The Effect of Incorporating Active Learning In Calibration Exercises On Intra and Interrater Reliability Among Dental Hygiene Faculty

A THESIS SUBMITTED TO THE FACULTY
OF THE DIVISION OF DENTAL HYGIENE SCHOOL OF DENTISTRY
UNIVERSITY OF MINNESOTA

BY:

Bridget Hotzler

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTERS OF SCIENCE IN DENTAL HYGIENE

Dr. Christine Blue

September 2019

ACKNOWLEDGEMENTS

I would like to thank and acknowledge Dr. Christine Blue for the assistance, guidance and uplifting support in completing my thesis.

DEDICATION

This thesis is dedicated to my husband Bryce Hotzler for all of his tremendous support and sacrifices he made to make this dream of mine happen. Thank you for being my rock and believing in everything I do.

TABLE OF CONTENTS

List of Tables

List of Figures

SECTION 1

INTRODUCTION

Developing fair and uniform grading practices is a constant challenge in health professional education, as faculty bring varying educational backgrounds and practice experiences. Given the ultimate goal of professional education is to prepare competent graduates ready to provide quality care to the public, the concern of grading inconsistencies and lack of standardization among faculty is well-founded (1,2). Lack of calibration often diverts students from directing their focus on patient care and redirects to satisfying the evaluating instructor (1,3). Efforts to increase intra- and interrater reliability have been shown to increase student learning, lessen frustration, and advance the quality of health care (1,3-5). Therefore, educators strive to attain consistency and reduce variability via uniformly applied assessment criteria. (1-4,6-7).

Inconsistent application of examination criteria can be reduced through calibration and standardization exercises (1-5,7). Scruggs et al. found methods such as faculty workshops, instrumentation sessions, task analysis, video recordings and group discussions are key to the development of standards for assessment and calibration of examiners (4). Although faculty calibration holds great significance and value in dental education, its success has been hard to achieve (2-4,7). Over time, faculty tend to drift from uniform application of grading criteria demanding calibration exercises be done frequently. Courts focused on these challenges and found that a process must be in order to effectively communicate grading standards to examiners, as standards need to be followed in order to achieve an acceptable level of calibration (5).

Theoretical Framework

Constructivist theory provides the theoretical framework for this study. Constructivist learning theory emphasizes learners structure their own understanding. Given new learning experiences, learners replace or modify existing viewpoints or prior knowledge with deeper and more skilled levels of understanding (8-11). During the past decade, constructivist theory has been embraced in the form of active learning. Teaching methods supporting active learning ask students to make associations between new knowledge and their existing mental representations (8-10).

Incorporating active learning into the classroom has shown to improve learning, retention and promote a deeper development in thinking and writing skills (8-11). Active learning requires the use of higher-order thinking, and studies have shown students become more engaged in course content (8,10-11,). Utilization of active learning has been strongly advocated if a course aims to encourage long-term retention of material, stimulate students toward further knowledge, allow students to apply information in new situations, or to advance students' intellectual skills (8-10). If active learning is associated with learning at a higher cognitive level and increased retention, the application of active learning to faculty calibration warrants exploration (9). Immersing faculty in the task assigned to students may foster a deeper understanding of the knowledge and skills required which may, in turn, clarify grading standards and criteria resulting in higher levels of intra and inter-examiner agreement (8-9).

Purpose of the Study

The purpose of this study was to investigate if incorporating active learning into calibration exercises increases the level of intra and interrater reliability among dental hygiene faculty.

Statement of the Problem

For optimal learning faculty assessment and feedback relayed to students must be consistent and reliable (1-3,6,7). Standardizing assessment of students' clinical skills and knowledge within dental education helps minimize grading inconsistencies, reduces faculty bias, and improves student learning (1,3,5). However, research has shown that high levels of calibration are hard to achieve and sustain (1,3-4,7). Additionally, empirical evidence on the effectiveness of specific calibration methods is limited. The active learning theory suggests individuals learn through building their own knowledge, and connecting existing ideas and experiences to form new or enhanced understanding (8-11). Given active learning methods are associated with higher levels of learning and increased retention, engaging faculty in the performance of the same skill/task in identical testing has the potential to increase the consistency of grading among faculty (8-11).

Significance of the Study

Providing alternative ways to conduct and improve faculty calibration holds great significance within dental educational programs. Consistent evaluation and grading practices may reduce student frustration and instill confidence in faculty skills and knowledge (1-2,5). Studies have shown that inadequate levels of calibration may lead to unsatisfactory student education, lessen students'

gratification regarding their education and ultimately influence patients' experience and care (1-2,4-5,7). Calibration studies in dental hygiene education are underrepresented and are needed to understand how to appropriately and effectively reduce grading inconsistencies and increase reliability when assessing student performance (1-5,12). The findings of this study will add to the body of knowledge and provide evidence to strengthen calibration methods for faculty responsible for educating future oral health professionals.

Research Question

What is the effect of incorporating active learning in calibration exercises on intra- and interrater reliability among dental hygiene faculty?

Hypothesis

There is no difference in intra and inter-examiner reliability scores of dental hygiene faculty participating in active learning calibration exercises verses traditional methods of calibration.

SECTION 2

REVIEW OF THE RELATED LITERATURE

Calibration refers to training exercises in which standardization among evaluators is accomplished (1-2,4-5,12). Faculty calibrate to achieve consistent evaluation of skills and procedures performed by students (2,5,13). Higher levels of uniformity achieved via calibration results in fairer and reliable grading (2,5,13). The goal of calibration is to achieve high levels of examiner agreement through the use of criterion-based standards and to be able to reproduce those standards in different situations (2). In other words, to achieve consistent and reliable assessment, instructors must understand chosen criteria, apply the criteria in an identical way each time a student's skill is evaluated and make comparable qualitative conclusions based on those standards (1).

For clinical assessment to accurately express student performance, it must be valid and reliable (1-2, 4-5). Validity is the most significant factor for assessment and is described as the degree to which evaluation measures the concept for which knowledge is attained (4-6). Validity encompasses both the assessment criteria (do assessment measurements that are used correctly assess the knowledge/skill being measured?) and the evaluators (do the evaluators apply the measurements exactly as specified?) (1,4-6). Reliability is also essential for accurate evaluation. When multiple examiners are involved, two types of reliability must be measured: intra and interrater reliability. Intra-rater reliability is defined as the agreement of the evaluators with themselves or degree of agreement throughout repetitive administrations of an assessment completed by the same rater (4,13-14). Interrater reliability is the ability of evaluators to consistently apply measurement criteria

when provided with the same circumstances for evaluation (4,14).

Grading inconsistencies among faculty during clinical assessment is often a direct result of lack in calibration. In order to appraise the current literature on faculty calibration within dental educational programs, a search for articles from 1985 to 2017 in the US was performed. The search included the use of two electronic databases, PubMed and Google Scholar. Search terms included calibration, standardization, interrater reliability, active learning, constructivism, dental education, clinical application and motivational interviewing. Journal articles found using these search terms were further filtered by reviewing the abstracts and included if they held significance to the study.

Relationship Between Calibration and Student Learning

Although calibration is most often associated with faculty, the impact and effects of calibration directly influence students' attitudes and values (1-4,6,15). Scoring inconsistencies resulting from poor calibration have shown to affect not only grades but overall learning development (1-2,4). A study by Scruggs et al. indicated that North American dental students identified inconsistent clinical feedback as one of the major obstacles in achieving clinical competence (4). Another study by Jacks et al. revealed that lack of calibration among dental faculty was "a significant source of trouble, worry, and discomfort, a major source of anger, and one of the primary reasons for abandonment of a quest for excellence and resignation to just getting by" (12). Recurrent discrepancies in assessment have been shown to minimize students' passion to learn, reduce student satisfaction with the learning environment, and ultimately affect patient care (1-2,4). Studies have

concluded that inconsistencies and variations among faculty during assessment can source the feelings of frustration within students and create an undesirable and negative learning environment (1-2,4,12,15). Scruggs et al. reported students feeling directly affected when criteria for evaluation is unclear and when faculty are inconsistent with their clinical feedback (4). Studies evaluating student perspectives on clinical feedback found students felt feedback was more credible when evaluators had been trained with exercises that fostered subject mastery (2,4,12,13).

Several studies assert the reliability of evaluators is compromised when instructor feedback varies with similar student performance (1-2,4,7,12,15). Students begin to view faculty as incompetent when continuous variations in grading and evaluation practices exist. (1,4). Recurrent situations involving inadequate calibration can alter the integrity of clinical instruction, discourage student learning and decrease motivation to improve skills (1-2,4,12). Jacks et al. provided evidence that students rely on evaluators' feedback and utilize this information "to make appropriate alterations in their next attempt in order to achieve a higher level of performance" (12). When valid evaluations are absent, confusion can occur and students are given a misconception of what is expected of them (1-2,4,12).

Calibration Challenges

Striving to reach an acceptable level of calibration is difficult, but even more challenging to maintain. Dental and dental hygiene faculty are often faced with the challenges of inconsistencies within evaluations due to variations in clinical judgement. Faculty members attend different schools at different times and

have been exposed to diverse clinical experiences (13). Additional obstacles to faculty calibration include, years of experience among faculty, part-time verses full-time faculty, if calibration training is mandatory or optional, and reimbursement for participation if calibration exercises are held after hours (1-5,7,12,13). A significant body of literature discussed the challenges associated with implementing calibration exercises, including the need for human subjects, coordinating large numbers of participants, and the question of how long intra and interrater reliability can be sustained (1,2,5-7,13).

Also, the effectiveness of a particular calibration method is often times unknown. A study by Partido et al. investigated calibrating calculus detection utilizing typodonts (2). The study reported difficulties with the calibration exercises due to the unrealistic nature of simulated calculus as compared to authentic calculus (2). Another study by Courts et al. asserted a well-defined, post-examination analysis should follow any calibration in order to validate that acceptable standardization was attained (5). The literature suggests that a well-developed calibration protocol consists of: 1) established criteria for evaluating scholarly or clinical performance, 2) subsequent assessment of the evaluators applying the clinical evaluation protocol and, 3) assessing the consequences of the calibration protocol in regards to student competence (2).

Research to date is inconclusive on length of time needed for faculty recalibration, as it may depend on the particular skill set needing to be calibrated (13). Studies have described preservation of enhanced interrater reliability over pre-training assessments ranging from ten weeks to one year (7,12,16). Haj-Ali et al

conducted a study to investigate the immediate effects of calibration when held to a gold standard in regards to Class II amalgam preparation and found that calibration could be sustained for a ten-week period (7). Jacks et al. discussed the capacity of evaluators to evaluate dental hygiene students' SOAP notes and assessed the sustainability long-term. This study concluded that faculty-maintained calibration (as measured by the gold standard) for one year following the calibration workshop (12). The paucity of research on frequency and sustainability of calibration is evidence that more research needs to be conducted.

Methods of Calibration

Evident in the literature is a gap in knowledge pertaining to which calibration methods are most effective in achieving high intra- and interrater reliability (1,3-5). Among the most common calibration methods is the application of communication and verbalization in groups (4-5). Among the approaches that have been investigated are workshops, instrumentation sessions, task analysis, communication or discussions in groups and videotapes (4). Gathering the evaluation team together for discussion-based calibration exercises has shown to improve the level of uniform feedback (4,7,13). Calibration workshops hold the highest success when performed outside of the educational environment and staged within a positive and nonthreatening atmosphere to lessen any potential influential factors (4,7). Workshops provide a mechanism of joining raters together as an assessment team and training them together. As new faculty join the team, training in the company of at least one experienced faculty member improves rater reliability (3-4). Haj-Ali et al. investigated immediate effects of calibration on

interrater agreement when evaluating Class II amalgam preparations (7). Raters gathered together, tested initial interrater agreement prior to any calibration to obtain a baseline. Next, raters participated in a workshop scenario consisting of calibration exercises involving group discussion and the development of an acceptable gold standard (GS). Immediately following calibration, the raters were given ten prepped teeth to individually evaluate. The study concluded that interrater agreement improved as a result of the calibration exercises and because the criteria evaluated had a GS (7).

Calibration sessions should involve a diversity of discrimination exercises, use checklists or rubrics, and engage evaluators in hands-on activities (13). Research verifies that students feel evaluators are more reliable when they can clearly understand and can articulate expectations and discrepancies in a uniform way (1-2,4,7). Related disciplines have revealed that the quality of the performance outcome is significantly improved when students can correctly self-evaluate their product and comprehend their progress (7). Students pursue "knowledge of results" from their faculty, and therefore, it is imperative that any feedback provided is consistently accurate (7).

Goolsbee et al. found using an audience response system (clickers) improved calibration among faculty with regard to caries risk assessment (17). Likewise, Metz et al. investigated the use of clickers at quarterly calibration sessions over a 12-month period (18). The results showed this practice united faculty members and allowed for instantaneous feedback. The instant feedback helped individual faculty members to assess their performance in relation to fellow

faculty and improved interrater reliability and resulted in more positive student opinions of faculty uniformity (17-18). Lastly, the use of videotapes or recording devices can aid evaluators in visualizing and identifying specific standards within the presentation. An advantage of calibrating with video recordings is the ability to watch and evaluate students' performance multiple times and discuss grading discrepancies until a consensus has been reached (4). Utilizing videos for calibrating also provides benefits for uncovering inconsistencies among faculty, beyond psychomotor deficiencies, such as body language, miscommunication and other nonverbal interferences (4).

Calibration Gaps Within Dental Education

Numerous studies confirmed substantial discrepancy in assessment and clinical decision making among health care education faculty (1,4,7,16,19-21). This finding was true for dental faculty with regard to clinical decision making (1,3-7). Many studies have found discrepancies among dental faculty during the assessment of periodontitis diagnosis, treatment planning, calculus detection, cavity preparation assessment, radiographic interpretation, periodontal probing (1,16,19-21). Consequently, efforts have been devoted to identifying effective strategies to improve the level of interrater agreement. Two studies conducted in the area of calculus detection were found in the dental hygiene literature. Garland and Newell investigated the effectiveness of utilizing a training program to improve intra and interrater reliability when grading deposit detection (1). Dental hygiene faculty were asked to detect simulated calculus on typodonts using an 11/12 explorer (1). Although the results were not statistically significant with regard to increasing

interrater reliability, the study still stressed other positives that came from the calibration exercise (1). Benefits to faculty included becoming more conscious of their own exploring skills and agreed with the students' assertion that there was lack of uniformity among faculty (1).

A randomized experimental pilot study by Partido et al. sought to determine if incorporating periodontal endoscopy into calibration exercises would increase both inter and intra-rater scores during calculus detection (2). Evaluators used an 11/12 explorer to detect simulated calculus on three typodonts. The treatment group had an additional two-hour calibration session using both the 11/12 explorer and dental endoscopy for calculus detection. A significant difference was found between pretest and posttest mean kappa averages for the treatment group vs. the control group. Additionally, the investigators found calibration training increased rater agreement among new and veteran faculty (2).

Active Learning Applied to Calibration

Active learning suggests that students need to be immersed in activities to learn as opposed to passively sitting and listening in a lecture (8-11,22-23). Active learning places emphasis on creating learning activities that reinforce course content and develop students' skills (9). Active learning rather than passive absorption has been shown to accelerate learning (8-10). Given active learning activities have shown to increase understanding and lead to deeper learning, active learning may hold significance in the context of faculty calibration (8-10). Engaging faculty in performing a procedure required of students may increase their knowledge and deepen their understanding of the procedure they will evaluate, which may in turn,

improve intra and interrater reliability. Therefore, the purpose of this study was to investigate if the integration of active learning into calibration exercises increased the level of intra and interrater reliability among dental hygiene faculty.

SECTION 3

SUMMARY

Purpose/Objective: The purpose of this study was to investigate if incorporating active learning into calibration exercises increased the level of intra- and interrater reliability among dental hygiene faculty.

Methods: The study used a two-group randomized experimental design with a convenience sample consisting of ten dental hygiene faculty members from the division of dental hygiene at the University of Minnesota (n=10). Baseline training in motivational interviewing (MI) was provided to all faculty at a day-long continuing education course. One month later, all faculty viewed three videos of students performing MI during an OSCE and graded their performance using a standardized grading rubric. The treatment group then engaged in the identical motivational interviewing OSCE required of the students. One month later, both study groups viewed the same three videos and graded the students' MI performance using the identical standardized grading rubric. (See Appendix C).

Results: The overall intra-rater reliability was calculated using Cohen's Kappa statistic, pre-and post-intervention for both the control and treatment groups. Results revealed moderate to weak intra-rater reliability for both groups (.494). Fleiss' kappa statistic was used to assess interrater reliability. The treatment group achieved higher levels of agreement verses the control group on six of the ten questions. Only one question (See Figure 1: R06) had perfect or near perfect agreement in both study groups.

Conclusion: There was no statistically significant difference found in intra- and interrater reliability scores between the control and treatment groups following an active learning intervention. Even though statistical significance was not achieved,

individual faculty data suggests active learning did have an effect on the faculty in the treatment group. Post intervention, faculty in the treatment group had greater variations in scores indicating the experience had challenged their frame of knowledge and may have become more empathetic to the challenges of motivational interviewing having conducted an MI session themselves.

MANUSCRIPT

This manuscript will be submitted to the *Journal of Dental Education.*

Introduction and Literature Review

Given the ultimate goal of professional education is to prepare competent graduates ready to provide quality care to the public, the concern of inconsistencies and lack of standardization among faculty is well-founded (1,2). Calibration refers to training exercises in which standardization among evaluators is accomplished (1,2,4-5,12). The goal of calibration is to achieve high levels of examiner agreement through the use of criterion-based standards, and to be able to reproduce those standards in different situations (2). When multiple examiners are involved, two types of reliability must be measured: intra- and interrater reliability. Intra-rater reliability is defined as the agreement of the evaluators with themselves or degree of agreement throughout repetitive administrations of an assessment completed by the same rater (4,13-14). Interrater reliability is the ability of evaluators to consistently apply measurement criteria when provided with the same circumstances for evaluation (4,14).

Interrater agreement directly influences students' attitudes toward learning (1-4,6,15). Scoring inconsistencies resulting from poor calibration have shown to

affect not only grades but overall learning development (1,2,4). Several studies assert the reliability of evaluators is compromised when instructor feedback varies with similar student performance (1,2,4,7,12,15). Students begin to view faculty as incompetent when continuous variations in grading and evaluation practices exist. (1,4). Recurrent situations involving inadequate calibration can alter the integrity of clinical instruction, discourage student learning and decrease motivation to improve skills (1,2,4,12). Additionally, recurrent discrepancies in assessment have been shown to minimize students' passion to learn, reduce satisfaction with the learning environment, and affect patient care (1,2,4). Lack of calibration often diverts students from directing their focus on patient care and redirects their attention to satisfying the evaluator (1,3). A study by Jacks et al. revealed that lack of calibration among dental faculty was "a significant source of trouble, worry, and discomfort; a major source of anger; and one of the primary reasons for abandonment of a quest for excellence and registration to just getting by" (12). Numerous studies confirmed substantial discrepancy in assessment and clinical decision making among health care education faculty (1,4,7,16,19-21). Discrepancies among dental faculty have been shown within periodontitis diagnosis, treatment planning, calculus detection, cavity preparation assessment, radiographic interpretation, and periodontal probing (1,16,19-21).

Calibration Methods

Evident in the literature is a gap in knowledge pertaining to which calibration methods are most effective in achieving high intra and interrater reliability (1,3-5). Consequently, efforts have been devoted to identifying effective calibration

strategies to improve the level of interrater agreement. Among the most common calibration methods is the application of communication and verbalization in groups (4,5). Discussion-based calibration exercises have shown to improve the level of uniform feedback given by faculty (4,7,13). Goolsbee et al. and Metz et al. found using audience response systems (clickers) improved calibration among faculty (17,18). Metz and colleagues found instant feedback helped individual faculty members to assess their performance in relation to fellow faculty, improved interrater reliability and positively affected student opinions of faculty uniformity (17-18). The use of videotapes or recording devices has also shown to aid evaluators in visualizing and identifying specific standards within the presentation (4). An advantage of calibrating with video recordings is the ability to watch and evaluate students' performance multiple times and discuss grading discrepancies until consensus has been reached (4). Using videos for calibrating also provides benefits for uncovering inconsistencies among faculty, beyond psychomotor deficiencies, such as body language, miscommunication and other nonverbal interferences (4).

Simulation is a calibration method that engages faculty in performance of the skill to be evaluated. Garland and Newell used typodonts with artificial calculus to improve intra and interrater reliability among dental hygiene faculty with regard to calculus detection (1). Although the results were not statistically significant with regard to increasing interrater reliability, the study still stressed other positives that came from the calibration exercise such as becoming more conscious of their own exploring skills (1). A randomized experimental pilot study by Partido et al. sought to determine if incorporating periodontal endoscopy into calibration exercises would

increase both inter and intra-rater scores during calculus detection (2). Evaluators used an 11/12 explorer to detect simulated calculus on three typodonts. The treatment group had an additional two-hour calibration session using both the 11/12 explorer and dental endoscopy for calculus detection. A significant difference was found between pretest and posttest mean kappa averages for the treatment group vs. the control group. Additionally, the investigators found calibration training increased rater agreement among new and veteran faculty (2).

Calibration Challenges

Striving to reach an acceptable level of calibration is difficult, but even more challenging to maintain. Faculty members attended different schools at different times and were exposed to diverse clinical experiences resulting in grading inconsistencies due to variations in clinical judgement (13). Additional obstacles to calibration include years of experience among faculty, part-time verses full-time faculty, mandatory or optional calibration, and reimbursement for participation if calibration exercises are held after hours (1-5,7,12-13). A significant body of literature points to the challenges associated with implementing calibration exercises including, the need for human subjects, coordinating large numbers of participants, the sustainability of intra and interrater reliability (1,2,5-7,13).

Research to date is inconclusive on length of time for faculty recalibration, as it may depend on the particular skill set needing to be calibrated (13). Studies have described preservation of enhanced interrater reliability over pre-training assessments ranging from ten weeks to one year (7,12,16). Haj-Ali et al conducted a study to investigate the immediate effects of calibration when held to a gold

standard in regards to Class II amalgam preparation and found that calibration could be sustained for a ten-week period (7). Jacks et al. discussed the capacity of evaluators to evaluate dental hygiene students' SOAP notes and assessed the sustainability long-term. This study concluded that faculty-maintained calibration (as measured by the gold standard) for one year following the calibration workshop (12). The paucity of research on frequency and sustainability of calibration is evidence that more research needs to be conducted.

Active Learning Applied to Calibration

Active learning suggests that students need to be immersed in activities to learn as opposed to passively sitting and listening in a lecture setting (8-11,22-23). Active learning theory suggests individuals learn through building their own knowledge, connecting existing ideas, knowledge and experiences to form new or enhanced understanding (8-11). Given active learning activities have shown to increase understanding and lead to deeper learning, active learning may hold significance in the context of faculty calibration (8-10). Engaging faculty in the assessments they require of their students is a potential calibration strategy that may increase consistency of grading among faculty. Therefore, the purpose of this study was to investigate the effect of an active learning calibration exercise on the level of intra- and interrater reliability among dental hygiene faculty.

Methods and Materials

This study was approved by the University of Minnesota's Institutional Review Board (IRB), study identification number STUDY00003240. This study used a two-group randomized experimental design to evaluate the effect of incorporating

active learning into faculty calibration exercises on intra and interrater agreement among dental hygiene faculty. The study was conducted at the University of Minnesota during May, 2018- August, 2018. A convenience sample consisting of ten dental hygiene faculty members from the division of dental hygiene at the University of Minnesota (n=10) were used. All dental hygiene faculty members had consented to participate and successfully completed a day-long continuing education course on motivational interviewing prior to being able to participate in the study. The continuing education session was taught by an expert from outside of the University of Minnesota in May, 2018. The course served as the baseline training session on motivational interviewing for the dental hygiene faculty. Following the baseline training session, all dental hygiene faculty that participated attended a traditional calibration session which involved watching a video of a student using motivational interviewing with a standardized patient during an objective structured clinical examination (OSCE). All faculty graded the student's performance using a standardized grading rubric with defined criteria. (See Appendix C). Next, faculty discussed the ratings they had assigned for each criterion with the goal of reaching consensus on how to grade the performance of MI based rubric criteria.

Following completion of the calibration exercise, stratified randomization was used to assign faculty to the control or intervention group in an effort to obtain balance with respect to various levels of teaching experience among faculty participants. One week later, faculty in the intervention group performed the identical motivational interviewing session on the same standardized patients in the same environment as the students. One month later, all faculty met again to view three

different student videos using motivational interviewing on standardized patients and scored the students using the same grading rubric. Two months later, all faculty watched the same three students' videos and assessed student performance using the same grading rubric.

Statistical Analysis- The Cohen's kappa statistic was used to determine intra and interrater reliability for both the control and intervention groups. The kappa statistic is frequently used to test inter-rater reliability, as it is a more robust measure than percent agreement alone because it considers the agreement occurring by chance (13). Cohen's kappa uses a scale from -1 to +1: scores closer to +1 reveal a higher level of agreement and vice versa.

Instrument- A pre-existing grading rubric used for evaluating student performance of MI. The grading rubric was developed by dental hygiene faculty and had been validated over time. The grading rubric (Appendix C) had ten areas of evaluation and used a three-point Likert scale (1= poor, 2= fair and 3=good).

Operational definitions:

Objective structured clinical examination (OSCE): an assessment where faculty observe students' performance in a number of clinical knowledge and skill domains (24).

Motivational interviewing: Motivational interviewing is a directive, client-centered counselling style for eliciting behavior change by helping clients to explore and resolve ambivalence. It is most centrally defined not by technique but by its spirit as a facilitative style for interpersonal relationship (25).

Standardized patients: Skilled performers who act as patients during an interview,

physical examination, or OSCE and portray realistic patient characteristics and presentations of disease (26).

Active Learning: activities designed to involve an individual(s) in doing things and thinking about what they are doing (9).

Calibration: degree to which individual's judgements about their understanding, capability, competence, or preparedness correspond to the understanding, capability, competence, or preparedness they actually manifest (27).

Interrater reliability: the degree of agreement between data collectors (28).

Intra-rater reliability: consistency of a rater with oneself (4).

Results

All ten dental hygiene faculty members from the University of Minnesota's Division of Dental Hygiene consented to participate and completed the study. Eight faculty members held full time positions (four in each group) and two faculty members held part time positions (one in each group). Years of clinical dental hygiene experience ranged from less than one year to twenty plus years amongst all faculty. Teaching experience was similar between study groups; the treatment group had three faculty with less than one year to five years of experience; the control group had two. Both study groups had one faculty with six to eleven years of teaching experience; the treatment group had one faculty with twenty plus years of teaching experience, whereas the control group had two.

The overall intra-rater reliability was calculated using Cohen's Kappa statistic, pre- and post-intervention for both the control and treatment groups (3 students × 10 raters × 10 questions) rated twice (July and August 2018). Results revealed moderate to weak

intra-rater reliability for both groups (.494). The kappa score for the treatment group was .491 and the control group had a kappa score of .485. These scores revealed no statistically significant difference in intra-rater reliability between the treatment and control groups. Figure 1 provides data for each rater and score per question. For both study groups, the highest level of *intra*-rater reliability was associated with student A; the representation for July and August is almost identical for both the control and treatment group. The lowest agreement between the control and treatment groups was seen when comparing faculty ratings for student C. Further evaluation of each individual faculty's *intra*-rater reliability scores is presented in Table 1. The treatment group's kappa scores ranged from .318 to .645; the control group's scores ranged from .267 to .598. These scores demonstrated moderate to weak agreement with oneself and the difference in scores between groups was not statistically significant. Tables 2a-2c. provide the scores given by each rater for each student video session. Faculty in the treatment group had wider variations in scoring pre vs. post intervention. Within the treatment group, raters went up or down two points on the Likert scale on five occasions. Post-intervention, three of five faculty in the treatment group gave higher ratings to student B and C. The control group had three separate occasions where the score went up or down two points on the Likert scale.

Fleiss' kappa statistic was used to assess interrater reliability. Table 3 shows how consistent raters were on each of the ten questions. The treatment group achieved higher levels of agreement verses the control group on six of the ten questions, but not at a level of statistical significance. Only one question (See Figure 1: R06) had perfect or near perfect agreement in both study groups.

Discussion

The purpose of this study was to investigate the effect of incorporating active learning into calibration exercises on intra and interrater reliability. Overall, the intervention did not increase the level of agreement among dental hygiene faculty when assessing students using motivational interviewing on standardized patients as part of an OSCE. Results showed moderate to weak intra- and interrater reliability for faculty in both study groups. This result was a surprise to faculty, as they had conducted calibration sessions related to the assessment of MI many times in the past. Past calibration sessions had required faculty to view student videos, rate student performance and discuss grading discrepancies until consensus was reached. Consequently, prior to this study, faculty had the perception that there was a high level of agreement when grading the students' use of MI. This finding revealed the shortcoming of discussion-based calibration exercises. Faculty often leave a calibration session confident they will uniformly apply the grading criteria, but when the time comes to evaluate, they drift back to what is familiar and rely on past knowledge and grading practices they have been accustomed to using. Without sound data post calibration, the level of intra- and interrater reliability is often unknown. The body of literature reinforces this phenomenon, as past research has revealed mixed results on the effects of calibration. While calibration has resulted in improvement in interrater reliability, this finding is not consistent (7,12,16). It is clear in the literature that grading variations between raters after calibration exercises are present, and little is known about the calibration methods that are most effective (1-7, 16,19-21).

Even though statistical significance was not achieved, individual faculty data suggests active learning did have an effect on the faculty in the treatment group. Across all students, faculty in the treatment group had more variations in their ratings post intervention. After performing MI themselves with a standardized patient, faculty may have become less confident in their ability to grade this skill. Additionally, the majority of faculty in the treatment group leaned toward easier grading for students B and C, indicating they may have become more empathetic to the challenges of motivational interviewing having conducted an MI session themselves. Following the study, faculty members in the treatment group reported having gained the "students' perspective" and gained a better understanding of expectations for demonstrating MI in an OSCE setting. Studies evaluating student perspectives on clinical feedback found students felt feedback was more credible when evaluators had been trained with exercises that fostered subject mastery (2,4,12,13). Perhaps, the active learning activity, provided faculty with the opportunity to replace old knowledge and experiences with new knowledge and experiences resulting in new or enhanced understanding of the MI assessment. The MI OSCE experience may have served to 'level-set' and deepen faculty's understanding of the skill they were grading.

Interestingly, the highest rating consistency among faculty was seen for student A. This student earned the highest overall score by faculty in both study groups. Potentially, the better the student performs, the less variance in grading because a faculty member easily can more readily observe the desired student behaviors. A high level of agreement among faculty in both groups was also seen for

question six which presented a binary choice. The criteria did not allow for faculty interpretation. Both of these findings suggest revisions to the grading criteria may be warranted. Grading criteria may need to be written in more objective terms and be less open to interpretation.

Calibration holds significance in academia, especially in the health professions. Inadequate calibration may lead to inferior learning, lessen students' gratification regarding their education, and ultimately influence patient experience and care (1-2,4-5,7). To date, research studies assessing the impact of interrater and intra-rater reliability on student learning and satisfaction have presented conflicting findings. However, whatever the research studies reveal does not diminish the need for calibration, and these efforts must be continued. In order for new habits to take effect and remain intact, faculty must be repeatedly reminded of what has been learned, what is agreed upon, and the need to ensure that students are taught the same way, every day, by every faculty member.

Faculty in this study assumed they were calibrated, but in reality, they were not. This study demonstrated the importance of using data to confirm rater agreement post calibration. If active learning is associated with learning at a higher cognitive level and increased retention, the application of active learning to faculty calibration warrants further exploration (9). The small convenience sample was a limitation for this study; larger samples are needed. Educational institutions must continue to test specific calibration methods to establish evidence-based calibration protocols. Longitudinal assessment of intra and interrater reliability among dental and dental hygiene faculty should also be the focus of further research.

Conclusion

This experimental study investigated incorporating active learning into calibration exercises and its effect on intra- and interrater reliability among dental hygiene faculty. No statistical significance difference was found in intra- and interrater reliability scores between the control and treatment groups following an active learning intervention. However, performing motivational interviewing appeared to have an effect on faculty in the treatment group, as the range of intra-rater reliability scores grew wider indicating they became less confident in their ratings after the active learning exercise.

SECTION 4

TABLES

Table 1: Cohen's kappa for intra-rater reliability by rater

Rater ID	Group	Cohen's Kappa score
0	Treatment	0.439
1	Treatment	0.318
2	Treatment	0.579
3	Treatment	0.427
4	Treatment	0.645
5	Control	0.598
6	Control	0.520
7	Control	0.267
8	Control	0.423
9	Control	0.534

Tables 2a-2c: Rater scores for each video session

Table 2a

Student A Session 1: 7/24/18	Rater ID	R01		R02		R03		R04		R05		R06		R07		R08		R09		R10		Total score for July	Total score for Aug	Easier, Harder, same	Number of Variations per rater	Number of variations per group
Session 2: 8/27/18																										
Treatment Group	0	3	3	3	3	2	3	3	3	3	3	3	3	2	3	3	3	3	2	3	3	28	29	E	3	9 variations within treatment group
	1	3	3	3	3	3	3	2	2	3	1	3	3	3	3	3	3	3	3	3	3	29	27	H	1	
	2	3	3	3	3	3	3	3	3	3	3	3	3	2	3	3	3	3	2	3	3	29	29	S	2	
	3	3	3	2	3	3	3	3	3	3	3	3	3	3	2	3	3	3	3	3	3	29	29	S	2	
	4	3	2	3	3	3	3	3	3	3	3	3	3	3	3	3	3	2	2	3	3	29	28	H	1	
Control Group	5	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	30	30	S	0	6 variations within control group
	6	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	2	2	3	3	29	29	S	0	
	7	3	3	3	3	3	2	3	3	3	3	3	3	2	3	3	3	2	2	3	3	28	28	S	2	
	8	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	30	30	S	0	
	9	3	3	3	3	3	3	2	3	2	3	3	3	1	2	2	3	3	3	3	3	25	29	E	4	
Number of variations per question		1		1		2		1		2		0		5		1		2		0					15 total variations	

Table 2b

Student B 7/24/18	Rater ID	R01		R02		R03		R04		R05		R06		R07		R08		R09		R10		Total score for July	Total score for Aug	Easier, Harder, same	Number of Variations per rater	Number of variations per group
8/27/18																										20 variations within treatment group
Treatment Group	0	2	3	2	2	1	2	2	1	1	1	3	3	1	1	1	1	2	2	2	2	17	18	E	3	
	1	3	2	2	2	1	2	1	1	2	3	3	3	1	2	1	3	2	3	1	2	17	23	E	7	
	2	3	2	1	1	2	1	1	1	1	1	1	3	1	1	1	1	2	2	1	1	14	14	S	3	
	3	2	2	1	2	2	3	2	2	2	2	3	3	1	1	2	1	2	3	1	1	18	20	E	4	
	4	2	2	2	2	2	1	2	2	2	1	3	3	1	1	2	2	2	2	2	3	20	19	H	3	
Control Group	5	2	2	2	2	2	2	2	2	2	2	3	3	1	2	1	2	3	2	2	2	20	21	E	3	16 variations within control group
	6	3	3	2	1	1	1	1	1	2	2	3	3	1	1	3	1	3	3	3	2	22	18	H	3	
	7	2	2	2	3	2	2	2	3	3	3	3	3	2	2	3	3	2	3	2	2	23	26	E	3	
	8	2	3	1	1	2	1	3	2	2	2	3	3	2	1	3	3	2	2	2	3	22	21	H	5	
	9	2	2	3	3	3	2	2	2	2	2	3	3	2	1	1	1	2	2	1	1	21	19	H	2	
Number of variations per question		4		3		7		3		2		1		4		4		4		4					36 total variations	

Table 2c

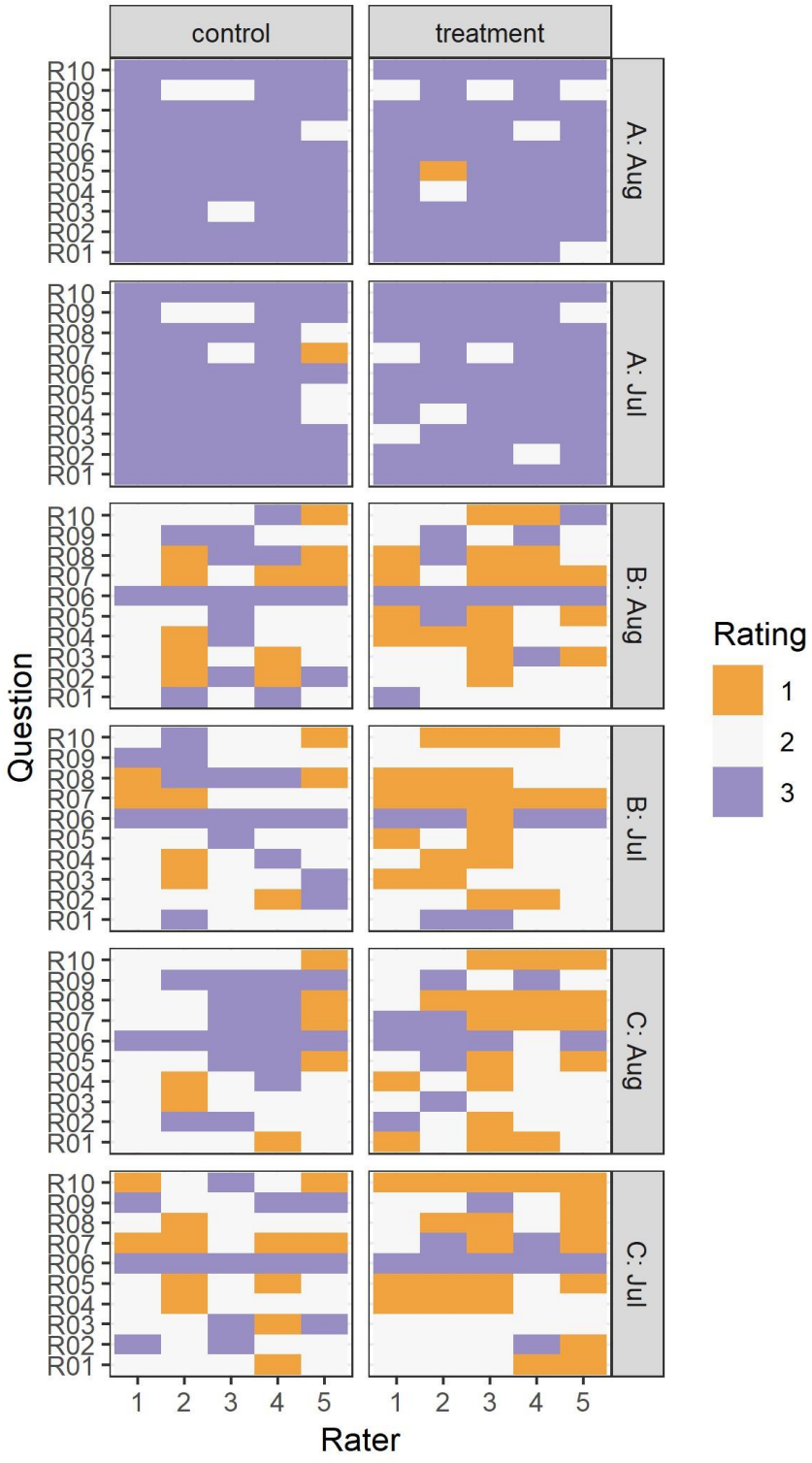
Student C 7/24/18	**Rater ID**	**R01**		**R02**		**R03**		**R04**		**R05**		**R06**		**R07**		**R08**		**R09**		**R10**		**Total score for July**	**Total score for Aug**	**Easier, Harder, same**	**Number of Variations per rater**	**Number of variations per group**
8/27/18																										
Treatment Group	**0**	2	1	2	3	2	2	1	1	1	2	3	3	2	3	2	2	2	2	1	2	18	21	E	5	21 variations among treatment group
	1	2	2	2	2	2	3	1	2	1	3	3	3	3	3	1	1	2	3	1	2	18	24	E	5	
	2	2	1	2	1	2	2	1	1	1	1	3	3	1	1	1	1	3	2	1	1	17	14	H	3	
	3	1	1	3	2	2	2	2	2	2	2	3	2	3	1	2	1	2	3	1	1	21	17	H	5	
	4	1	2	1	2	2	2	2	2	1	1	3	3	1	1	1	1	1	2	1	1	14	17	E	3	
Control Group	**5**	2	2	3	2	2	2	2	2	2	2	3	3	1	2	2	2	3	2	1	2	21	21	S	4	24 variations within control group
	6	2	2	2	3	2	1	1	1	1	2	3	3	1	2	1	2	2	3	2	2	17	21	E	6	
	7	2	2	3	3	3	2	2	2	2	3	3	3	2	3	2	3	2	3	3	2	24	26	E	6	
	8	1	1	2	2	1	2	2	3	1	3	3	3	1	3	2	3	3	3	2	2	18	25	E	5	
	9	2	2	2	2	3	2	2	2	2	1	3	3	1	1	2	1	3	3	1	1	21	18	H	3	
Number of variations per question		3		6		5		2		6		1		6		5		7		4					**45 total variations**	

Table 3: Fleiss' kappa for inter-rater reliability

Question	Overall Score	Treatment Group Score	Control Group Score	Difference between groups
R01	0.402	0.306	0.462	.156
R02	0.259	0.247	0.193	.054
R03	0.311	0.387	0.226	.161
R04	0.244	0.189	0.200	.011
R05	0.271	0.258	0.259	.001
R06	-0.026	-0.053	NA	NA
R07	0.177	0.247	0.122	.125
R08	0.274	0.399	0.173	.226
R09	0.000	0.042	-0.154	-.112
R10	0.400	0.477	0.306	.171

FIGURES

Figure 1: Representation of overall data for both treatment and control group



SECTION 5

PRACTICAL APPLICATION

The importance calibration holds in both student and faculty success is widely recognized in dental education. However, achieving a desired level of rater agreement has many challenges in the educational environment. It has been found that no matter how often calibration occurs or how engaging the instructor is, educators tend to revert back to the principles they are most comfortable with. This study investigated active learning as a new calibration method. Engaging faculty in the task/procedure required by their students may encourage "re-learning" of the procedure. The active learning component of the calibration exercise may assist faculty in replacing old knowledge with new and be an important step in "level-setting" and prevent faculty drift. A significant finding from this study was the perception of high rater agreement among faculty when in fact there was not. Calibration strategies aimed at consensus building may not always be effective, and therefore, interrater reliability should be measured to determine if a high level of agreement was achieved. Standardization of grading practices has the ability to impact student learning and therefore, the quality of care they will provide as future health care providers. The importance of calibration and achieving higher levels of agreement is not limited to only dental education. Utilizing the knowledge gained from past and present studies pertaining to calibration challenges provides opportunity to better educate the future educators in all aspects of health care. More research needs to be conducted on this specific issue in order to understand the best practices.

SECTION 6

APPENDICES

APPEDIX A: CONSENT FORM

Using Active Learning to Improve Intra and Inter-rater Reliability Among Dental Hygiene Faculty

You are invited to participate in a research study that uses active learning to investigate the effects it may have on intra and inter-rater reliability. You were selected as a possible participant because you are a current dental hygiene faculty member employed by the University of Minnesota School of Dentistry. Before agreeing to participate, it is important that you read and understand the following explanation of the proposed study. After reading the consent form, please ask researcher(s) to explain any information or details that you may have inquiries about.

Investigators:
The study is being conducted by:
Christine Blue BSDH, MS, DHSc, Associate Professor and Director Division of Dental Hygiene
Bridget Hotzler (Student investigator) RDH, MSDH student

It is primarily funded by the School of Dentistry Primary Care Department

Study Purpose:
The purpose of this study is to evaluate the effectiveness of active learning as it relates to the level of inter and intra-rater reliability among dental hygiene faculty. This study is significant as it may provide dental educators a superior alternative way to conduct calibration exercises. By utilizing and incorporating active learning intra and inter-rater reliability may increase. Providing alternative ways to conduct and improve calibration holds great significance within the dental hygiene program.

Study Procedures:
If you agree to participate in this study, we ask you to go verify that you can commit to the following details:

1. Attend Baseline:
    a. Attend and complete the continuing education (CE) course held on May 4th. This CE will thoroughly discuss motivational interviewing and will serve as the baseline training expected from all participants.
2. Verify Calibration dates:
    Specific days and times have been set aside for mandatory calibration sessions. These sessions will consist of watching student motivational interviewing objective structured clinical exams (OSCE), discussions, debriefing. These days are critical and attendance of all participants is expected.

    a. Thursday, June 21st 2018 from 1:00-2:30pm
    b. Wednesday, June 27th from 11:00-11:50am & 1:00-4:00pm (This date will be specific to the treatment group only) If participating in treatment group you will be performing the identical OSCE as the students.

c. Thursday, July 26th from 12:30-2:00pm
d. Monday, August 27th time TBD

Risks of Study Participation
This is a minimal risk study. The study will not ask you to perform or participate in any activities that will cause harm to you or others around you. You will not be asked to do anything that is not already expected from you as a dental hygiene faculty member.

Benefits of the Study Participation
Benefits of participating in this study include increased confidence in grading student performance of Motivational interviewing, consistent grading practices and improved student learning.

Study Compensation
There is no compensation for participating in this study.

Research Related Injury
This is a minimal risk study with minimal risk for injury. In the event that this research activity results in injury please inform the study investigator(s) as soon as possible, however, understand that the University and or the research investigators are **NOT** responsible or liable for any necessary treatment.

Confidentiality
The data from this research will be stored using Box security data storage system. All information will be un-identifiable and will NOT contain any personal information.

Voluntary Nature of the Study
Participation in this study is voluntary. Your decision whether or not to participate in this study will not affect your current of future relations with the University or the School of Dentistry. If you decide to participate, you are free to withdraw at any time.

Contacts and Questions
The researchers conducting this study are:

Christine Blue BSDH, MS, DHSc Phone: 612-625-5954 Email: bluex005@umn.edu
Bridget Hotzler RDH, MSDH student Phone: 763-412-6876 Email: mcgui182@umne.du

You are encouraged to ask any questions you may have now or at any time during the study.

Statement of Consent
I have read and reviewed the above information. I have asked questions and have

received answers. By signing I authorize and consent to my participation in the study.

Signature of participant _____
Date: _____

Signature of Investigator _____
Date: _____

APPENDIX B: Instrument (Standardized Grading Rubric)

Student Name_____
Evaluator/ Faculty Member_____

Self-Care Checklist for MI

Criteria Score:	**Good 3**	**Fair 2**	**Poor 1**
Rapport (R01)	Introduces self, role, brief small talk & behavior is engaging	Introduction, is short and very minimal small talk	Jumps right into self-care discussion
Reflections/Listening (R02)	Uses reflections to demonstrate listening, does not interrupt or change the subject	Some reflections are used, student minimally interrupts or changes the subject	Minimal reflections are used, students interrupts or changes the subject
Questions (R03)	Uses open ended questions more than closed ended	Uses about the same amount of closed ended questions as open ended questions	Most questions were closed ended
Holding back expertise/ Resist righting reflex (R04)	Did not lecture Did not try to "fix" the patient Conversational	Some lecturing occurred, occasional attempt to "fix" patient	Lectured without pausing Tried to "fix" the patient Non-conversational
Collaboration (R05)	Collaborated with the patient by eliciting ideas for change	Collaborated on some ideas for change	Did not collaborate
Asking Permission (R06)	Asks permission at least one time		Fails to ask for permission
Eliciting understanding (R07)	Checked understanding by asking mostly open-ended questions like what are your thoughts? What do you make of that? How does that sound?	Checked understanding by asking both open and closed ended questions.	Checked understanding by asking mostly closed ended questions like: Does that make sense? or Do you understand?
Emphasizes person's choice to change or not to change (R08)	Discusses change but allows the patient to make his/her own decisions	Briefly discusses change but does not provide detail or allow patient to share their entire thoughts	Makes recommendations and does not allow the patient to choose
Provides accurate content (R09)	Content provided was accurate and not confusing	Content was slightly confusing or aspects were inaccurate	Content was confusing and/or inaccurate.

Elicits summary from the patient. Student discusses next steps (R10)	Asks the patient to recap or summarize the discussion	Asks the patient if they understand what was discussed, but fails to have them summarize	Does not ask for a patient summary but instead provides the summary to the patient.

COMPREHENSIVE LIST OF REFERENCES

1. Garland KV, Newell KJ. Dental hygiene faculty calibration in the evaluation of calculus detection. J Dent Educ. 2009;73(3):383–389.

2. Partido BB, Jones AA, English DL, Nguyen CA, Jacks ME. Calculus detection calibration among dental hygiene faculty members utilizing dental endoscopy: a pilot study. J Dent Educ. 2015;79(2):124–32. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/25640616

3. Dicke NL, Hodges KO, Rogo EJ, Hewett BJ. A survey of clinical faculty calibration in dental hygiene programs. JDH. 2015;89(4):264–273.

4. Scruggs RR, George MC. Faculty Calibration in Clinical Education: A Review of the Literature. Educ Dir Dent Hygiene.1985;10(4):15–21.

5. Courts FJ. Standardization and calibration in the evaluation of clinical performance. J Dent Educ. 1997;61(12):947–950.

6. Knight GW. Towards Faculty Calibration. J Dent Educ. 1997 Dec;61(12):941-946.

7. Haj-Ali R, Feil P. Rater Reliability: Short- and Long-Term Effects of Calibration Training. J Dent Educ. 2005 Dec;70(4):428-433.

8. Bransford JD, Brown AL, Cocking RR. How People Learn: Brain, mind, experience, and school. Expanded Edition.Washington, D.C.:National Academy Press;2000.

9. Bonwell CC, Eison AJ. Active Learning: Creating Excitement in the Classroom. ASHE-ERIC Hight Education Report. 1991;(1);2-121.

10. Finkelstein A, Ferris J, Weston C, Winer L. Research-Informed Principles for (Re)designing Teaching and Learning Spaces. Journal of Learning Spaces. 2016;5(1):26-40.

11. Carr R, Palmer S, Hagel P. Active Learning: The importance of developing a comprehensive measure. Active Learning in Higher Education. 2015;(16):173-186.

12. Jacks ME, Blue C, Murphy D. Short- and long-term effects of training on dental hygiene faculty members' capacity to write SOAP notes. J Dent Educ. 2008;72(6):719–24. https://doi.org/72/6/719 [pii]

13. McAndrew M. Faculty Calibration: Much Ado About Something. J Dent Educ. 2016;80(11):1271-1272.

14. Landis JR, Koch GG. The Measure of Observer Agreement for Categorical Data. IBS. 1977 March;33(1):159-174.

15. Guild R. Questionnaire studies at three schools of dentistry. J Dent Educ. 1996;30(4):344-353.

16. Lanning SK, Best AM, Temple HJ, Philip SR, Carey A, McCauley LK. Accuracy and consistency of radiographic interpretation among clinical instructors in conjunction with a training program. J Dent Educ. 2006;70(5):545-557.

17. Goolsby SP, Young DA, Chiang HK, Carrico CK, Jackson LV, Rechmann P. The Effects of Faculty Calibration on Caries Risk Assessment and Quality Assurance. J Dent Educ. 2016;80(11):1294-1300.

18. Metz MJ, Metz CJ, Durski MT, Aiken SA, Mayfield TG, Lin W. A Training Program Using an Audience Response System to Calibrate Dental Faculty Members Assessing Student Clinical Competence. J Dent Educ. 2016;80(9):1109-1118.

19. Lanning SK, Pelok SD, Williams BC, Richards PS, Sarment DP, Oh T, McCauley LK. Variation in periodontal diagnosis and treatment planing among clinical instructors. J Dent Educ. 2005;69(3):325-337.

20. Sharaf AA, AbdelAziz AM, El Meligy OA. Intra-and inter-examiner variability in evaluating preclinical pediatric dentistry operative procedures. J Dent Educ. 2007;71(4);540-544.

21. Seabra RC, Costa FO, Costa JE, VanDyke T, Soares RV. Impact of clinical experience on the accuracy of probing depth measurements. Quintessesnce Int. 2008;39(4):559-565.

22. Handelsman J, Miller S, Pfund C. Scientific Teaching. 1st Edition. New York:W.H. Freeman and Company;2007.

23. Freeman S, Eddy SL, McDonough M, Smith MK, Okoroafor N, Jordt H, Wenderoth MP. Active learning increases student performance in science, engineering, and mathematics. PNAS. 2014; USA (111):8410-8415.

24. Wardman MJ, Yorke VC, Hallam JL. Evaluation of multi-methods approach to the collection and dissemination of feedback on OSCE performance in dental education. Eur J Dent Educ. 2017 April; (e203-e211). DOI:10.1111/eje.12273.

25. Rollnick S, Miller WR. What is Motivational Interviewing? BABCP. Oct

1995;23(4):325-334. https://doi.org/10.1017/S135246580001643X

26. Brender E. Standardized Patients. JAMA. 2005 Nov;294(9):1172.

27. Alexander PA. Calibration: What is it and why it matters? An introduction to the special issue on calibrating calibration. Learning and Instruction. 2013;4:1-3.

28. McHugh ML. Interrater reliability: the kappa statistic. CSMBLM. 2012 Aug;22(3);226-82.