency-based evaluation is the optimal method of clinical evaluation according to the author. Three areas which Mackenzie stresses to include in clinical evaluation methods are quality, quantity, and the need for performance criteria (37).

Quality is defined by Mackenzie as an assessment of variety and degree of excellence (37). Tasks performed by students must be defined and measured procedurally. Students can then give attention to areas being measured which are known to them. For example, evaluators must be sure to not assess the quality of restorations on aesthetics alone; product function must be strongly considered. The final product must be only a small part of the complete evaluation. Mackenzie recommends process evaluation due to its ability to identify student learning difficulties that prevent them from reaching an acceptable final product. Students also should be shown what an acceptable product looks like and the acceptable process for attaining it. A definition of right and wrong during a process will enable
students to develop self-evaluation skills to criticize their own work. A videotaping of procedures may facilitate critical self-evaluation.

Quantity of evaluation addresses the number of times a student must perform a skill in order to be considered competent (41). When learning a skill for the first time, initial experience of acceptable and unacceptable performance eventually will reduce the number of trials the student requires to master the skill. Difficulty of a given task must be analyzed on the basis of the ease of the following: the possibility that an error will be made, the number of times the procedure will be encountered during daily practice, the frequency and extent of deviating from the norm, clarity and complexity of cues, the necessity for timing the task, the value of speedy performance, precision degree, number of steps involved, and how much the procedure overlaps with previously learned skills (41). When all these factors are taken into consideration and tailored to individual students, Mackenzie concluded that an estimation can be made on quantity in relation to task repetition based on student ability.

The third area stressed by Mackenzie is how perceptions of performance criteria are affected by individual values and beliefs (42). Keeping subjectivity away from evaluation is pertinent, though often an exercise in futility. Mackenzie describes subjectivity in evaluation as no
agreement among examiners and the true ability of students cannot be measured. Objectivity allows for fairness and makes room for evaluators to agree on a list of set criteria (42). To decrease the amount of errors present while evaluators use performance criteria, Mackenzie suggested the following: 1) make observation criteria clear and definite, 2) allow provision of feedback to increase student learning, 3) decrease the number of criteria needed to determine competence, 4) allow room for improvement through evaluator assessment of criteria, 5) stress the new obligation of health care professionals to be quality assurance managers requiring evaluation skills to assure acceptable products, and 6) monitor inter- and intra-rater reliability to assure reliability and validity of evaluation instruments.

Mackenzie states that performance criteria present in a skill evaluation form should be carefully selected to include the goals and purposes of the procedure (1973, 43). Mackenzie also contends that raters should be able to decipher the meaning of each criteria. These criteria should be detectable via observational techniques and standardized through training so that subjectivity is not present in the evaluation of students. Standardized models providing examples of each level of the criteria and relevant, as well as irrelevant aspects of the evaluation, assists evaluators in deciding how the skill was executed. Mackenzie believes
that instructors should be certified to evaluate student competence. Overly precise or numerous criteria which wastes faculty time and energy should be avoided in evaluation forms. Observational aids such as calibrated instruments (templates, probes, and stopwatches) will increase reliability and objectivity. Lastly, grading should facilitate instructional decision making to determine the presence and degree of correctness or incorrectness.

McCann, Maddock, and Schneiderman studied the effects of a competency-based grading system on dental hygiene students' motivational levels during clinical procedures (332). Nineteen first year dental hygiene students were randomly assigned to one of two groups, one evaluated on a letter-grade basis and the other evaluated on a pass-fail basis. The scaling and polishing procedure of each student's third, fourth, and fifth "light" calculus patient of the semester was graded by the same clinical evaluator whose performance was proven to be consistent through documented intra-rater reliability. All other areas in the student's treatment of patients were graded by their section instructor except for scaling and polishing which were assessed together. More than four total errors in the scaling/polishing category would constitute a failing grade. Only students knew their group status. When the standardized evaluator assessed scaling and polishing, students recorded the grade assigned by the evaluator based
on their group assignment and number of errors based on a conversion chart. After six months, a meeting was held where all students were required to complete a questionnaire to measure motivational effects of grades on clinical performance (332). For ethical reasons, students were given a choice of what type of grade, letter or pass-fail, they would receive as a final grade. Student's t-tests compared pre-clinic scores, college grade-point average, and dental hygiene grade-point average between the two groups to determine if any would effect clinical ability; no statistically significant results were obtained in the study. Further, a student's t-test determined no difference in over-all performance as determined by final grade in the two groups.

Based on the questionnaire, the two groups perceived grading methods and motivational level differently. The letter grade group perceived grades as "fairly important," "while the pass-fail group determined that grades "were hardly important at all," (335). The group graded by letters decided that on the average they would have been better motivated to perform well in clinic if a pass-fail method was used in their evaluation. The pass-fail students would have been motivated only to a small extent if they were given letter grades. Overall, the letter group decided grades were more important and motivational than passing or failing.
The authors concluded that clinical performance is not influenced by grading method in comparison with classroom grading. The factor which most strongly motivated students to perform well in clinic was not determined in the study. Although both methods can be used in clinic, the authors recommended that pass-fail, or competency-based evaluation, replace letter grading in dental hygiene clinical education (337).

Competency-based evaluation does not allow for grade inflation. Pass-fail systems also develop intrinsic motivation moving the student toward critical self-evaluation. Disadvantages to this grading method are that faculty must encourage students to excel in clinic and that pass/fail clinic grades have no universal interpretation. The authors cited threats to internal validity of the study as faculty intervention, novelty effect, and the presence of similar characteristics when dealing with intact groups (337).

Summary

The literature is divided on whether calibrating evaluators plays any role in achieving standardization. Studies are available which have added to the premise that calibration is equated with an increase in intra- and inter-rater reliability or rater standardization (Edwards, Morse, and Mitchell; Fleiss and Chilton; Fleiss et al.; Natkin and Guild; Reed, Feil, and Greer). Others, state that
standardization cannot help but be increased following rater training (Mackenzie et al.; Mackenzie 1974). Others show no evidence that faculty should be calibrated prior to student evaluation due to the lack of improvement in rater reliability (Weaver and Saeger; Fuller; Mescher and Kerber). The review of the literature is inconclusive regarding the value of calibration in faculty standardization.
CHAPTER 3
Methods and Materials

The purpose of this investigation was to determine the effects of calibration on dental hygiene faculty standardization. Twelve pair-matched subjects were randomly assigned to one of two groups; an experimental group which was calibrated and a control group which was not calibrated. Subjects completed instrumentation skill evaluation forms while participating in a mock clinical evaluation exercise. Completed skill evaluations were statistically analyzed for answer key agreement and patterns of rater reliability to form conclusions about standardization of evaluators following calibration.

Sample Description
Subjects were drawn from a convenience population of dental hygiene educators with at least one semester of clinical teaching experience. Ranges of teaching experience included less than one year (n=3), one year (n=2), two years to four years (n=5), five years to seven years (n=0), eight years to ten years (n=0), and more than 10 years (n=2). Although a majority of the subjects were faculty members and graduate students of Old Dominion University, previous faculty members and dental hygiene educators employed at other institutions were invited to participate. All
subjects held a minimum of a baccalaureate degree, were females except for one male, and were between the ages of 22 and 45.

**Research Design**

The randomized, matched subjects, posttest-only research design was utilized in this investigation (Darby and Bowen 68). Pair-matching of subjects was implemented for the purposes of this investigation to control for small sample size, absence of a pre-test, and to maintain group equivalency. Subjects were pair-matched based on years of clinical teaching experience and randomly assigned to the experimental or control groups via coin-toss. Clinical teaching experience was categorized as less than one year, one year, two years to four years, five years to seven years, eight years to ten years, and more than ten years. Group A (experimental) participated in receiving the independent variable, calibration, prior to completing the posttest; skill performance criteria contained on the instrumentation skill evaluation form represented the posttest. Group B served as the control group, not having been calibrated prior to administration of the post-test. The dependent variable in this investigation was standardization as measured by answer key agreement, intra-rater reliability, and inter-rater reliability. Anonymity of responses was maintained by a research assistant assigning each subject a letter code.
Methodology

Subjects were mailed an information packet containing a cover letter explaining their obligations to the study (see Appendix A), informed consent materials (see Appendix B), and instrumentation textbook materials relating to use of universal instruments. The cover letter requested all participants to familiarize themselves with the instrumentation textbook materials prior to the mock clinical evaluation exercise to achieve similar baseline knowledge of universal instrumentation skills. Completed forms required for pair-matching subjects based on experience and informed consent materials that were returned by mail were collected by a research assistant to further insure anonymity.

Pair-matching, random assignment of the subjects to the experimental or control groups, and assignment of letter code to subjects were performed by a research assistant. Subjects were randomly assigned to one of two groups; Group A (n=6), received the experimental treatment, calibration; Group B (n=6) served as the control group, and thus, did not receive calibration. Group B was asked to arrive at the mock clinical exercise one hour after Group A so that the latter group could participate in the calibration exercise.

The method of calibration used in the study drew upon suggestions found in the review of the literature (Edwards, Morse, and Mitchell 695; Mackenzie et al. 285; McNabb and
Zarkowski 25). Group A, the experimental group, participated in the calibration session for one hour prior to the arrival of Group B. To facilitate training, subjects in Group A were supplied with a Columbia 13/14 curet, a typodont, and a copy of the instrumentation textbook. The text was referred to throughout the calibration session. Subjects were instructed on how to complete the instrumentation skill evaluation form, instructed in all aspects of each skill performance criteria needing explanation, and viewed an instructional videotape. The videotape, "An Edge on Success: Techniques for the Successful Use of Universal Instruments," demonstrated correct skill performance. The principal investigator also demonstrated incorrect behaviors for each skill performance criteria using a typodont and the instrument being evaluated. Discussion followed in order to clarify ambiguities and address questions in the context of the skill evaluation. Appendix D contains a detailed lesson plan of the calibration exercise. The principal investigator was responsible for calibrating the group. Calibration of subjects proceeded for one hour at which time Group B was asked to join Group A to begin the mock clinical evaluation exercise.

The mock clinical evaluation exercise was conducted simultaneously for both groups. Although the composition of Groups A and B was obvious to the principal investigator at
the mock clinical evaluation exercise, the letter code provided anonymity for the subjects completing instrumentation skill evaluation forms. Prior to viewing the vignettes, subjects who had not signed their informed consent forms were asked to do so at this time. Subjects viewed the video vignettes through the use of a projection television.

The mock clinical evaluation exercise consisted of four vignettes depicting a clinician portraying a student performing instrumentation skills while activating a Columbia 13/14 curet on a live patient. Each vignette was approximately one minute in length and subjects were given two additional minutes to complete their evaluation. Subjects evaluated each vignette before the next one was shown. The second vignette (Vignette II) was shown again as the fourth vignette (Vignette IV) to provide an opportunity for testing intra-rater reliability; the subjects were unaware that the fourth vignette was a copy of the second vignette. Vignettes I and III were not considered in data analysis. Appendix E describes each vignette shown during the mock clinical evaluation exercise. Instrumentation was performed on the mandibular right posterior facial tooth surfaces and was selected for ease in filming and to more clearly demonstrate instrumentation skills for the viewer. Academic Television Services of Old Dominion University was consulted for expertise in filming small intraoral action in
Subjects evaluated universal curet instrumentation skills during the mock clinical exercise. According to Nield, the universal instrument requires more skill when compared to the Gracey curet in relation to angulation factors present in the two classifications. Gracey curets are said to be self-angulating; universal curets require the skill of the clinician to determine proper angulation (Nield and Houseman V). Although any universal curet could have been used, the Columbia 13/14 was chosen for its strict adherence to principals of universal curet design as well as its acceptability and familiarity by most clinicians. Subjects were instructed to evaluate the instrumentation skill performance seen on each vignette as they would in an average student instrumentation skill assessment setting.

Instrumentation skill evaluation forms were collected by a research assistant following completion of each vignette so that the subjects could not change their responses based on student comparisons. Subjects were given the opportunity to request a copy of the results as outlined in the informed consent materials and then dismissed.

Mackenzie and co-authors state that if rater reliability is between 80-100 percent, evaluators can be considered standardized (287). The authors contended that a range of 80-100 percent agreement in instructors' assessment of student ability is acceptable. The range between
instructors’ scores permits an allowable percentage of disagreement in the overall skill assessment. Results of the study were compared to the 80-100 percent agreement range as a measure of standardization based on Mackenzie’s recommendations.

**Protection of Human Subjects**

Review and acceptance of the research design and concept was obtained from the Committees for the Protection of Human Subjects in the Old Dominion University School of Dental Hygiene and Dental Assisting and the College of Health Sciences.

1. **Subject Population**: Subjects were selected from a population of dental hygiene educators who reside in close proximity to the research location. Requirements for subject selection included attainment of a Bachelor of Science degree as the minimum education level and a minimum of one semester experience as a clinical instructor in an accredited dental hygiene program.

2. **Potential Risks**: The principal investigator did not foresee any physical, social or psychological risks to the subjects. A principal investigator blind to the subjects' individual responses assisted in preventing any potential bias in data analyses (Darby and Bowen 46; Tuckman 96).

3. **Consent Procedures**: Subjects were provided with informed consent materials that clearly outlined their obligations to the study and emphasized that they would not
encounter any physical, psychological or social dangers by participating in this investigation. Consent materials are located in Appendix B. Subjects were asked to sign the informed consent prior to the mock clinical exercise if they had not already done so; copies were provided.

4. **Procedures to Maintain Confidentiality:** Results remained anonymous through the use of an letter code rather than name identification on the data collection instrument. A research assistant was responsible for assigning the letter code to each subject and of informing each subject of their code by including it on the letter of introduction. The principal investigator and statistician viewed only the codes after that time. Results were reported only in group form and subjects' responses were in no way connected with their identity.

5. **Potential Benefits:** This investigation can provide the research community with information pertaining to accurate assessment of student skills, fair grading, and calibration's necessity in objective evaluation techniques.

6. **Risk/Benefit Ratio:** The benefits resulting from this research far outweighed the minimal risks to subjects. All precautions to maintain confidentiality of the letter code identifying subjects and raw data were practiced.

**Instrumentation**

By permission of the author, the instrumentation skill evaluation form titled, "Skill Evaluation: Universal
Instruments, Mandibular Sextant," was used by the subjects as the data collection instrument and served as the dependant variable measure (Nield, 1990). Appendix C contains the instrumentation skill evaluation form used in this investigation. The instrument skill evaluation form contained the following areas: the student's and instructor's name; the date; a breakdown of the areas of the mouth to be evaluated; directions for using the form; 14 skill performance criteria to be evaluated by scoring a "C" for correct skill performance or an "I" for improvement needed; and three columns for student self-check, evaluator assessment, and evaluator re-check. Space was provided for the instructor's written comments and clarification of these comments on the back of the form. For the purposes of the investigation, only the evaluators' assessment column was utilized by the subjects during the mock clinical exercise.

Four instrumentation skill evaluation forms were provided for each subject during the mock clinical exercise. Subjects recorded their letter code in lieu of their names on the "instructor" line. Twenty-four completed instrumentation skill evaluation forms were available for analysis; twelve from Vignette II were analyzed for answer key agreement and inter-rater reliability while twelve from Vignette IV were compared with the data gathered from Vignette II to determine intra-rater reliability. Each instrumentation skill evaluation form was assessed for
completeness and accuracy; improperly completed forms were not considered in the data analysis. Answers contained on complete, acceptable forms were compared to an answer key of errors that the clinician actually committed and which the subject should have discerned while viewing the mock clinical exercise (see Appendix E).

The following is a description of each of the fourteen skill performance criteria; #1: "uses all criteria for patient and operator positioning,"; #2: "uses all criteria for grasp,"; #3: "uses all criteria for mirror and finger rests,"; #4: "uses correct working end," #5: "inserts at zero degrees, if appropriate,"; #6: "handle and shank parallel to long axis of tooth,"; #7: "uses tip-third (toe third) of cutting edge,"; #8: "establishes correct tooth-to-cutting edge angulation,"; #9: "rolls handle to maintain adaptation,"; #10: "uses controlled working strokes,"; #11: "uses appropriate stroke length,"; #12: "applies stroke pressure in coronal direction only,"; #13: "uses overlapping strokes,"; #14: "uses wrist/arm activation to produce controlled strokes" (Nield and Houseman 311). Criteria #1 and #3 were not assessed due to inability to include positioning and use of the mouth mirror on the videotape.

Statistical Analysis

Subjects were requested to record a "C" (correct) or "I" (incorrect) during their assessment of the twelve criteria of the videotaped skill performance. Data
collected from this investigation, therefore, was discrete and nominal and recorded in group form only. Analysis of data included a z-test for both answer key agreement and intra-rater reliability, and the chi-square test of independence for inter-rater reliability.

The z-test determines whether the mean performance of two groups is significantly different when a sample is less than 20 (Popham and Sirotnik 124). The chi-square test of independence examines whether or not there is a relationship between and among variables; the direction of the relationship is determined by analyzing the cell frequencies. Two-tailed tests were applied in all instances to test the null hypotheses. Alpha level was set at 0.05.

Of critical importance to the interpretation of results is the differentiation of answer key agreement and rater reliability. Intra- and inter-rater reliability were not based on agreement of the subjects' answers with the answer key, but rather on the agreement of responses among and between raters. Intra-rater reliability of subjects was determined without consideration for whether or not the subject agreed with the answer key, as was inter-rater reliability. Only their responses were assessed.

Answer key agreement was based on dividing the total number of correct responses for the group by the total number of possible responses (72). This value was used in the z-test formula to derive a test statistic. Intra-rater
reliability was determined by dividing the number of times each member of the group scored Vignettes II and IV identically by the total number of possible agreements (72). This product was used in the z-test formula to arrive at a test statistic. Inter-rater reliability was calculated by determining the frequency of agreement among evaluators for their recorded assessments of Vignette II. These frequencies constituted the data to be inserted into the chi-square cells. The control and experimental groups' responses were assessed to determine if their reliability and answer key agreement were between 80 and 100 percent. Percentage range was established based on the work of Mackenzie et al. (287).
CHAPTER 4

Results and Discussion

The purpose of this investigation was to determine the impact of calibration on dental hygiene faculty standardization. Twenty-six dental hygienists with clinical teaching experience were invited to participate. Five dental hygienists failed to respond to the request while six were unable to participate due to scheduling conflicts. Fifteen subjects participated in the mock clinical evaluation exercise. Due to the uneven number, one subject was asked to serve as a backup in the event that another subject was absent; the backup subject was called to participate. Half of the subjects were exposed to the experimental treatment, calibration, prior to instrumentation skill evaluation. Following completion of the mock clinical evaluation exercise by the experimental and control groups, instrumentation skill evaluation forms were evaluated for completion. One form from each group was found to be completed incorrectly, therefore, data from these forms were not considered in analysis. Statistical analysis included responses made by six subjects from the experimental group (calibrated), and six from the control group (non-calibrated). Table 2 illustrates the clinical
Table 2
Clinical Teaching Experience of Subjects by Group

<table>
<thead>
<tr>
<th></th>
<th>Calibrated</th>
<th>Non-calibrated</th>
</tr>
</thead>
<tbody>
<tr>
<td>less than one year</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>one year</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>two years to four years</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>five years to seven years</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>eight years to ten years</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>more than ten years</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td><strong>6</strong></td>
<td><strong>6</strong></td>
</tr>
</tbody>
</table>
teaching experience of subjects by groups.

Results

The investigation was designed to test three hypotheses. Each was aimed at determining the effects of calibration on standardization of evaluators during assessment of instrumentation skills. Standardization was measured via answer key agreement, intra-rater reliability, and inter-rater reliability.

The first hypothesis states, "There is no statistically significant difference at the 0.05 level of significance between calibrated and non-calibrated evaluator groups as measured by skill performance criteria answer key agreement during a mock clinical evaluation exercise." Table 3 illustrates the number of subjects from calibrated and non-calibrated groups achieving answer key agreement based on the established criteria. Established criteria contained on the instrumentation skill evaluation form are presented in Appendix C. Results from calibrated subjects revealed answer key agreement was achieved by four subjects for skill performance criterion two, six subjects for criterion four, two subjects for criterion five, two subjects for criterion six, four subjects for criterion seven, four subjects for criterion eight, two subjects for criterion nine, six subjects for criteria ten and eleven, one subject for criterion twelve, and four subjects for criteria thirteen and fourteen.
Table 3

Number of Calibrated and Non-calibrated Subjects Achieving Answer Key Agreement for Individual Criteria

<table>
<thead>
<tr>
<th>Criteria Item*</th>
<th>Calibrated Group (n=6)</th>
<th>Non-calibrated Group (n=6)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>7</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>8</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>9</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>10</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>11</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>12</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>13</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>14</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>TOTAL</td>
<td>45</td>
<td>42</td>
</tr>
</tbody>
</table>

*Criteria items #1 and #3 were not analyzed.
Results from non-calibrated subjects revealed answer key agreement was achieved by one subject for criterion two, six subjects for criterion four, two subjects for criterion five, three subjects for criterion six, five subjects for criterion seven, two subjects for criterion eight, three subjects for criterion nine, five subjects for criterion ten, four subjects for criterion eleven, two subjects for criterion twelve, three subjects for criterion thirteen, and six subjects for criterion fourteen. Calibrated subjects achieved higher answer key agreement than the non-calibrated group for five of twelve skill performance criteria (two, eight, ten, eleven, and thirteen), concurrently, non-calibrated subjects achieved higher answer key agreement than the calibrated group for five criteria (six, seven, nine, twelve, and fourteen). Overall, calibrated subjects achieved perfect answer key agreement for three of the twelve skill performance criteria (four, ten, and eleven); whereas, non-calibrated subjects achieved perfect answer key agreement for two of the twelve criteria (four and fourteen).

Table 4 illustrates total answer key agreement scores for both groups. Raw data demonstrates that of 72 possible correct responses, calibrated subjects answered correctly for 45 responses; non-calibrated subjects answered correctly for 42 responses. A z-test was applied to determine a statistically significant difference between the
## Table 4

Total Answer Key Agreement Scores for the Calibrated and Non-calibrated Groups

<table>
<thead>
<tr>
<th>Responses Compared with Answer Key</th>
<th>Total Criteria(n)</th>
<th>Responses Achieving Answer Key Agreement(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Agreed</td>
<td>Disagreed</td>
<td></td>
</tr>
<tr>
<td>Calibrated</td>
<td>45</td>
<td>27</td>
</tr>
<tr>
<td>Non-Calibrated</td>
<td>42</td>
<td>30</td>
</tr>
<tr>
<td>TOTAL</td>
<td>94</td>
<td>50</td>
</tr>
</tbody>
</table>

\[z\text{-value} = 0.5116\]

\[p\text{-value} = 0.305*\]

*not significant at the 0.05 alpha level
experimental and control groups. No statistically significant difference was revealed between calibrated and non-calibrated groups for answer key agreement ($z=0.5116; p=.305$). The null hypothesis was retained at the 0.05 alpha level.

The second hypothesis states, "There is no statistically significant difference at the 0.05 level of significance in faculty standardization between calibrated and non-calibrated evaluator groups as measured by skill performance criteria intra-rater reliability during two consecutive identical mock clinical evaluation exercises."

Table 5 illustrates the number of subjects from both the calibrated and non-calibrated groups achieving intra-rater reliability based on individual criteria. Results from calibrated subjects revealed intra-rater reliability was achieved by five subjects for criteria two and four; two subjects for criterion five; five subjects for criteria six, seven, and eight; three subjects for criterion nine; six subjects for criteria ten and eleven; five subjects for criterion twelve; four subjects for criterion thirteen; and three subjects for criterion fourteen.

Results from non-calibrated subjects revealed intra-rater reliability was achieved by four subjects for criterion two, five subjects for criterion four, four subjects for criterion five, one subject for criterion six,
### Table 5

**Number of Calibrated and Non-calibrated Subjects Achieving Intra-rater Reliability for Individual Criteria**

<table>
<thead>
<tr>
<th>Criteria Item*</th>
<th>Calibrated Group (n=6)</th>
<th>Non-calibrated Group (n=6)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>6</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>8</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>9</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>10</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>11</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>12</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>13</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>14</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td><strong>54</strong></td>
<td><strong>46</strong></td>
</tr>
</tbody>
</table>

*Criteria items #1 and #3 were not analyzed*
four subjects for criterion seven, three subjects for criterion eight, four subjects for criterion nine, five subjects for criterion ten, four subjects for criterion eleven, five subjects for criterion twelve, four subjects for criterion thirteen, and three subjects for criterion fourteen. The calibrated and non-calibrated groups had similar intra-rater reliability for four criteria (four, twelve, thirteen, and fourteen). Calibrated subjects had higher intra-rater reliability than the non-calibrated group for six of the twelve skill performance criteria (two, six, seven, eight, ten, and eleven). Non-calibrated subjects had higher intra-rater reliability than the calibrated group for two of the twelve criteria (five and nine). The calibrated group achieved perfect intra-rater reliability for two of the twelve criteria (ten and eleven); the non-calibrated group did not achieve perfect intra-rater reliability for any criteria.

Table 6 illustrates total intra-rater reliability scores for both groups. Raw data showed that of 72 possible correct responses, calibrated subjects achieved intra-rater reliability 54 times. Non-calibrated subjects achieved intra-rater reliability 46 times. Application of the z-test revealed no statistically significant differences between the calibrated and non-calibrated evaluator groups for intra-rater reliability (z= 1.45; p=.0738). The
Table 6

Total Intra-rater Reliability Scores for the Calibrated and Non-Calibrated Groups

<table>
<thead>
<tr>
<th></th>
<th>Reliable(n)</th>
<th>Not Reliable(n)</th>
<th>Total Criteria(n)</th>
<th>Intra-rater Reliability(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Calibrated</td>
<td>54</td>
<td>18</td>
<td>72</td>
<td>75</td>
</tr>
<tr>
<td>Non-Calibrated</td>
<td>46</td>
<td>26</td>
<td>72</td>
<td>64</td>
</tr>
<tr>
<td>TOTAL</td>
<td>100</td>
<td>44</td>
<td>144</td>
<td></td>
</tr>
</tbody>
</table>

z-value = 1.45

p-value = .0738*

*not significant at the 0.05 alpha level
null hypothesis was retained at the 0.05 alpha level.

The third hypothesis states, "There is no statistically significant difference at the 0.05 level of significance in faculty standardization between calibrated and non-calibrated evaluator groups as measured by skill performance criteria inter-rater reliability during a mock clinical evaluation exercise." Table 7 illustrates calibrated and non-calibrated subjects responses for each skill performance criteria. Results from calibrated subjects revealed for criterion two, two subjects responded "correct," and four subjects responded "incorrect"; for criterion four, all subjects responded "correct"; for criterion five, two subjects responded "correct," and four subjects responded "incorrect"; for criteria six and seven, four subjects responded "correct," and two subjects responded "incorrect"; for criteria eight and nine, two subjects responded "correct," and four subjects responded "incorrect"; for criteria ten and eleven, all subjects responded "incorrect"; for criterion twelve, five subjects responded "correct," and one subject responded "incorrect"; for criterion thirteen, two subjects responded "correct," and four subjects responded "incorrect"; and for criterion fourteen, four subjects responded "correct," and two subjects responded "incorrect."

Non-calibrated subjects responses revealed for
Table 7
Calibrated and Non-calibrated Subjects' Responses to Criteria Items

<table>
<thead>
<tr>
<th>Criteria Item*</th>
<th>Calibrated Group (n=6)</th>
<th>Non-calibrated Group (n=6)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>&quot;C&quot;</td>
<td>&quot;I&quot;</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>6</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>7</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>8</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>9</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>10</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td>11</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td>12</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>13</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>14</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>TOTAL</td>
<td>33</td>
<td>39</td>
</tr>
</tbody>
</table>

*Criteria items #1 and #3 were not analyzed.

^ indicates higher inter-rater reliability between the calibrated and non-calibrated groups.

"C" indicates response of "correct"
"I" indicates response of "incorrect"
criterion two, five subjects responded "correct," and one subject responded "incorrect"; for criterion four, all subjects responded "incorrect"; for criterion five, two subjects responded "correct," and four subjects responded "incorrect"; for criterion six, three subjects responded "correct," and three subjects responded "incorrect"; for criterion seven, five subjects responded "correct," and one subject responded "incorrect"; for criterion eight, four subjects responded "correct," and two subjects responded "incorrect"; for criterion nine, three subjects responded "correct," and three subjects responded "incorrect"; for criterion ten, one subject responded "correct," and five subjects responded "incorrect"; for criterion eleven, two subjects responded "correct," and four subjects responded "incorrect"; for criteria twelve, four subjects responded "correct," and two subjects responded "incorrect"; for criterion thirteen, three subjects responded "correct," and three subjects responded "incorrect"; and for criterion fourteen, all subjects responded "correct."

The groups achieved similar inter-rater reliability for three criteria (four, five, and eight). Calibrated subjects achieved higher inter-rater reliability for five criteria (six, nine, ten, twelve, and thirteen). Non-calibrated subjects achieved higher inter-rater reliability for three
criteria (two, seven, and fourteen). Perfect inter-rater reliability was achieved by calibrated subjects for three criteria (four, ten, and eleven), while non-calibrated subjects achieved perfect inter-rater reliability for two criteria (four and fourteen).

The chi-square test of independence was applied to the data to determine if there was a relationship between calibration and inter-rater reliability. Chi-square was 2.80 (df=1; p=.0475); based on these results, the null hypothesis was rejected. Statistical computation revealed a significant relationship existed between calibration and inter-rater reliability, however, the non-calibrated subjects responded more similarly than calibrated subjects. Table 8 summarizes the statistical analysis.

Discussion

Results of this study provide a basis for discussing the effect of calibration on faculty standardization as measured by answer key agreement, intra-rater reliability and inter-rater reliability. The experimental treatment consisted of calibrating a group of evaluators prior to assessment of instrumentation skills. Skill performance evaluation assessments collected during a mock clinical evaluation exercise were statistically analyzed. Of three null hypotheses investigated, two were retained, one was rejected.
Table 8
Chi-square Analysis for Inter-rater Reliability

<table>
<thead>
<tr>
<th></th>
<th>Answered &quot;C&quot; (n)</th>
<th>Answered &quot;I&quot; (n)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-Calibrated</td>
<td>44</td>
<td>28</td>
</tr>
<tr>
<td>Calibrated</td>
<td>33</td>
<td>39</td>
</tr>
</tbody>
</table>

\[ \chi^2 = 2.08 \]
\[ df = 1 \]
\[ p-value = 0.0475^* \] (two-tailed test)

*statistically significant at the 0.05 level
Analysis of the ability of calibrated and non-calibrated evaluators to achieve answer key agreement revealed no statistically significant differences existed between the groups. Results suggest that calibration prior to instrumentation skill assessment will not enable evaluators to distinguish between correct and incorrect instrumentation skill performance.

Results might have been influenced by chance agreement because only two assessment responses (i.e. correct or incorrect) were available for selection by subjects. Mackenzie, et al. state that when evaluators have only two assessment responses to choose from, evaluators can be reliable at least 50 percent of the time if they make guesses and evaluate nothing (287). They also suggest that with a small number of assessment responses, there will be an unavoidable increase in the percent of agreement among evaluators.

Additionally, extraneous variables might have affected the results of answer key agreement. For example, experience of evaluators may have played a role in their ability to assess instrumentation skills in a coincidentally standardized manner. Daniel, et al. give evidence that experience in clinical evaluation result in standardized assessment of student skill performance (412). Fuller suggests that rater reliability increases as evaluator
experience increases (21). Experience in evaluation also was discussed by Biller and Kerber who give evidence that a dental hygiene educator's best calculus detection skills occur within nine months to five years in clinical teaching (209). Further, the educational backgrounds of evaluators might have affected the results (Biller and Kerber 209). Subjects in this investigation were pair-matched based on years of clinical teaching experience to control for this variable.

Additionally, results may have been affected by the discernability of the various criteria (see Appendix C for specific criteria). For example, subjects may have easily discerned if an incorrect working end was utilized but had difficulty noticing if the instrument was inserted at zero degrees. Criterion four, "uses correct working end," was easily discriminated by all subjects in both the calibrated and non-calibrated groups. Criteria ten, "uses controlled working stroke," and eleven, "uses appropriate stroke length," were easily discerned by the calibrated group. Criterion fourteen, "uses wrist/arm activation to produce controlled working strokes," was easily discerned by the non-calibrated group. The slightly higher ability of calibrated subjects to achieve answer key agreement may have occurred because of criteria that were more easily discernable.
The nationally accepted instrumentation skill evaluation form chosen for the purposes of this investigation might not have been a valid or reliable data collection instrument. Flaws could have existed within the skill performance criteria resulting in ambiguities and therefore, misinterpretation by subjects. For the sake of providing students with valid and reliable clinical skill assessment, future research should determine skill performance criteria with unquestionable validity and reliability for instrumentation skill evaluation forms.

The small number of subjects participating in this investigation also could have masked statistically significant differences. Analysis of data was based on six instrumentation skill evaluation forms containing twelve skill performance criteria for a total of 72 possible responses for each group. All attempts were made to involve as many dental hygiene educators as subjects as possible. When comparing this study to others addressing the effect of calibration on standardization, this investigation employed a relatively large sample population.

Results of this investigation can be generalized only to a population with similar characteristics and when using identical calibration exercise, vignettes, and answer key. If clinicians not involved with this study were to view the vignettes and develop a new answer key, it is possible that
results might vary from this study. The design of this study can be applied easily to other dental hygiene programs using the textbook, *Fundamentals of Dental Hygiene Instrumentation* by Nield and Houseman and a two assessment point, competency based evaluation system.

One quantifiable method of determining whether subjects achieved standardization is to apply the 80-100 percent agreement range based on the work of Mackenzie, et al. (287). Figure 1 illustrates the ability of calibrated and non-calibrated subjects to achieve standardization based on 80-100 percent answer key agreement. No subjects in either group obtained this range. Results from the calibrated group revealed two subjects achieved 75 percent answer key agreement, one subject achieved 67 percent answer key agreement, one subject achieved 58 percent answer key agreement, and two subjects achieved 50 percent answer key agreement. Results from non-calibrated subjects revealed one subject achieved 75 percent answer key agreement, two subjects achieved 67 percent answer key agreement, one subject achieved 58 percent answer key agreement, one subject achieved 50 percent answer key agreement, and one subject achieved 33 percent answer key agreement.

These results suggest that when measuring standardization based on 80-100 percent answer key agreement, the highest score of 75 percent was obtained by
Figure 1

Calibrated and Non-calibrated Subjects
Achieving Standardization Based on
80-100 Percent Answer Key Agreement

Percent of Answer Key Agreement
(n= 12)

■ =Calibrated Group
(n= 6)

□ =Non-Calibrated Group
(n= 6)
two calibrated subjects, while the lowest score of 33 percent was obtained by a non-calibrated subject. It would appear that overall, more calibrated subjects were able to obtain higher scores than non-calibrated subjects, although these differences were not statistically significant. This information provides more evidence that calibration does not make a difference in ability to be standard because no subjects achieved the 80-100 percent range.

The second hypothesis examined if statistically significant differences existed in standardization between calibrated and non-calibrated evaluators as measured by intra-rater reliability. Analysis of data resulted in failure to reject the null hypothesis. Results suggest that evaluators will be reliable within themselves whether they are calibrated or not. The potential for chance agreement may have influenced the results of this investigation of intra-rater reliability. In support of this phenomenon, Mackenzie states that if only two assessment responses exist, and if evaluators can prove to be consistent over time in their assessments when using only two responses, then they may be considered standardized (1974, 219). Perhaps several subjects ascertained that the identical vignette was shown twice and recalled their first assessment while responding to the second viewing of the vignette. Similarly, with answer key agreement, the subtlety of errors present in the vignettes may not have been detected easily
by both the calibrated and non-calibrated groups.

Another limitation affecting the validity of the results might have been the characteristics of the subjects. Results suggest that the groups had a somewhat evenly distributed range of experience which may have controlled for any possible variations in intra-rater reliability. In future investigations of this nature, it is suggested that subjects be pair-matched to account for experience with process evaluation and not solely for years of experience.

Figure 2 represents the ability of subjects to achieve standardization based on 80-100 percent intra-rater reliability. Three calibrated subjects obtained scores between 80 and 100 percent while one non-calibrated subject obtained a score within this range. Results from calibrated subjects revealed one subject achieved a high score of 92 percent intra-rater reliability, two subjects achieved 83 percent intra-rater reliability, two subjects achieved 67 percent intra-rater reliability, and one subject achieved 58 percent intra-rater reliability. Results from the non-calibrated subjects revealed one subject achieved 83 percent intra-rater reliability, two subjects achieved 75 percent intra-rater reliability, one subject achieved 58 percent intra-rater reliability, one subject achieved 50 percent intra-rater reliability, and one subject achieved 42 percent intra-rater reliability.

By applying 80-100 percent intra-rater reliability as
Figure 2

Calibrated and Non-calibrated Subjects
Achieving Standardization
Based on 80-100 Percent Intra-Rater Reliability

Number of Subjects

Percent of Intra-rater Reliability
(n= 12)

■ =Calibrated Group
(n= 6)

□ =Non-Calibrated Group
(n= 6)
another measure of standardization, 50 percent of the calibrated subjects achieved standardization while only one subject from the non-calibrated group achieved standardization. This result suggests that calibrated evaluators achieved greater intra-rater reliability than non-calibrated counterparts. Although no statistically significant differences were revealed, these results suggest that calibration might have had an effect on intra-rater reliability.

The third hypothesis was designed to determine if statistically significant differences existed in the ability of the calibrated and non-calibrated groups to achieve standardization as measured by inter-rater reliability. Results suggest that there is a relationship between calibration and inter-rater reliability. Results indicate that the non-calibrated group was more successful in achieving inter-rater reliability. Calibration had no effect on increasing inter-rater reliability among evaluators in this investigation.

Possible explanations for these results include the likelihood that calibrated subjects may have experienced the Hawthorne effect in the form of test anxiety. This anxiety might have caused the group to be unreliable in their judgments. Calibrated subjects also may have become overwhelmed with information as a result of the calibration exercise, or "read into" the skill performance criteria.
Additionally, the calibrated group may have analyzed the ability of the student portrayed on the vignettes more critically than did evaluators in the non-calibrated group as a result of participating in the calibration exercise. The greater numbers of responses of "incorrect" for skill performance criteria within the calibrated group could be evidence of this phenomenon.

The elements of chance and whether or not the skill performance criteria were easily distinguishable when subjects viewed the vignettes cannot be overlooked when interpreting these results. Possibly, subjects in the non-calibrated group might have been coincidentally reliable among each other where the calibrated subjects were not reliable among each other. Those subjects with experience in pre-clinical instrumentation evaluation may have had an advantage over those subjects without this type of experience or those more experienced in product evaluation. Furthermore, evaluators experienced in task analysis, or clinical skill performance evaluation, coincidentally may have been in the non-calibrated group, thereby, increasing the inter-rater reliability of the group due to their pre-clinical experience. Some evaluators in the non-calibrated group might have been inherently "task analysis" oriented which may have caused them to achieve higher inter-rater reliability.

The chi-square test of independence chosen as the
statistical test for this investigation might not have been sensitive enough to discriminate fine differences in inter-rater reliability due to the small sample size. Traditionally, correlation coefficients such as Pearson's Product Moment correlation or Spearman's Rank Order Correlation tests are used as tests of inter-rater reliability. Chi-square was selected as the most appropriate test for its ability to provide the analysis of information in support of the research question pertaining to inter-rater reliability.

The lack of higher inter-rater reliability in the calibrated group during this investigation conflicts with the work of Edwards, Morse, and Mitchell (695); Fleiss, et al. (608); Fleiss and Chilton (605); and Reed, Feil, and Greer (556), where inter-rater reliability was shown to improve following calibration. The lack of higher inter-rater reliability for calibrated subjects in this investigation is supported by investigations by Fuller (22); Hinkleman and Long (18); Natkin and Guild (160); and Weaver and Saeger (255), where calibration did not effect inter-rater reliability. The findings of Mescher and Kerber also support results of this investigation. They found that discrepancies in assessment of clinical skill performance exist when more than one evaluator assesses student skills (86).

A calibration exercise procedure such as a discussion
of discrepancies in judgement among evaluators until reaching a consensus agreement should be considered for future investigations in order to increase inter-rater reliability. This technique was used in investigations by Edwards, Morse, and Mitchell (694); Fleiss, et al. (606); Fuller (24); Mackenzie, et al. (285); McNabb and Zarkowski (23); Natkin and Guild (156); and Reed, Feil, and Greer (555). Mackenzie, et al. also suggest that frequent calibration will form changes in behavior (288). Perhaps variations in the calibration exercise would yield different results.

Previous studies investigating the effect of calibration on inter-rater reliability address length and frequency of evaluator calibration. Weaver and Saeger discovered in an extensive literature review of calibration programs that in the majority of studies where inter-rater reliability was increased, an active training element existed where evaluators practiced their assessment skills for several days or weeks (225). Yet, in their investigation, calibration was conducted for four hours without increasing inter-rater reliability (253). Fuller calibrated subjects for two hours without increasing inter-rater reliability (21). Natkin and Guild had no increase in inter-rater reliability after eight meetings of an unspecified length of time (158).

Conversely, Fleiss, et al. found evaluators to be
reliable after a two day calibration exercise (606) and after a meeting each week for eight weeks Bazan and Seale calculated that evaluators were only 70 percent reliable during product evaluation (727). Mackenzie, et al. calibrated subject-evaluators weekly for one and one half hours for four months, however, the authors did not report reliability calculations (285). Supported by these investigations, the one hour calibration exercise utilized in the study under investigation would not be expected to improve inter-rater reliability.

Reed, Feil, and Greer discuss the differences between inter-rater agreement and inter-rater reliability (556). They state that reliability measurements will be low when variance and error are small. However, subjects can be in acceptable agreement even though analysis suggests low inter-rater reliability when calculated statistically. They contend that there is a difference between reliability and agreement and that researchers should be specific as to which variable they are testing. For the purposes of this investigation, agreement among evaluators was operationally defined as inter-rater reliability. Perhaps if inter-rater agreement had been explored, results of the investigation could have been different.

Results of the calibrated subjects’ abilities to achieve inter-rater reliability may have been affected by factors contributing to disagreement identified by
Mackenzie, et al (285). Possibly flaws existed in the operational definitions of the skill performance criteria resulting in checkpoint ambiguity. For example, subjects may have found the criteria measuring controlled strokes and strokes overlapping to be ambiguous. Unspecified exceptions may not have been covered sufficiently during the calibration exercise. For example, parallel position of the handle and shank may have been difficult to achieve if the patient's mouth did not open wide and could count as an exception. The vignettes may have been unclear visually to some of the subjects, and resulted in discrepancies in visual acuity. Some subjects may have given the student the benefit of the doubt resulting in degrees of leniency. Lastly, when bringing such a diversified group of professionals together, there will be bound to be differences in educational background. Many of these factors were difficult to control.

A lack of increase in inter-rater reliability can be expected from replication studies if evaluators in the experimental group experience the same limiting factors as in this investigation. Results of the investigation are inconclusive as to whether subjects in the calibrated group actually became aware of discriminations and ambiguities for more effective evaluation when assessing the use of the universal curet as a result of the calibration exercise. Results of the effects of calibration on inter-rater
reliability suggest that awareness of discriminations and ambiguities were not enhanced by calibration, but rather hampered by the application of the independent variable.

Overall, inter-rater reliability among subjects is presented in Figure 3. Results revealed that calibrated subjects achieved 100 percent inter-rater reliability for three criteria, 83 percent inter-rater reliability for one criteria, and 66 percent inter-rater reliability for eight criteria. Non-calibrated subjects achieved 100 percent inter-rater reliability for two criteria, 83 percent inter-rater reliability for three criteria, 66 percent inter-rater reliability for four criteria, and 50 percent inter-rater reliability for three criteria. When considering perfect reliability, calibrated subjects were able to achieve perfect reliability for three criteria; non-calibrated subjects achieved perfect reliability for two criteria. When considering the ability of the calibrated evaluators to achieve a range of 80-100 percent standardization, this measure of standardization was reached for four criteria; the non-calibrated subjects obtained this measure for five criteria. Results of determining whether subjects were standardized based on their ability to achieve 80-100 inter-rater reliability resulted in certain subjects being standardized.

When individual subject inter-rater reliability was examined, several evaluators were found to be consistently
Figure 3
Calibrated and Non-Calibrated Subjects
Achieving Standardization
Based on 80-100 Percent Inter-rater Reliability

*Criteria 1 and 3 were not utilized
reliable as a pair. Two calibrated evaluator pairs were reliable with each other at a level of 86 percent. If evaluators show consistency in evaluation over time, they may be considered to have high inter-rater reliability.

Mackenzie suggests that certification of evaluators should be performed in order to make valid judgments of student skill performance (1974, 219). Certification can provide a mechanism for ensuring that high reliability and agreement exists in student skill assessment. Due to the questionable benefits of calibration when measuring inter-rater reliability obtained from this investigation, Mackenzie’s suggestion might be beneficial to faculty members assessing pre-clinical instrumentation skills. Based on the results of this investigation, certification of evaluators may or may not include calibration. Perhaps other methods of ensuring reliability and agreement of evaluators could be explored in future investigations. Certification of evaluators could be used as a faculty development approach or as a requirement for new members of clinical dental hygiene programs.

The success of the calibration exercise is difficult to determine when analyzing individual instrumentation criteria. One can only speculate its success. Calibrated subjects achieved higher answer key agreement, intra-rater reliability or inter-rater reliability for the select criterion addressed below. Calibration may have had a very
positive effect on criterion four, ten, and eleven as calibrated subjects achieved perfect answer key agreement, intra-rater reliability, and inter-rater reliability. These criteria address correct working end, controlled strokes, and appropriate stroke length respectively. Calibration also seems to have effected answer key agreement and intra-rater reliability for criterion two, addressing grasp, intra-rater reliability and inter-rater reliability for criterion six addressing ability to keep the curet parallel, intra-rater reliability for criterion seven addressing area of the cutting edge being used, answer key agreement and intra-rater reliability for criteria eight addressing angulation, inter-rater reliability for criterion twelve addressing use of coronal pressure, and answer key agreement and inter-rater reliability for criterion thirteen addressing overlap of strokes.

Results suggest that further research should investigate methods of increasing standardization of those criteria which non-calibrated subjects achieved higher answer key agreement, intra-rater reliability or inter-rater reliability. Criterion nine, addressing handle rolling in which non-calibrated subjects achieved better scores in all three measures and criterion fourteen, addressing wrist/arm activation in which were noted perfect answer key agreement and inter-rater reliability. Further, non-calibrated subjects had higher inter-rater reliability for criterion
two, addressing grasp; intra-rater reliability for criterion five, addressing insertion at zero degrees; answer key agreement for criterion six, addressing ability to keep the curet parallel; inter-rater reliability for criterion seven, addressing area of the cutting edge being used; inter-rater reliability for criterion eight, measuring correct angulation; and answer key agreement for criterion twelve, addressing coronal pressure.

Limitations of this investigation must be discussed when interpreting results. A possible limitation of the calibration exercise was the absence of a measurement of learning resulting from the application of the independent variable, calibration. Perhaps a pretest, posttest design could be implemented to assess effects of calibration on improving evaluation skills. A number of studies contained in the literature review employed similar pretest, posttest methodology (Daniel, et al., 311; Fleiss and Chilton 605; Hinkleman and Long 15; Reed, Feil, and Greer 555; Weaver and Saeger 254). Inclusion of learning evaluation in the research design would increase generalizability of investigation results.

Pair-matching was an attempt to equalize subject experience within the groups. Experience previously has been shown to impact on evaluators' ability to assess student performance in a more standardized manner (Biller and Kerber 209; Daniel, et al. 412; Fuller 21). Table 2
illustrates that although pair-matching was attempted, groups were not even. Only subjects with one year or more than ten years were perfectly pair-matched, where the categories of less than one year and two to four years of teaching experience contained odd numbers of subjects. Further manipulation of subjects participating in the investigation so that years of clinical teaching experience is disclosed prior to an invitation to participate could pair-match groups for more accurate group equivalency.

Another limitation to this investigation was the involvement of the principal investigator in application of the experimental variable. Walsh and Phelps-Sandall state that subjectivity is decreased if the principal researchers involvement is minimal (14). In this investigation, the principal investigator was responsible for conducting the calibration exercise.
CHAPTER 5
Summary and Conclusions

An investigation was conducted to determine the effects of calibration on evaluator standardization as measured by answer key agreement, intra-rater reliability and inter-rater reliability. Twelve dental hygiene educators were pair-matched and randomly assigned to a control or experimental group based on years of clinical teaching experience. Subjects in the experimental group were calibrated for one hour which included demonstration and discussion of correct and incorrect instrumentation skills, skill performance criteria contained on a nationally accepted instrumentation skill evaluation form, and areas of possible misinterpretation and ambiguity. After participating in the calibration exercise, it was assumed that the subjects in the experimental group were standardized for evaluation of the Columbia 13/14 curet.

Following the calibration exercise, subjects in the experimental group were asked to join subjects in the control group who were not calibrated, for a mock clinical evaluation exercise. The mock clinical evaluation exercise consisted of instructions on how the instrumentation skill evaluation form was to be completed and a videotape viewing
of four vignettes, shown with the aid of a projection television. Subjects were instructed to evaluate each vignette as if assessing pre-clinical instrumentation skills. Subjects viewed a vignette, were given two additional minutes to complete the instrumentation skill evaluation form based on their assessment of what they had just seen, and then forms for the vignette were collected by a research assistant. This procedure was repeated until all four vignettes were viewed and evaluated by subjects.

Data used in statistical analysis consisted of the subjects' responses on the various skill performance criteria contained on the instrumentation skill evaluation form. Only the second vignette was considered for data analysis; the fourth vignette was a duplicate of the second vignette repeated in order to explore intra-rater reliability. The two responses from which the subjects could choose were "correct" or "incorrect." Answer key agreement was measured by comparing calibrated and non-calibrated subjects' responses with a pre-determined answer key. A z-test was applied for statistical analysis. Intra-rater reliability was measured by comparing each subject's responses to the second and fourth vignette. A z-test was employed for statistical analysis. Inter-rater reliability was measured by calculating the number of times subjects in each group responded "correct," and the number of times subjects in each group responded "incorrect," and
applying the chi-square test of independence to this ratio.

Null hypotheses were retained for the effect of calibration on standardization as measured by answer key agreement and the effect of calibration on standardization as measured by intra-rater reliability. The null hypothesis was rejected for the effect of calibration on standardization as measured by inter-rater reliability. Results of statistical analysis via chi-square revealed a relationship between calibration and inter-rater reliability. Upon interpretation of group results, it was determined that the non-calibrated group achieved inter-rater reliability more often than the calibrated group.

An unhypothesized area of interest was the ability of calibrated and non-calibrated subjects to obtain 80-100 percent agreement in the three measures of standardization. Evaluators did not have to be perfectly reliable to be considered standardized. Results of this measure of standardization revealed that no subjects from either group were able to obtain this range when measuring answer key agreement, three calibrated subjects and one non-calibrated subject achieved this range when measuring intra-rater reliability, and four calibrated subjects and five non-calibrated subjects achieved this measure of standardization. Results of this investigation could have been affected by the small sample size, chance agreement,
experience of the evaluators, discernability of the various criteria, and effectiveness of the calibration exercise.

Conclusions

Based on the discussion of this investigation, the following conclusions are offered:

1. Calibration of evaluators had no statistically significant effect on their ability to achieve answer key agreement.

2. Calibration of evaluators had no statistically significant effect on intra-rater reliability.

3. Calibration of evaluators had no statistically significant effect on inter-rater reliability.

4. Applying a range of 80-100 percent reliability as an acceptable measure of standardization, evaluators were not standard when considering answer key agreement. Several evaluators achieved standardization when applying the 80-100 percent range to intra-rater reliability and inter-rater reliability.

The following suggestions for future research to improve evaluator standardization are offered:

1. The investigation should be repeated to include other dental hygiene programs to increase the sample size and determine validity of this investigation.

2. Years of experience in clinical teaching should be correlated with the evaluators’ ability to achieve answer key agreement, intra-rater reliability, and inter-rater
reliability.

3. Type of evaluator experience, such as experience with process evaluation in pre-clinic versus product evaluation at higher learning levels, should be determined prior to calibration. Subjects should be pair-matched based on this variable.

4. The calibration exercise should be focused on criteria that evaluators may find easy to misinterpret or ambiguous in order to improve answer key agreement and reliability.

5. A third assessment point should be added such as "excellent," to the two assessment point, competency-based evaluation system of "correct," or "incorrect," to reduce the variable of guessing and add an assessment of quality to the product or process being evaluated.

6. Survey research should be conducted to determine evaluator calibration methodology utilized by dental hygiene programs nationwide.

7. Various methods of calibrating examiners should be investigated with the use of a non-calibrated control group to determine the effectiveness of the various methods.

8. Knowledge attained as a result of the calibration exercise should be measured to assess the effectiveness of the calibration exercise.

Within the limits of the study, findings suggest that calibration has no effect on standardization of evaluators
when measured by answer key agreement, no effect on standardization when measured by intra-rater reliability, and a questionable effect when measuring inter-rater reliability. Results suggest that method of calibration utilized may affect standardization of evaluators for assessment of pre-clinical instrumentation skills. Other methods of improving faculty standardization must be explored to ensure that dental hygiene students will benefit from their education through fair, objective, and unbiased skill performance evaluation.
BIBLIOGRAPHY


Mackenzie, R.S. "Defining Clinical Competence in Terms of Quality, Quantity, and Need for Performance Criteria." Journal of Dental Education 37, no. 9(1973): 37-44.


University of Texas at San Antonio, Health Science Center, Office of Educational Resources, Dental Hygiene Education. *An Edge on Success: Techniques for the Successful Use of Universal Instruments*, (1986).

APPENDIX A

Letter of Introduction
April 21, 1991

Ms. , RDH, BS
9522 Sturgis Street
Norfolk, VA 23503

Dear Ms. ,

I am a master's degree candidate in the School of Dental Hygiene at Old Dominion University conducting my thesis research in the area of dental hygiene educational methods. Due to your past experience as a clinical instructor, I am respectfully asking you to participate in my research along with twenty other dental hygiene clinical instructors. I urge you to accept this invitation due to the limited number of professionals who qualify as subjects.

The project will require that you review the enclosed materials and be present for a mock clinical evaluation exercise. The study will require approximately two hours of your time in Room 245 of the Technology Building in addition to the time it will take you to review these materials. It is important to remember that you will not be evaluated, only the forms which you will be filling out. The forms will be identified with an alphabet code rather than your name to assure confidentiality. Please memorize your code which is contained on the informed consent abstract.

A tentative scheduling of the data collection day is early May. I will be calling you within seven days of mailing out this packet to answer any questions you might have regarding this study and your participation. Please complete and return the attached time availability schedule by crossing out times which are inconvenient for you as well as the form stapled onto it requesting your years of clinical teaching experience. Hours after 3 PM are preferable. Please return the time availability schedule along with your signed informed consent materials at your earliest convenience; preferably within one week. Bring this entire package with you on the decided upon day. You will be contacted when an agreeable date has been determined. I am looking forward to the opportunity to work with you. Thank you for taking the time to review these materials.

Sincerely,

Diane-Marie Smela, RDH, BS
Master's Degree Candidate
School of Dental Hygiene
Old Dominion University
Please complete and return this form with the time availability schedule and your informed consent materials.

Name ________________________________

Years of clinical teaching experience (please check the category which best describes you).

_____ less than one year

_____ one year

_____ 2-4 years

_____ 5-7 years

_____ 8-10 years

_____ 11 or more years
APPENDIX B

Informed Consent Materials
Informed Consent Abstract

Because of your past experience as a clinical instructor/evaluator in a dental hygiene educational program, you are being asked to participate in this research project.

The principal investigator does not foresee any risks to the physical, psychological or social welfare of any subject participating in this study.

The principal investigator assures that all precautions will be taken to prevent the publicizing of individual measurements. The measurement instrument completed by you, the subject, will be identified only by coded data. The principal investigator will remain blind.

The study will assess the use of dental hygiene instrumentation skill evaluations by clinical instructors. You have been assigned to GROUP ___. Your code is ____. Please use the code on all skill evaluation forms which you will complete on the day of the actual research. Because this investigation relies on the Nield instrumentation text, it will be necessary for you to review some important information. The skill which will be assessed is instrumentation utilizing a Columbia 13/14 curet. Please take the time to review the following enclosed materials until you are reasonably familiar with them:


This study will add to the dental hygiene body of knowledge by providing educators with a method of calibrating faculty evaluators for standardization and rater reliability.

The research may be subject to an Institutional Board Review for which you may be asked to appear.

By signing the attached consent form, you are agreeing to participate in this study.
INFORMED CONSENT FORM

Project Name: "Faculty Standardization of Dental Hygiene Instrumentation Skill Evaluation: A Calibration Exercise."
Investigator: Diane-Marie Smela, RDH, BS
Date: 

This is to certify that I, ______________________ hereby agree to participate as a volunteer in a scientific investigation as a part of the educational and research program of Old Dominion University, under the supervision of Diane-Marie Smela.

The investigation and the nature of my participation has been described and explained to me, and I understand the explanation (see attached abstract). I understand that I am one of the twenty dental hygiene educators chosen because of my past experience in clinical teaching. I further understand that I may withdraw from the project at any time, without penalty or prejudice.

I have been afforded an opportunity to ask questions concerning the purpose of this project and all such questions have been answered to my satisfaction. I understand that should I have additional questions in the future about this project or the manner in which it is conducted, I may contact Diane-Marie Smela, the principal investigator, at (804) 683-5234 or Deborah Bauman, thesis director, at (804) 683-5150.

I understand that any data or answers to questions will remain confidential with regard to my identity. I further understand that no data which can be identified with me will be released to persons outside the research team which consists of Ms. Smela, Ms. Bauman and their research assistant, Kristin Hamman, without the team first obtaining my written permission.

I acknowledge that I was informed about any possible risks and benefits to my health and well being that may be associated with my participation in this research (see attached abstract).

I understand that no medical or psychological assistance will be made available to me by either Old Dominion University or any member of the research team as a result of any physical or emotional harm I may experience as a result of this research project.

I acknowledge that I have been advised of how I may obtain a copy of the results of this research project and that upon my making such a request, a copy will be provided without charge.

I have been informed that I have the right to contact the Old Dominion University Institutional Review Board for Protection of Human Subjects should I wish to express any opinions regarding the conduct of this study.
APPENDIX C

Instrumentation Skill Evaluation Form
SKILL EVALUATION: UNIVERSAL INSTRUMENTS, MANDIBULAR SEXTANTS

Student: _______________________

Instructor: _______________________

Date _______________________

Area 1 • anterior, facial aspect: instrument(s) evaluated _______________________

Area 2 • anterior, lingual aspect: instrument(s) evaluated _______________________

Area 3 • right posterior, facial: instrument(s) evaluated _______________________

Area 4 • right posterior, lingual: instrument(s) evaluated _______________________

Area 5 • left posterior, facial: instrument(s) evaluated _______________________

Area 6 • left posterior, lingual: instrument(s) evaluated _______________________

DIRECTIONS: For each sextant evaluated, indicate:

C for correct skill performance or
I for improvement needed

---

<table>
<thead>
<tr>
<th>CRITERIA</th>
<th>AREA 1</th>
<th>AREA 2</th>
<th>AREA 3</th>
<th>AREA 4</th>
<th>AREA 5</th>
<th>AREA 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Uses all criteria for patient and operator positions</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Uses all criteria for grasp</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Uses all criteria for mirror and finger rests</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Uses correct working end</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. Inserts at zero degrees, if appropriate</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. Handle and shank parallel to tooth's long axis</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7. Uses tip-third (toe third) of cutting edge</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8. Establishes correct tooth-to-cutting edge angulation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9. Rolls handle to maintain adaptation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10. Uses controlled working strokes</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11. Uses appropriate stroke length</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12. Applies stroke pressure in a coronal direction only</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13. Uses overlapping strokes</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14. Uses wrist/arm activation to produce controlled strokes</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
APPENDIX D

Calibration Session Lesson Plan
Calibration Session
Lesson Plan

Audience: Clinical Dental Hygiene Educators.


Subject: Seminar for Standardized Use of Instrumentation Skill Evaluation.

Bibliography:


Supporting Concepts: Participants will have experience in clinical evaluation. Participants should have read and familiarized themselves with pages 287-309 of the Nield text.

Materials: 6 typodonts
6 Columbia 13/14 curets
6 copies of Nield text
"An Edge of Success: Universal Instrumentation."
instructional videotape

Time: 60 minutes

Objectives: Upon completion of this calibration seminar, the clinical dental hygiene educator will be able to:


2. accurately define the 14 skill performance criteria.

3. appreciate the benefits of calibration to standardize clinical evaluation.

4. evaluate the student in a simulated evaluation setting on video vignettes with 90% rater reliability.

Mood: informal, yet serious
Instructional Set

Introduction: Good afternoon. I would like to begin by thanking you for agreeing to participate in my thesis research. Before we begin, I would like to collect your informed consent. If you haven’t already signed it, please do so now. This portion of the study should take approximately one hour.

Rationale: It is believed by researchers that calibration prior to evaluation will decrease the amount of subjectivity present in clinical evaluation. Valid, reliable and unbiased clinical evaluation of student ability will result in the graduation of competent clinicians from dental hygiene programs. The study which you are about to participate in will add to the body of knowledge relating to this point.

Instructional Objective: At the conclusion of the hour, you will be able to use the skill evaluation for the universal curet in the mandibular arch with a high degree of intra- and inter-rater reliability.
Content Outline: Procedures/Activities

I. Use of Skill Evaluation
   trans: the skill evaluation
   A. Completing the ID info.
      1. Student:
         a. write "video 1,2,3 or 4"
      2. Instructor:
         write your code
      3. Date:
         today’s date
      4. Area#
         write "Columbia 13/14" next to appropriate area

   B. Criteria Area
      1. match area # being evaluated
         grade only in that area
      2. mark in "I" (instructor) area
      3. C=correct
         I=improvement needed

   C. The Back Side
      trans: the back side
      1. student= video #
      2. instructor comments
         a. feel free to write comments
            to aid student in improving skill
         b. draw sketches as desired
         c. clarify "I" ‘s

III. Skill Performance Criteria Breakdown
FEEL FREE TO ASK QUESTIONS AS THEY COME UP.

1. USES ALL CRITERIA FOR PATIENT-OPERATOR-POSITIONING
   **DO NOT EVALUATE
   provide demo

2. USES ALL CRITERIA FOR GRASP
   pp. 68-76
   modified pen
   encourage subjects
   see photo page 70
   to demo

3. USES ALL CRITERIA FOR MIRROR AND FINGER RESTS
   **DO NOT EVALUATE

4. USES CORRECT WORKING END
   p. 292
   face should be curved toward tooth surface
   partially visible in mesial direction
   is face is very visible= incorrect
   shouldn’t switch working ends to scale distal aspect
5. INSERTS AT ZERO DEGREES IF APPROPRIATE  
   p. 214  
   toe 1/3 in contact with tooth  
   face flat against surface zero degrees between face and tooth  

6. HANDLE AND SHANK PARALLEL TO TOOTH'S LONG AXIS  

7. USES TOE 1/3 OF CUTTING EDGE  
   p. 213  
   roll handle to maintain cutting edge to all areas  

8. ESTABLISHES CORRECT TOOTH-TO-CUTTING EDGE ANGULATION  
   p. 216  
   more than 45° less than 90°  
   ideal 70-80°  

9. ROLLS HANDLE TO MAINTAIN ADAPTATION  
   p. 173  
   adaptation= manner in which instrument is brought into contact with tooth surface and maintained during instrumentation  
   roll between index finger and thumb  

10. USES CONTROLLED WORKING STROKES  
    instrument shouldn’t fly off of tooth  

11. USES APPROPRIATE STROKE LENGTH  
    p. 210  
    to maintain control, short stroke length is necessary 2-3mm.  

12. APPLIES STROKE PRESSURE IN A CORONAL DIRECTION  
    never apical  
    reposition with pressureless grasp  
    pull up firmly  
    very different from the way you learned  

13. USES OVERLAPPING STROKES  
    do not miss any area  

14. USES WRIST/ARM ACTIVATION TO PRODUCE CONTROLLED STROKES  
    p. 167  
    fingers should remain motionless  
    digital motion not acceptable  
    rocking/ rotating
IV. Demonstration, Discussion and Questions

get feedback

demonstrate non-mastery behaviors

answer questions

CLOSURE

Summary: That concludes the calibration portion of today's activities. Let's review some identification criteria.

Any questions?

Appreciation: Thank you for your attention and participation. I would encourage you to take a ten minute break and then report back to this room for the next portion of the study.

1. While Group A is on break, ask Group B to enter room
   a. collect informed consent
   b. get settled
   c. random seat assignments

2. Instructions to Both Groups:
   a. Please do not discuss anything among yourselves
   b. Try to make comments in space provided
   c. Place your code on the "Instructor" line on all four immediately
   d. Place the vignette number on "Student" line
   e. Write: "Columbia 13/14" in"Area" line
   f. do not evaluate criteria #1 or #3

3. Commence video viewing

4. Collect skill evaluations after each vignette

5. Thank you. You may leave.
APPENDIX E

Answer Keys and Descriptions of Vignettes
**VIGNETTE I**

**SKILL EVALUATION OF UNIVERSAL INSTRUMENTATION, MANDIBULAR SEXTANT, RIGHT POSTERIOR FACIAL**

**SKILL EVALUATION: UNIVERSAL INSTRUMENTS, MANDIBULAR SEXTANTS**

<table>
<thead>
<tr>
<th>Student VIGNETTE I</th>
<th>Instructor</th>
<th>Date</th>
</tr>
</thead>
</table>

**DIRECTIONS:** For each sextent evaluated, indicate:

- C for correct skill performance or 1 for improvement needed

<table>
<thead>
<tr>
<th>CRITERIA</th>
<th>S = self-evaluation</th>
<th>R = re-evaluation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Uses all creases for patient and operator positions</td>
<td>NA</td>
</tr>
<tr>
<td>2</td>
<td>Uses all creases for grasp</td>
<td>C</td>
</tr>
<tr>
<td>3</td>
<td>Uses all creases for mirror and lighter rests</td>
<td>NA</td>
</tr>
<tr>
<td>4</td>
<td>Uses correct working end</td>
<td>I</td>
</tr>
<tr>
<td>5</td>
<td>Inserts at zero degrees, if appropriate</td>
<td>I</td>
</tr>
<tr>
<td>6</td>
<td>Handle and shank parallel to tooth's long axis</td>
<td>I</td>
</tr>
<tr>
<td>7</td>
<td>Uses two-thirds third of cutting edge</td>
<td>C</td>
</tr>
<tr>
<td>8</td>
<td>Establishes correct tooth-to-cutting edge angulation</td>
<td>I</td>
</tr>
<tr>
<td>9</td>
<td>Rolls handle to maintain adaptation</td>
<td>C</td>
</tr>
<tr>
<td>10</td>
<td>Uses controlled working strokes</td>
<td>I</td>
</tr>
<tr>
<td>11</td>
<td>Uses appropriate stroke length</td>
<td>I</td>
</tr>
<tr>
<td>12</td>
<td>Applies stroke pressure in a coronal direction only</td>
<td>I</td>
</tr>
<tr>
<td>13</td>
<td>Uses overlapping strokes</td>
<td>I</td>
</tr>
<tr>
<td>14</td>
<td>Uses wrist arm activation to produce controlled strokes</td>
<td>C</td>
</tr>
</tbody>
</table>

Student operator will improperly instrument the right posterior facial resulting in the following errors:

- Criterion #4: uses incorrect working end
- Criterion #5: instrument not inserted at zero degrees
- Criterion #6: handle and shank not parallel
- Criterion #8: tooth to cutting edge angulation incorrect
- Criterion #10: working strokes not controlled
- Criterion #11: stroke length incorrect
- Criterion #12: stroke pressure in apical direction
- Criterion #13: strokes not overlapping
VIGNETTE II

SKILL EVALUATION OF UNIVERSAL INSTRUMENTATION, MANDIBULAR SEXTANT, RIGHT POSTERIOR FACIAL

SKILL EVALUATION: UNIVERSAL INSTRUMENTS, MANDIBULAR SEXTANTS

<table>
<thead>
<tr>
<th>Student: VIGNETTE II</th>
<th>Instructor:</th>
<th>Date:</th>
</tr>
</thead>
</table>

AREA 1 = anterior, facial aspect: instrument(s) evaluated
AREA 2 = anterior, lingual aspect: instrument(s) evaluated
AREA 3 = right posterior, facial: instrument(s) evaluated
AREA 4 = right posterior, lingual: instrument(s) evaluated
AREA 5 = left posterior, facial: instrument(s) evaluated
AREA 6 = left posterior, lingual: instrument(s) evaluated

DIRECTIONS: For each sextant evaluated, indicate:
C for correct skill performance or I for improvement needed

<table>
<thead>
<tr>
<th>CRITERIA:</th>
<th>S = self-evaluation</th>
<th>I = instructor-evaluation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Uses all crowns for correct and operator positions</td>
<td>AREA 1</td>
<td>AREA 2</td>
</tr>
<tr>
<td>2. Uses all crowns for grasp</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Uses all crowns for molar and lingual edges</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Uses correct working end</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. Inserts at 90 degrees, if appropriate</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. Handle and shank parallel to tooth's long axis</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7. Uses tip-thirteenth third of cutting edge</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8. Establishes correct tooth-to-cutting edge angulation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9. Rolls handle to maintain angulation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10. Uses controlled working strokes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11. Uses approximate stroke length</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12. Applies stroke pressure in a coronal direction only</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13. Uses overlapping strokes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>14. Uses wrist, arm activation to produce controlled strokes</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Student operator will improperly instrument the right posterior facial aspect resulting in the following errors:

Criterion #2: student uses incorrect grasp
Criterion #6: handle and shank not parallel to long axis of tooth
Criterion #8: incorrect tooth-to-cutting edge angulation
Criterion #10: working strokes not controlled
Criterion #11: stroke length incorrect
Criterion #12: stroke pressure in coronal direction
Criterion #13: strokes are not overlapping
VIGNETTE III

SKILL EVALUATION OF UNIVERSAL INSTRUMENTATION, MANDIBULAR SEXTANT, RIGHT POSTERIOR FACIALS

SKILL EVALUATION: UNIVERSAL INSTRUMENTS, MANDIBULAR SEXTANTS

Student: ________________  INSTRUCTOR: ___________________________

Date: ________________

DIRECTIONS: For each sextant evaluated, indicate:

C for correct skill performance or
R for improvement needed

<table>
<thead>
<tr>
<th>CRITERIA</th>
</tr>
</thead>
<tbody>
<tr>
<td>S = non-evaluation</td>
</tr>
<tr>
<td>1 = re-evaluation</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>AREA 1</th>
<th>AREA 2</th>
<th>AREA 3</th>
<th>AREA 4</th>
<th>AREA 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>S I R S I R S I R S I R S I R</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

1. Uses all crevices for patient and operator positions
2. Uses all crevices for grasp
3. Uses all crevices for molar and lingual roots
4. Uses correct working end
5. Inserts at zero degrees, if appropriate
6. Handles and shaft parallel to tooth’s long axis
7. Uses top-third toe thrust of cutting edge
8. Establishes correct tooth-to-cutting edge angulation
9. Rotates handle to maintain occlusion
10. Uses controlled working strokes
11. Uses appropriate stroke length
12. Applies stroke pressure in a coronal direction only
13. Uses overlapping strokes
14. Uses wrist/arm activation to produce controlled strokes

Student operator will improperly instrument the right posterior facial aspect resulting in the following errors:

no errors
Student operator will improperly instrument the right posterior facial aspect resulting in the following errors:

Criterion #2: student uses incorrect grasp
Criterion #6: handle and shank not parallel to long axis of tooth
Criterion #8: incorrect tooth-to-cutting edge angulation
Criterion #10: working strokes not controlled
Criterion #11: stroke length incorrect
Criterion #12: stroke pressure in coronal direction
Criterion #13: strokes are not overlapping
APPENDIX F

Raw Data Organizational Charts
GROUP A
INDIVIDUAL SUBJECT RESPONSE
CORRECT IN COMPARISON TO ANSWER KEY

<table>
<thead>
<tr>
<th>CRITERIA NUMBER</th>
<th>L</th>
<th>Z</th>
<th>K</th>
<th>J</th>
<th>P</th>
<th>R</th>
<th>O</th>
<th>ANSWER</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
GROUP B
INDIVIDUAL SUBJECT RESPONSE
CORRECT IN COMPARISON TO ANSWER KEY

<table>
<thead>
<tr>
<th>CRITERIA</th>
<th>SUBJECT CODE</th>
<th>ANSWER</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>X A G E M F C</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### GROUP A INTRA-RATER COMPARISON

**VIGNETTES II AND IV**

<table>
<thead>
<tr>
<th>SUBJECT CODE</th>
<th>L</th>
<th>Z</th>
<th>K</th>
<th>J</th>
<th>P</th>
<th>R</th>
<th>O</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>II</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>IV</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>II</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>IV</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>II</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>IV</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>II</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>IV</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>II</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>IV</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>II</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>IV</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>II</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>IV</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>II</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>IV</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>II</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>IV</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>II</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>IV</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>II</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>IV</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>II</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>IV</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### GROUP B INTRA-RATER COMPARISON

**VIGNETTES II AND IV**

<table>
<thead>
<tr>
<th>CRITERIA AND VIGNETTE NUMBERS</th>
<th>SUBJECT CODE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>X</td>
</tr>
<tr>
<td>2 II</td>
<td></td>
</tr>
<tr>
<td>2 IV</td>
<td></td>
</tr>
<tr>
<td>4 II</td>
<td></td>
</tr>
<tr>
<td>4 IV</td>
<td></td>
</tr>
<tr>
<td>5 II</td>
<td></td>
</tr>
<tr>
<td>5 IV</td>
<td></td>
</tr>
<tr>
<td>6 II</td>
<td></td>
</tr>
<tr>
<td>6 IV</td>
<td></td>
</tr>
<tr>
<td>7 II</td>
<td></td>
</tr>
<tr>
<td>7 IV</td>
<td></td>
</tr>
<tr>
<td>8 II</td>
<td></td>
</tr>
<tr>
<td>8 IV</td>
<td></td>
</tr>
<tr>
<td>9 II</td>
<td></td>
</tr>
<tr>
<td>9 IV</td>
<td></td>
</tr>
<tr>
<td>10 II</td>
<td></td>
</tr>
<tr>
<td>10 IV</td>
<td></td>
</tr>
<tr>
<td>11 II</td>
<td></td>
</tr>
<tr>
<td>11 IV</td>
<td></td>
</tr>
<tr>
<td>12 II</td>
<td></td>
</tr>
<tr>
<td>12 IV</td>
<td></td>
</tr>
<tr>
<td>13 II</td>
<td></td>
</tr>
<tr>
<td>13 IV</td>
<td></td>
</tr>
<tr>
<td>14 II</td>
<td></td>
</tr>
<tr>
<td>14 IV</td>
<td></td>
</tr>
</tbody>
</table>
GROUP A INTER-RATER COMPARISON
VIGNETTE II

<table>
<thead>
<tr>
<th>CRITERIA NUMBER</th>
<th>L</th>
<th>Z</th>
<th>K</th>
<th>J</th>
<th>P</th>
<th>R</th>
<th>O</th>
<th>RELIABLE</th>
<th>PERCENT</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## GROUP B INTER-RATER COMPARISON
### VIGNETTE II

<table>
<thead>
<tr>
<th>CRITERIA NUMBER</th>
<th>SUBJECT CODE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>X</td>
</tr>
<tr>
<td>2</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td></td>
</tr>
</tbody>
</table>

| #TIMES GROUP RELIABLE PERCENT |
|------------------------------|-----------------|
| X   | A   | G   | E   | M   | F   | C   |