# Determining Reliability and Validity of the Modified Gingival Index when Reviewed in an Image-based Survey Format

by

Samantha Eve Heron

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science

Medical Sciences – Dental Hygiene

University of Alberta

## *Abstract*

Gingivitis affects 50-90% of the Canadian adult population, with the accumulation of dental plaque being the most common cause. Clinically, gingivitis is characterized by red, swollen gums, bleeding upon stimulation, and a loss of stippled, knife-edged margins. Gingivitis will not necessarily develop into periodontal disease, but gingivitis always precedes periodontitis. Gingivitis can be diagnosed by inserting an instrument into the gingival sulcus to measure a bleeding response, performing tissue biopsies or by analyzing gingival crevicular fluid, however, in clinical research, the Modified Gingival Index (MGI) is a commonly used method for assessing the gingiva without performing invasive periodontal probing. The MGI is used to determine the inflammatory status of the gingiva by performing only a visual examination. A score of zero to four is provided for each tooth based on the physical appearance of the gingival surfaces and ranges from healthy (zero) to severe gingivitis (four). Although the MGI classification is a frequently used measure for diagnosing gingival disease, it is unclear whether the index is reliable and valid for research use when the gingiva of subjects is provided via an image-based survey format.

A total of 72 participants were recruited from the general dental hygiene clinic as well as the dental and dental hygiene student population of the School of Dentistry in the Kaye Edmonton Clinic at the University of Alberta. Anterior intraoral photographs of the subjects' teeth and gingiva were taken and used to create an online image-based survey administered through REDCap. Dentists, dental hygienists, and non-clinician researchers were asked to evaluate each photo by assigning a score based on the MGI definitions.

A cumulative logistic mixed-effects regression model was used to determine if one job type was more or less likely to assign a higher or lower MGI score to each subject. McFadden's pseudo-R2 was calculated from the regression model to establish the effect of job type and years of practice in assigning MGI scores. Krippendorff's alpha was used to determine reliability within- and across-groups of reviewers, and the ordinal alpha measured internal consistency of the index.

Overall, the results of this study indicate there is no consistency or agreement among dentists, dental hygienists, and non-clinical researchers when classifying the health of the gingiva based on the MGI criteria. This suggests that the MGI may not be a reliable or valid research instrument when subjects are presented in an image-based survey. The lack of agreement among reviewers suggests that the visual component of classifying gingival inflammation based on colour, contour, consistency, and texture may not be sufficient as when a bleeding component is included. Furthermore, the classification of severity of gingival inflammation may not be appropriate as the reviewers did not agree in assigning mild, moderate, or severe scores.

## *Preface*

This thesis is an original work by Samantha Eve Heron. The research project, of which this thesis is a part, received research ethics approval from the University of Alberta Research Ethics Board, Project Name "Salivary biomarkers of gingivitis", No. Pro00075629, October 15, 2017.

*"A mind is like a parachute.*
*It doesn't work if it's not open"*
Frank Zappa

*"If we knew what it was we were doing, it would not*
*be called research, would it?*
Albert Einstein

*"We do not need magic to transform our world; we carry*
*all the power we need inside ourselves already:*
*we have the power to imagine better"*
J.K. Rowling

*"Happiness can be found even in the darkest of times,*
*if one only remembers to turn on the light"*
Albus Dumbledore

## *Acknowledgements*

I want to thank my supervisors, Dr Ava Chow and Dr Minn Yoon for the support and mentorship. Thank you to my supervisory committee for the feedback and constructive criticism.

Thank you to the Dental Hygiene Graduate Research Fund for funding my research and the Dental Hygiene Graduate Research Stipend. A huge thank you to SHOFU Dental Corp for lending me the EyeSpecial C-II camera. I am eternally thankful to Matthew Pietrosanu of the Training and Consulting Centre (TCC) with the Department of Mathematical and Statistical Sciences for analyzing and helping interpret the statistics.

Thank you to all of my research participants and survey responders. I am grateful to all of the friends and colleagues that took my extensive survey. Thank you to all the clinical instructors for helping me find suitable participants.

I am thankful for my parents and my sisters - Candice, Kim, and Vanessa, for the immense support. Thank you to my friends, especially Nadia, Nicole, and Rebecca. A big thank you to everyone at Glenora for providing a supportive and loving environment. Finally, thank you to my partner, Dan, who provided me with an endless supply of support. I am so grateful for you.

# *Table of Contents*

# *List of Tables*

# *List of Figures*

## *List of Abbreviations*

ANOVA: analysis of variance

BI: bleeding index

BOP: bleeding on probing

BTI: bleeding time index

CAL: clinical attachment loss

CFA: confirmatory factor analysis

CI: confidence interval

EFA: exploratory factor analysis

EIBI: Eastman interdental bleeding index

GBI: gingival bleeding index

GI: gingival index

ICC: intra-class correlation

IUA: interpretations/use argument

MGI: modified gingival index

MPBI: modified papillary bleeding index

PBI: papillary bleeding index

PBS: papillary bleeding score

PD: pocket depth

PMA: papillary marginal attachment

QGBI: quantitative gingival bleeding index

REDCap: research electronic data capture

SBI: sulcus bleeding index

## *Glossary of Terms*

In the context of this thesis, terms are defined as follows:

**Alignment**: ensuring reviewers are consistent in how they differentiate scores from each other with proper training or comparison to a gold standard

**Consistency**: correlation among scores assigned by raters

**Construct**: a phenomenon of interest that is measured directly or indirectly

**Gingivitis**: inflammation of the gingiva predominantly caused by plaque biofilm

**Index**: a research tool used to generate valid, reliable, and reproducible information that can be applied from population to population using the same defined standards and definitions

**Internal consistency**: indicates correlation between different items in an index and whether the index appropriately measures the construct of interest

**Periodontitis**: an infection resulting in attachment loss and destruction of the bone surrounding the teeth

**Reliability**: the extent to which a measure is consistently or repeatedly performed each time it is used

**Stability**: how constant scores remain from one occasion to another

**Utility**: a subjective measure that determines how practical an index is when used in a real-world setting

**Validity**: evaluates the extent to which an index measures what it is intending to measure.

**Quality**: the reliability and validity of an index

# *Chapter 1: Introduction*

Gingivitis, or inflammation of the gums, is a common oral disease present in upwards of 60% of the adult population, with some studies suggesting close to 100% (1-4). The most common cause of gingivitis is the accumulation of dental plaque biofilm along the gingival margins of the teeth, prompting swelling and irritation of the gingiva (5-8). Gingivitis will not necessarily progress into periodontal disease, but gingivitis always precedes periodontitis (7). Periodontitis results in attachment loss and destruction of the bone surrounding the teeth and is the most common cause of tooth loss in adults (9). Clinicians diagnose gingivitis as red, enlarged attached and free gingiva, bleeding upon stimulation, and a loss of stippled knife-edged gingival margins (6, 8, 10-13). Researchers evaluate the health of the gingiva in several different ways, such as by measuring gingival crevicular fluid or performing gingival biopsies. However, a more straightforward technique used to diagnose gingivitis can be employed through the use of a gingival index, such as the Modified Gingival Index (MGI) (14-19).

According to Russell, an index is a "*number that defines the relative status of a population on a graduated scale, with definite upper and lower limits, for comparison with other populations classified by the same criteria and methods*" (20). An index should generate valid, reliable, and reproducible information that can be applied from population to population using the same defined standards and definitions. Developing an index requires a rigorous process of generating, testing, and re-testing to create an instrument that is specific enough to address the attribute in question, but broad enough that it will not leave out essential information. Once an index is developed, tests

of validity and reliability must be performed to determine the quality of the instrument (21). Reliability examines the reproducibility of an instrument every time it is used. Validity determines if the instrument is accurately measuring what it is intended to measure (21).

An increasingly common way of presenting oral health conditions is through digital photographs (22-26). A digital imaging system is advantageous as it allows for images to be examined at different time points and different geographic locations, as well as the ability to create a permanent database for future reassessment (27). Further, the examiner bias of in-person clinical examinations is reduced since photos can be randomized and anonymized. Several research studies report that photo-based examinations are equally reliable to a clinical examination (26, 28-31). As the MGI is predominantly used in research studies, it is unclear what the reliability of the MGI is when used across multiple researchers. Previous studies that have used the MGI typically only use one or two reviewers (14, 17-19, 32-35). Despite the MGI being used in multiple clinical trials since its introduction, the MGI has not been validated to classify the health of the gingiva when presented in a photo-based format.

This gap in knowledge led to the development of this study's research question: *is the modified gingival index a reliable and valid research instrument when used in photo-based evaluations?* This question is investigated by collecting anterior intraoral photos of subjects presenting with either healthy gingiva or varying degrees of gingival inflammation. These photos are displayed on a secure online platform and distributed to dentists, dental hygienists, and non-clinical researchers to examine and score according to the criteria of the MGI. Dentists and dental hygienists are included as it is expected that

these oral health professionals should accurately assign the score of the MGI to each photograph. These scores will form a baseline assessment to which non-clinical researchers can be compared. It is hypothesized that the researchers scores will agree with the oral health professionals as a result of the presentation of exemplar images of each gingival inflammation classification and an overview of the MGI provided at the start of the survey. The reliability of the MGI scores are determined across the groups of reviewers as well as within each group of reviewers (i.e., dentists, dental hygienists, non-clinical researchers). Based on the results of the reliability tests, the validity of the MGI as a photo-based research instrument can be determined. Specific research questions that are addressed in this study are:

1) Does job type impact how MGI scores are assigned?
2) How consistent is the MGI score assignment within a given profession?
3) Does the length of time in clinical practice impact MGI score assignment?

# *Chapter 2: Literature Review*

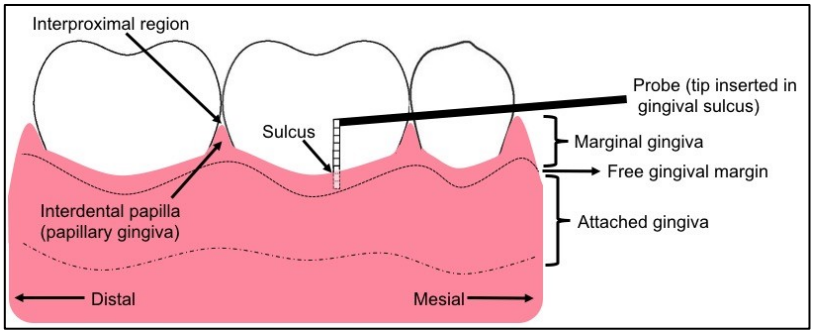
The following literature review provides an overview of the background supporting the development and conclusions of this research study. The way that gingivitis is diagnosed in a clinical setting differs from the way it is determined in a research setting. Clinicians tend to rely on the presence of bleeding to indicate inflammation, whereas researchers employ a method that can be quantitatively measured (36, 37). However, as gingivitis does not have a diagnostic gold standard measurement, it can be challenging to use an index that cannot be validated against this gold measure. An overview of index development is followed by a discussion of reliability, validity, and utility/responsiveness of indices. The history of the MGI is discussed including the strengths and limitations, as well as previous studies that have examined the quality of the MGI. In this study, the quality of an index is defined by reliability and validity. This review concludes with recent implementations of the MGI in current clinical research studies to emphasize the importance of determining the validity and reliability of this index.

## *2.1 Background*

Gingivitis is a common oral disease present in an estimated 50-95% of the adult population (1-4). Multiple genetic and environmental risk factors can predispose a person to gingivitis, but the most frequent cause of the disease is the accumulation of dental plaque biofilm resulting in inflammation (5-8). Gingivitis can be diagnosed by examining several features of the gingiva including changes to the colour, contour, consistency,

texture, sulcular temperature, and presence of bleeding upon provocation (8-10, 12, 13, 38-40). Clinically, this is characterized by red, enlarged attached and free gingiva, bleeding upon stimulation, and a loss of stippled knife-edged marginal gingiva, as seen in Figure 1 (8-10, 12, 13, 38-40). Muhlemann et al. were the first to propose and later confirm that bleeding on probing is the first sign of gingival inflammation (13). Despite the hallmark features that help identify gingivitis, it is still a relatively subjective diagnosis. Fortunately, there have been several gingival indices developed over the past seventy years to help quantify the degree of gingivitis into mild, moderate, and severe, as well as to decrease the subjectivity involved in diagnosing.

*Figure 1: Tooth and gingival anatomy based on cited indices*




A commonly used index in research settings to diagnose gingivitis is the Modified Gingival Index (MGI), developed by Ralph Lobene in 1986 (41). Despite ubiquitous use of this index in clinical research settings for the past thirty years, there is little evidence to support the quality of the index in its use as a research instrument. Hefti et al. examined 85 studies from 1996 to 2009 that employed the use of the MGI and determined that 45% of the studies provided no information regarding reliability assessments and 31% only used one examiner when using the MGI (42).

## *2.2 The Modified Gingival Index*

The MGI is a modification of Harold Loe's gingival index (GI) from 1963 (43). The GI is used to assess the quantitative conditions of the gingiva visually as well as mechanically (44). This is accomplished by inserting a periodontal probe into the gingival sulcus and probing the mesial, buccal, distal, and lingual surfaces of each tooth to determine if bleeding is elicited, as seen in Figure 1 (43). This bleeding is caused by micro-ulcerations in the sulcular epithelium and, as previously noted, is a cardinal sign of gingivitis (38, 45). The GI assigns a score from zero to three per each gingival unit (buccal, lingual, mesial, and distal) of the individual tooth. These scores from the four areas of the tooth are added and divided by four to provide the GI for each tooth. Finally, the values of all teeth are added together, and divided by the number of present teeth in the subject to classify the subject with healthy, or with mild, moderate, or severe gingivitis (43). The scoring criteria for the GI are as follows:

**0: absence of inflammation**

**1: mild inflammation**: slight change in colour, and little change in texture

**2: moderate inflammation**: moderate glazing, redness, edema, and hypertrophy; bleeding on pressure

**3: severe inflammation**: marked redness and hypertrophy; tendency to spontaneous bleeding; ulceration

A score between 0.1-1.0 is classified as mild gingivitis; 1.1-2.0 indicates moderate inflammation; and 2.1-3.0 signifies severe inflammation (43). The GI is frequently used to classify the health of the gingiva when a bleeding response is measured, however, there are limitations with this index. Firstly, the GI is an invasive

index as it requires a periodontal probe to be inserted into the gingival sulcus to elicit bleeding. The different techniques used by the operators can impact the amount of bleeding and/or how long it takes for bleeding to occur (46). These different techniques include probe angle, probe force, probing pattern, and the training provided to each operator (46). These techniques can impact how the reviewers classify the gingiva based on the onset and amount of bleeding. Another limitation is the ability for subsequent reviewers to assess the gingiva becomes challenging if bleeding has already been elicited by the first reviewer. Periodontal probing performed by the first reviewer may induce bleeding, even if as minimal as mild dotting, and subsequent reviewers may observe heavier bleeding upon probing since the gingival tissue has already been disrupted (47).

The limitations of the GI prompted Lobene et al. to develop the MGI in 1986 based on the previous criteria of the GI (41). This revision increases the sensitivity of the index by redefining the scoring system for mild and moderate inflammation as well as creating a completely non-invasive index (41). The criteria of the MGI is based on the following five ordinal values (41):

**0: healthy**: absence of inflammation

**1: mild inflammation**: slight change in colour, little change in texture of any portion of but not the entire marginal or papillary gingival unit

**2: mild inflammation**: criteria as above but involving the entire marginal or papillary gingival unit

**3: moderate inflammation**: glazing, redness, edema, and/or hypertrophy of the marginal or papillary gingival unit

**4: severe inflammation**: marked redness, edema and/or hypertrophy of the marginal of the papillary gingival unit, spontaneous bleeding, congestion, or ulceration

The MGI can be used as a full mouth index or applied to select teeth to determine the health of the gingiva. The MGI specifically examines the attached and free gingiva, as well as the texture of the marginal gingiva, as seen in Figure 1. Occasionally, partial evaluations of selected teeth are performed to reduce the cost and complexity of research studies (48). The Ramfjord teeth are commonly used which include the following six teeth: maxillary right and mandibular left first molars (tooth 16 and 36), maxillary left and mandibular right first premolars (tooth 24 and 44), and maxillary left and mandibular right central incisors (tooth 21 and 41) (49). However, Bentley et al. determined that using the Ramfjord teeth resulted in lower bleeding and MGI scores than compared to using half-mouth or full-mouth examinations (48). Conversely, half-mouth evaluations produced similar results when compared to full-mouth evaluations when determining plaque, MGI, and bleeding scores (48). Full mouth-evaluations provide the most information; however, these evaluations can be time-consuming and fatiguing for both rater and subject (48). The Ramfjord teeth do not produce enough information; therefore, the most appropriate amount of teeth to use ranges between half- and full-mouth examinations (48).

Another visual index, the Papillary Marginal Attachment (PMA), has similarities to the MGI in terms of use and functionality (50). However, there are important ways in which they differ. The PMA index divides the buccal surface of the gingiva into three separate anatomical areas to score: the papillary gingiva, the marginal gingiva, and the

attached gingiva, also detailed in Figure 1 (50). Each area is assigned a score between zero and four based on the following criteria:

**0: no gingivitis**

**1: mild inflammation**: slight change in color and little loss of contour

**2: moderate inflammation**: swelling, glazing and redness. Tendency to bleed on slight pressure. Papillae or margins become blunted or rounded in contrast to normal tissue

**3: severe inflammation**: more welling and redness, pocket formation, spontaneous bleeding

**4: very severe:** any degree more severe than above, including ulceration and sloughing

The numerical values of all teeth are totalled together to calculate the PMA index score for each subject (50). Researchers have critiqued the PMA by drawing attention to the definition of these three anatomical areas and how they can be interpreted differently from clinician to clinician, however, the same could be said about the MGI and the associated tissue of interest (51). Based on the subjective clinical determination, the PMA determines the degree of inflammation based on the affected anatomical area of the gingiva, while the GI or MGI determine the severity of inflammation at each gingival site around each tooth or the oral cavity as a whole (41, 44, 51).

There are some limitations of the MGI that should be noted. One is the variability involved in visually determining the colour and texture of the gingiva, especially in subjects that present with pigmentation (48). In a study comparing African-American, Caucasian, and Chinese school children where the raters assessed the gingival margins

based on the colour, determined low interrater reliability, thought in part, to be due to differentially pigmented gingiva between the different races (52). Bentley et al. agree that identifying the colour change of the gingiva is challenging to determine in different races, as opposed to an objective bleeding index, such as the GI, that is easier to visualize in all subjects, regardless of pigmentation (48). Another limitation is the subjective nature of the MGI when evaluating the appearance of the gingiva including colour, texture, and consistency (53). Despite frequent use of the MGI in clinical research settings, it is unclear what the validity and reliability of the MGI are when presented as an image-based format.

## *2.3 Index Development*

An index is a proxy used to measure a variety of conditions or characteristics that cannot be directly measured by quantitative techniques (54-56). Table 1 lists a number of gingival indices that have been developed over the past seven decades to evaluate the degree and severity of gingivitis. A variety of indices examine the presence or absence of bleeding such as the Gingival Bleeding Index (GBI), or length of time in which bleeding occurs, like the Bleeding Time Index (BTI) (57, 58). Given the large number of available indices evaluating gingivitis, several differing only by slight variations, it is essential to understand the development of indices to determine which index is the most appropriate to use in a given situation. Regardless of the index used in a research study, an understanding of the index development process will allow for appropriate utilization and interpretation of the index (59).

Indices are fundamentally developed to measure various constructs that can be measured directly or indirectly (60, 61). Direct constructs are measured values while an indirect construct, is an unobservable phenomenon (60). Gingival indices measure direct constructs as they examine and measure the gingiva directly.

*Table 1: Indices used to measure gingival health*

| Index | Author(s) (Year) | Instrument Used | Graded Responses |
|---|---|---|---|
| Papillary Marginal Attachment (PMA) | Schour & Massler (1947) | Visual | 0-4 |
| Gingival Index (GI) | Loe (1967) | Probe | 0-3 |
| Sulcus Bleeding Index (SBI) | Muhlemann & Son (1974) | Probe; wait 30 seconds | 0-5 |
| Gingival Bleeding Index (GBI) | Carter & Barnes (1974) | Unwaxed dental floss (twice); wait 30 seconds | Dichotomous (bleeding: yes/no) |
| Bleeding Index (BI) | Edwards (1975) | Dental tape (twice); wait 15 seconds | Dichotomous (bleeding: yes/no) |
| Gingival Bleeding Index (GBI) | Ainamo & Bay (1975) | Probe 3 to 4 times; wait 10 seconds | Dichotomous (bleeding: yes/no) |
| Papillary Bleeding Index (PBI) | Muhlemann (1977) | Probe | 0-4 |
| Papillary Bleeding Score (PBS) | Loesche (1979) | Stimudent (wooden interdental aid) | 0-5 |
| Modified Papillary Bleeding Index (MPBI) | Barnett et al. (1980) | Probe; wait 30 seconds | 0-3 |
| Bleeding Time Index (BTI) | Nowicki et al. (1981) | Probe 1 to 2 times; wait 15 seconds | 0-4 |
| Eastman Interdental Bleeding Index (EIBI) | Caton & Polson (1985) | Wooden interdental aid | Dichotomous (bleeding: yes/no) |
| Quantitative Gingival Bleeding Index (QGBI) | Garg & Kapoor (1985) | Dental brush; wait 30 seconds | 0-3 |
| Modified Gingival Index (MGI) | Lobene et al. (1986) | Visual | 0-4 |

DeVellis suggests an eight-step process to developing an index (59). This guideline is detailed in Figure 2 and compared against two other common strategies in designing an index by Boateng et al. and Morgado et al. (54, 55). The first step is to state precisely what it is that is to be measured. DeVellis states, "*this may be as simple as a well-formulated definition of the phenomenon they seek to measure*" (59).

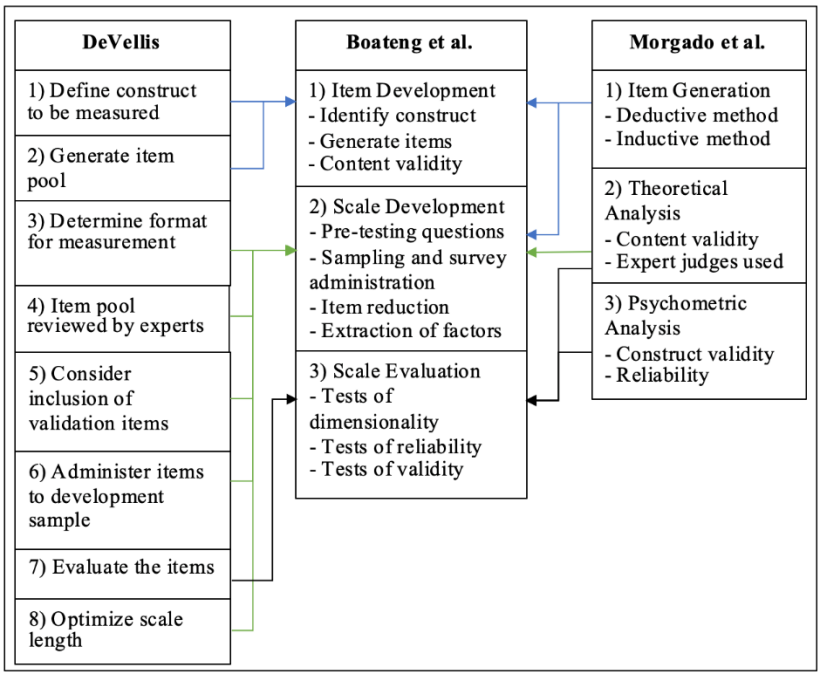
Furthermore, the researchers must describe how this new construct relates to and differs from existing concepts that are in use. Step two involves generating a pool of items that will become potential items in the index (59). Each item should be carefully constructed keeping the specific measurement goal in mind and ensuring clarity of expression. DeVellis stresses the importance of writing both positively and negatively worded items to avoid agreement bias. (62). If items are consistently worded to include the construct of interest (i.e., a positively worded item), DeVellis states that this may encourage the responders to agree with the item irrespective of their own opinion (59). Step three considers how to format the questions or statements of interest to present to responders (59). This includes formats such as the Likert scale which displays varying degrees of agreement, a visual analog scale where responses are presented on opposite ends of a continuum, a binary option such as "agree" or "disagree", or a semantic differential where a stimulus is represented by a list of adjective pairs on opposite ends of a continuum (59). Step four gathers all of the information collected and presents it to a group of people knowledgeable in the content area (59). This step allows experts to review the item pool and determine if the definition of the construct is appropriate. Also, these experts can evaluate the items for clarity and readability as well as to use their expertise of the phenomenon of interest to determine if any items are missing. The fifth step is to consider the inclusion of validation items that will help support construct validity. This form of validity, which will be discussed further in the next section, examines the correlation between the construct of interest and the variables that are related to the construct measured by the index (63). Step six involves administering a pilot of the index to a development sample. This step will help identify how large a sample size is appropriate

to determine statistically significant data. Future researchers that wish to employ this index will have an idea of how many subjects are required. Step seven involves evaluating the items in the index. If the index cannot be directly assessed against the true score, then inferences can be made using correlation tests (59). The higher the correlations are between two index items, the higher the reliability is between these two items (21). Other attributes to evaluate an index is item variance and item means (59). Finally, step eight in DeVellis' guidelines for developing an index is to optimize the index length. This step must find a balance between a long enough index to help indicate reliability but not too long that it becomes a burden on respondents (59).

Boateng et al. suggest a similar approach to developing and validating indices; however, this method involves three phases (54), as further outlined in Figure 2. The first phase of Boateng's et al. model is Item Development which includes identification of the construct and item generation as well as content validity. This phase is similar to the first two steps of DeVellis' guidelines where a construct is presented and defined followed by generating questions for the index to fit the defined construct of interest (54). The second phase is termed Scale Development and includes pre-testing questions, survey administration and sample size, item reduction analysis, and extraction of factors. This phase coincides with steps two through six, as well as step eight in DeVellis' guidelines. The final phase is Scale Evaluation which includes tests of dimensionality, tests of reliability, and tests of validity and is the seventh step in DeVellis' guidelines. A third strategy for developing an index is provided by Morgado et al. (55). Like Boateng et al., this guideline consists of three steps. Item generation consists of deductive and/or inductive methods to create an initial item pool (55). Deductive methods include

generating items that have come from rigorous literature reviews and previously developed indices (55, 64). Conversely, inductive methods create items based on information gathered from the target population including focus groups, expert panels, or interviews (55).

*Figure 2: Comparison of three different types of index development*




These three guidelines are compared in Figure 2 and illustrate that the index development process is complex, but typically follows a similar systematic procedure: 1) define the construct that is to be measured, 2) generate an item pool and have these items reviewed by an expert panel, 3) present these items to a development sample, 4) determine an adequate sample size, and 5) evaluate the index to determine sufficient validity and reliability. These guidelines differ in the ordering of steps, such as optimising index length being the last step in DeVellis' guideline where it falls in step two of Boateng's et al. and Morgado's et al. guidelines. An advantage of all three guidelines is the use of both inductive and deductive methods to generate an item pool.

Other index development guidelines in the past have used only one of the methods (64, 65). Using both methods is ideal as the quality of generated items will help define the construct of interest. A poorly defined construct may lead to confusion about what the construct is intended to measure and therefore, lead to an inadequately developed index (55). A disadvantage of Morgado et al. is the absence of tests of dimensionality that both DeVellis and Boateng et al. mention (54, 55, 59). Dimensionality examines the homogeneity of items, determining if a construct has a single facet or dimension (unidimensionality) or multiple ones (multidimensionality) that can account for correlation among index items (66). The advantage to performing tests of dimensionality allow for determining if the measurement of items is the same across two independent samples or across two separate time points from within the same sample (54). Further, dimensionality can help indicate the number of relevant scores that the index produces for each subject (59).

The most significant commonality between these three guidelines is the emphasis that they place on reliability and validity testing (54, 55, 59). These concepts and the importance of reliability, validity, and responsiveness/utility are discussed next.

### *2.3.1 Reliability*

Reliability is defined as the extent to which a measure is consistently or repeatedly performed each time it is used (21, 63). When determining reliability, measurement error, which can include an error in the subjects being examined, rater error, or error in the instrument itself, must be calculated to account for variability in the

data (21, 67, 68). Reliability can be calculated and defined by the following formula (42, 67, 69):

$$reliability = \frac{variability\ between\ study\ objects}{variability\ between\ study\ objects + measurement\ error}$$

Measurement error is the difference between the observed value and the true value while variability indicates how far data points diverge from the mean or other data points (67, 70). When applying the above equation, if a measurement error value is negligible, the reliability coefficient, or r-value, approaches one indicating a highly reliable measurement. Conversely, as a measurement error becomes significant, the r-value decreases indicating an increasingly unreliable measurement (59, 67, 69, 71). Reliability can be further supported by the following three concepts: internal consistency, interrater reliability, and stability (63, 70, 71).

Internal consistency measures how well different items measure a common construct (21, 63, 72). High internal consistency is favourable as it states that the items in the index have a high degree of correlation among each other (72). The most widely used tool for measuring internal consistency reliability is Cronbach's alpha, also called the coefficient alpha (21, 54, 63, 71, 72). However, Cronbach's alpha should only be used with continuous and non-dichotomous data as it is based on the Pearson covariance matrix which assumes data are continuous (73). A more accurate way to estimate the internal consistency of binary and ordinal responses is by using a polychoric correlation matrix as used in Zumbo's ordinal alpha (73). An alpha value $\geq 0.7$ and $<0.8$ is said to have an acceptable correlation to establish adequate internal consistency, a value $\geq 0.8$ and $<0.9$ has good internal consistency, and a value $\geq 0.9$ indicates excellent internal

consistency (71, 72). However, an alpha value that falls above 0.95 can indicate redundancy within the index (72). The split-half technique and Kuder-Richardson formula 20 (KR-20) can also be used to evaluate internal consistency. However, the split-half technique has been said to underestimate reliability by splitting the index in half, and the KR-20 formula is limited to dichotomous indices (21, 72).

Interrater reliability measures agreement across different raters using the same index (21, 63, 71, 72). It is important to calculate interrater reliability to determine consistency and agreement between scores assigned by raters, as this indicates the degree to which the data collected are accurate representations of the items measured (71, 74). It is assumed that these scores assigned by the raters are acquired completely independent of each other with no discussion or collaboration among the raters occurring (63).

Another assumption is that proper training of the index is provided to the raters before administration including examples of questions or items that can be expected. Without accounting for these assumptions, rater drift can occur (63). Rater drift is defined as variations in the scores assigned due to raters becoming more lenient or stringent in their scoring throughout the index (63). For categorical data, interrater reliability is determined by Cohen's kappa; however, it can only be applied when there are two raters (63, 71, 74). This measure takes into account the level of agreement as well as the agreement that can be expected by chance (63, 71). Landis et al. proposed a set of benchmark variables for the kappa statistic with between 0.81 and 1.0 defined as almost perfect agreement and a value between 0.61 to 0.80 as substantial agreement (75).

For continuous data, the level of agreement between raters is measured by intra-class correlation (ICC). Fisher introduced ICC in 1954 as a variation of Pearson

correlation coefficient (75). ICC is similar to Pearson correlation as they require variables that follow a linear relationship; however, Pearson correlation fails to take into account rater bias (76). Small rater bias is ideal as this indicates little difference in the means of the scores assigned between raters (76). ICC is calculated by estimating population variances based on the amount of variation in a given number of subjects (77). When interpreting an ICC, a low value may not only indicate a low degree of rater agreement but may also represent poor variability among subject samples or raters (77). Additionally, a low ICC value suggests that more subjects or raters are required for a given effect size to be statistically significant (78). The ICC does not have standard values for acceptable reliability; however, Koo et al. recommend that ICC values <0.50 indicate poor reliability, values between 0.51 and 0.75 indicate moderate reliability, values between 0.76 and 0.90 indicate good reliability, and ICC values above 0.91 indicate excellent reliability (77). Further, a 95% confidence interval is estimated when analyzing an ICC to identify the true range that the ICC value falls in (77). To summarise, ICC is an interrater reliability index for continuous data that reflects both agreement and degree of correlation between measurements.

For ordinal data, Krippendorff's alpha is commonly used to measure interrater reliability (79-82). This reliability coefficient is applicable when data sets are missing values, have small sample sizes, or if there are multiple raters or unequal sample sizes. (83). Krippendorff states that an alpha value above 0.8 can be considered good reliability and to only draw tentative conclusions between 0.667 and 0.8. This interrater reliability coefficient can be used for all levels of measurement including binary, nominal, ordinal,

interval, and ratio; however, Feng and Zhao et al. criticize that the alpha calculates lower coefficients despite when the levels of agreement are high (84, 85).

Stability is another concept that supports reliability. Stability of measurement is defined as how constant scores remain from one occasion to another (59). Stability is typically measured by test-retest reliability where the index is administered at two different time points to the same sample of raters, and the scores are compared between samples (63, 71). The term stability comes from minimizing fluctuations of the outcome due to daily or weekly changes and is the reproducibility of a set of results (71, 72). This type of reliability is based on the assumptions that no significant changes have occurred between the first test and the second; however, no test will receive the same results from test to test, so it is imperative to determine an acceptable level of error (21). The second test should not be given too soon to minimize overestimated reliability, where scores are affected by the memory of the first test, but also to diminish underestimated reliability, where too much time has passed, and the rater or subject have experienced changes such as further education or health status changes, respectively (71). Streiner et al. state that a retest interval of two to four weeks is typically seen (70). Methods of measuring stability often differ depending on the researcher; however, ICC and Pearson product-moment correlation appear to be the most common (54). (54). A shortcoming of test-retest reliability is determining if a low coefficient value obtained indicates an unreliable index, or if the index measures have changed over time (66).

Stability is further supported by the concept of intrarater reliability (71). This is similar to test-retest reliability; however, intrarater reliability assesses the reliability of repeated administrations of the index to the same rater to determine the variation that

occurs within the rater (70). Intrarater reliability is measured using similar tests as interrater reliability including ICC. Shrout et al. recommend a 2-way mixed effects model for intrarater reliability testing as it is not appropriate to assume one rater's score can be applied to a larger population of raters, as seen in a 1-way mixed effects model (77, 82). Additionally, for intrarater reliability measurements, absolute agreement over consistency should be chosen since measurements are unimportant if there is no agreement between repeated measurements (77). As intrarater reliability is measured using ICC, it is interpreted in the same manner with values greater than 0.90 indicating excellent reliability (77).

Reliability is defined as the degree to which measurements can be repeated under consistent conditions (77). It is supported by the concepts of internal consistency, interrater reliability, and stability. A highly reliable index indicates accuracy, reproducibility, and consistency from one testing occasion to another. However, a reliable index does not imply that it is also valid. Validity will be discussed further in the next section.

### *2.3.2 Validity*

Besides determining reliability, an index should also have the validity assessed. While reliability measures consistency and repeatability of an index, validity determines if the index truly yields what it is intended to measure (72, 79). Furthermore, to make accurate conclusions, the strengths and limitations of the index need to be clearly understood as well as the processes that were used to come to those conclusions. There are several ways to assess the validity, and a variety of approaches should be used when

evaluating an index. Educators first recognized criterion, content, and construct validity (70, 80, 86). More recently, validation studies focus on employing Messick's unified model which incorporates these three types of validation; however, the most recent change to validation takes into consideration Kane's argument-based approach to validation (70, 87).

When validity first became of interest in 1915, it was mainly used to look at how tests could predict future performances in a multitude of situations, such as university graduates and employment rates (88). This became known as criterion validity which takes an index and evaluates it for consistency with similar indices (72, 88). Criterion validity can be further broken down into two types known as concurrent validity and predictive validity depending on whether the criterion measure refers to a present or future state (72, 89, 90). Concurrent validity requires the index in question to be compared against a criterion measure that is known as the "gold standard" (21). The index under development should be applied against the gold standard to compare the results and determine agreement (63, 72). Developing a new index instead of the known gold standard can occur if the gold standard is too expensive, cumbrous, or invasive to use. The statistical test to measure concurrent validity is a correlation coefficient test with a high correlation suggesting good concurrent validity (72). Predictive validity is applied in a prospective study and allows an index to forecast future concepts such as behaviours, outcomes, or attitudes by administering the index in question at multiple time points (63, 89). Similar to concurrent validity, predictive validity is calculated as a correlation coefficient between the initial implementation of the index and the results of the second

administration of the same index (72). The challenge of predictive validity is that it typically demands a long study period, often making it impractical to employ.

To summarise, criterion validity compares an index to a measure that has already been established to be valid. Concurrent validity compares information from an instrument to the gold standard criterion at the time the index is administered while predictive validity compares measures at multiple time points (21). However, criterion validity cannot be appropriately validated if there is no standard reference, encouraging theorists to search for alternatives that could address this issue (91).

Content validity is defined as the degree to which items in an index accurately reflect the construct of interest (71). This can be done by having the content of the index reviewed by a set of experts in that area and/or compared to published literature (63, 72, 88, 92). This is done to ensure that the index has included all pertinent information (21). As a subjective approach to evaluation content validity, there is no "correct" answer, making it difficult for researchers to guarantee total content validity. This method of validity also contains face validity, a similar concept which examines if an index appears to measure what it purports to measure (21, 71). The difference between content and face validity is that content estimates how much the index represents the construct of interest where face validity only examines how reasonable an index appears to look. There are no statistical tests to measure content or face validity as they are presented as overall opinions from experts in the construct of interest (72). Despite not having a quantitative test to determine content validity, the information provided from this validity method can impact the conclusions that are drawn from the index.

Another validity method that does not use a standard reference is construct validity which became recognized by Cronbach et al. in 1955 (86). Construct validity is an indirect measure of validation that determines how appropriate the index is when being used and that it measures the construct it claims to be measuring (72, 88, 90). This type of validity is often not reported until the index has been tested for several years to accumulate evidence and observations (72). McDowell et al. mention that good studies of construct validity will: 1) define and justify the relevancy of hypotheses; 2) test stated hypotheses; and 3) disprove that the hypotheses measure anything else other than its intended measure (92). Construct validity is further composed of three facets: convergent, divergent, and factorial validity (21). Convergent validity tests the hypothesis that is designed to correlate with the constructs of interest being measured. Conversely, divergent validity, also known as discriminant validity, tests the hypothesis that is developed to not correlate with the index being measured (21). Alternatively, the index should not correlate with different or unrelated ones (90). Convergent validity assesses sensitivity while divergent validity tests the specificity of an index (21). The sensitivity of a test indicates the proportion of people that are positive cases who are correctly identified as such, where specificity implies the proportion of negative cases who are correctly identified as such (93). The final facet within construct validity is factorial validity. The statistical test, factor analysis, is used to divide larger sets of variables into smaller sets, called factors, with common characteristics (21). This way, factorial validity can examine how closely the factors being measured resemble each other in one or more themes, with each factor independent of the others (93). Additionally, this form of validation determines whether the items in the index all measure the same thing (21).

Construct validity cannot be established definitively as it is an ongoing process of developing and testing predictions that contribute to understanding the construct of interest (93).

In 1987, Messick published a research report stating that validity is not a unitary concept, but a unified, multi-faced concept that is based upon scientific inquiry (87). Messick suggested that the intent of validity "*is to account for consistency in behaviours or item responses, which frequently reflects distinguishable determinants*" (87). Since content and criterion-related validity both contribute to the interpretation of the index, these two concepts, according to Messick, should be enveloped under the umbrella term of construct validity. By combining these measures, a unified validity framework was conceived (87, 94). This framework is supported by evidence composed of five aspects of validity: content evidence, internal structure, relationships with other variables, response process, and consequences (86, 88). Content evidence, which incorporates the same concepts of content validity, ensures that the index reflects the construct that it intends to measure (95). Internal structure examines the relationship of individual index items, such as survey questions, with each other and with the overall construct of interest. An example of evidence that supports internal structure is determining interrater reliability (95). Relationships with other variables evaluates the *"degree to which these relationships are consistent with the construct underlying the proposed test score interpretations"* which can be assessed with correlation tests (95, 96). The response process assesses the degree to which the data received from the index reflects the construct of interest (95, 96). Evidence to support this includes rater training and performing quality control, such as video recording the raters. Finally, the fifth aspect of

Messick's framework is the impact of consequences which is defined as the impact of the index, whether negative or positive (95). This can include evidence such as subject experience.

Messick's framework is widely accepted by the American Educational Research Association, American Psychological Association, and National Council on Measurement in Education since 1999; however, several reviewers state that it provides incomplete guidance in prioritizing among different evidence sources and how this prioritization may change for different assessments (80, 89, 94-96).

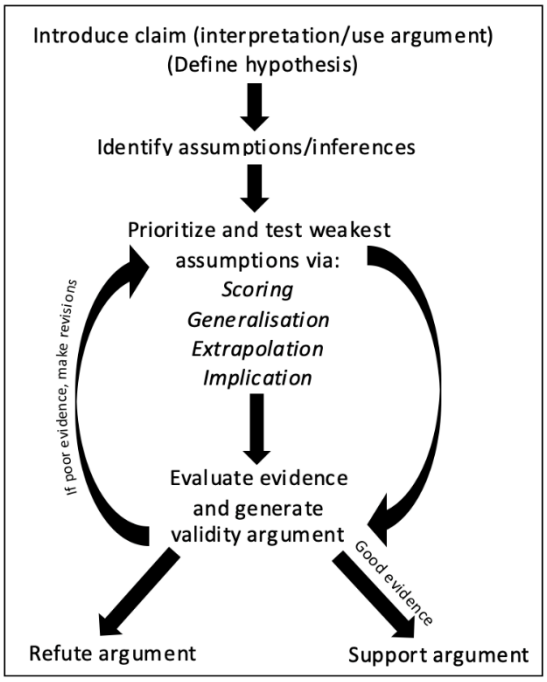
To address this problem, Kane introduced the argument-based approach to validation to address the process behind collecting validity evidence (89, 91). The premise involves defining and evaluating the claim being made (80). This idea is composed of two arguments within Kane's framework: the interpretations/use argument (IUA) and the validity argument (80). The primary purposes of the IUA are to identify and state any assumptions, to prioritize the weakest assumptions for further analysis, and to develop a plan for gathering evidence to support or refute each assumption (80, 94, 97). After collecting evidence, the assessors review the IUA and compare the results to the original assumptions to develop a validity argument describing how well the assumptions were supported or refuted (80, 97). Both the IUA and validity argument follow the scientific structure of developing and testing a hypothesis. Implementing this approach becomes challenging when attempting to identify weaknesses, gaps, and inconsistencies within a hypothesis and choosing the appropriate tests to evaluate these weaknesses. Kane suggests that researchers gather evidence across four key inferences, when appropriate, within the validity argument: “scoring”, “generalization”,

“extrapolation”, and “implications” (80). The “scoring” inference is achieved by evaluating assessment activities to determine if raters assign consistent scores after an observed performance, such as a multiple-choice questionnaire or a clinical testing session (86, 89, 94). Typically, a rubric or scoring card will be generated by a panel of experts so raters can appropriately assign scores based on scoring criteria (94). The “generalization” inference takes this sample of observations and determines if it can be applied to a "test-world universe" (94). This idea can be established by asking for experts' opinions and knowledge on the index, by reviewing past literature, or by continuing observations of similar indices (94). The “extrapolation” inference examines whether the observed index can relate to a real-world setting, while the “implications” inference determines the researchers' decision about the index (94). For example, if the evidence supports or refutes the interpretation of the index, this information may have an impact on society at large, stakeholders, educators, or learners. Figure 3 shows a flow chart amended from Cook et al. that illustrates Kane's argument-based approach to validation including the four inferences. Flaws can exist within the validity argument, such as citing interpretations as valid even though limited evidence has been evaluated, or proposing arguments and interpretations based on inappropriate validity evidence. Kane concludes that validation is a complex concept and not easy to accomplish, however, depending on the index interpretations and use of the data, validity can be reasonably achieved (89).

Messick shifted away from the idea of validity as a multi-faceted approach, and instead proposed a unified validity framework. Kane's approach to validation incorporated pieces of Messick's framework to allow new ways of organizing, evaluating, and presenting validation efforts. However, St-Onge et al. report that validity is still

poorly defined and is often improperly appraised (98). To claim strong validity, Cook et al. recommend using parts of both Messick and Kane's contemporary frameworks when collecting and interpreting data (86, 88, 91, 94, 95). It is important to ensure that assumptions are identified, stated, and evaluated before making claims about value or appropriateness of an index. To support claims of reliability and validity, responsiveness and utility of an index can help which will be discussed further in the next section.

*Figure 3: Kane's argument-based approach to validation*




*Adapted from Cook et al. (2012)*

### *2.3.3 Responsiveness and Utility*

Responsiveness examines how a measure can respond to change over time in regards to the construct of interest (63). Measuring responsiveness is critical to determine the changes that have resulted from the intervention and not "noise" that is due to measurement error (63). Kimberlin presents an example of responsiveness stating that using a scale built to weigh trucks cannot be used to weigh humans undergoing a new

weight-loss drug trial as the measurements will be too imprecise to identify small changes (63). In this scenario, the measurement will be valid but considered unreliable. Streiner questions whether responsiveness is a part of reliability, validity, or if it refers to a third, separate attribute (70). Moreover, Guyatt et al. state that when perfect validity is not achieved, responsiveness should be measured as a characteristic separate from reliability and validity (99). Measuring responsiveness is done by calculating the effect size, standardized response means, and/or the responsiveness statistic (93, 100).

Utility is a subjective measure that determines how practical an index is when used in a real-world setting (21). Bannigan et al. discuss that utility can be determined by considering the time it takes to administer the index, how easy it is to administer it and ensuring the language used to explain how the index works are appropriate and clear (21). Another important aspect of utility is to re-test the index in different settings to determine how a variety of people can use the measure (92). Both responsiveness and utility are often overlooked when developing and implementing an index, however, considering both concepts will add to the integrity of the index, ensuring that is appropriate to use and applicable in different settings.

To conclude, developing an index is a complex process that should be rigorously tested to ensure the reliability and validity of the index. The concepts of reliability and validity have been challenging due to evolving definitions and frameworks. Reliability ensures repeatability of research results and sets up a foundation for validity. Validity evaluates the extent to which an index measures what it intends to measure. Goodwin et al. summarise that "*an instrument cannot correlate with another external criterion (i.e., be valid) if it first does not correlate with itself (i.e., be reliable)*" (79). Regardless of the

misperception towards reliability and validity, the importance of implementing these concepts is an essential component of index development.

## *2.4 Clinical Uses of the MGI*

The MGI is commonly used in clinical research settings (14-16, 18, 19, 32) due to its non-invasive nature and repeatability by multiple reviewers (41). Indices which use a periodontal probe require training and achieving intra- and interrater calibration can be challenging. Differences in the amount of force exerted on the periodontal probe, the size of the probe, positioning of the probe, and disturbance of plaque and irritation of soft tissues upon repeated probing can all contribute to variability (41, 53). Though it can be argued that bleeding may be the earliest objective sign of gingivitis (53), Lobene et al. demonstrate that the MGI generates significant interrater correlation coefficients and can be considered comparable to the GI (40, 41). The anti-inflammatory efficacy of an antiseptic mouthwash was used to evaluate the reliability of the MGI over a six-month clinical trial (41). Sixty-seven subjects were examined and assigned MGI scores by three independent raters at four separate time intervals to determine the degree of agreement. The results concluded that the average interrater correlation coefficient using the Pearson Correlation measure was 0.81 indicating a positive linear relationship between raters (41). As mentioned earlier, a correlation coefficient between 0.75 and 0.90 is indicative of good reliability when a 95% confidence interval is used (77). This suggests good agreement between raters assigning MGI scores; however, the test-retest reliability results were not shown over the four separate time intervals. Overall, the authors

determined that the MGI has greater sensitivity at the lower end of the rating index and thus, can assess earlier signs of gingivitis.

The only validation that has been performed for this index was performed by Lobene et al. using an experimental gingivitis model to determine correlations between the MGI, GI, IBI, and PBI (47). Of the 99 subjects that completed the study, the highest correlations were found between the MGI and GI. The Pearson correlation coefficient between the MGI and GI produced a range of 0.844 to 0.932, followed by the MGI and PBI of 0.650 to 0.823, and MGI and IBI of 0.616 to 0.698 (47). This study demonstrates that the MGI strongly correlates with the GI and supports aspects of Messick's framework including content evidence, relationships with other variables, response process, and consequences. Internal structure does not appear to be examined as interrater reliability is not tested. The validity of this index is also supported through Kane's framework with evidence supporting scoring as the raters in this study were provided with a detailed description of how to score the gingiva based on each indices' criteria. Generalization, extrapolation, and implications cannot be determined based on the information provided from the study.

After Lobene et al. published the data supporting the validity of the MGI compared to the GI, the MGI became increasingly popular in clinical trials (14-16, 18, 19, 32, 34, 35, 47, 101-104). Barnett et al. used the MGI while testing computer-based thermal imaging techniques to determine if there were temperature differences between healthy gingiva and gingiva with increasing severities of inflammation (101). Ross et al. performed a double-blind, controlled clinical study to compare the effectiveness of 30 and 60 second Listerine rinses compared to a control mouthwash in inhibiting and

reducing plaque using an experimental gingivitis model where subjects were only accepted into the study if they had a mean MGI score ≥2.0 (104). Vastardis et al. implemented the MGI to compare the inflammatory state of gingivitis across HIV positive subjects with varying degrees of immunosuppression (35). More recently in 2018, Rai et al. compared MGI scores pre- and post-implementation of a customized toothbrush to determine decreased plaque scores and mean MGI scores (103).

These examples illustrate that the MGI has been used in a variety of research settings ranging in multiple topics; however, it is unclear if the quality of the MGI, including the reliability and validity of the index, is appropriate when presented in an image-based format to use in research studies.

## *2.5 Clinical versus Photographic Assessment*

There have been various research studies that have employed the use of photographs to diagnose various oral conditions. Practically, using a digital imaging system to assess oral healthcare is advantageous as it allows for an easy method of evaluation that can be performed by multiple reviewers at several different time points without the subject having to be present. Also, a permanent database can be created for additional research studies, and the photos can be scored blind.

Smith et al. demonstrate that an image analysis system where anterior intraoral photographs are taken with plaque disclosing solution on the upper and lower incisors is more sensitive in measuring plaque levels when compared to in-person traditional plaque measurement indices (24). After validating the use of the image analysis technique, Smith et al. investigated the reliability of measuring gingival inflammation through the

aforementioned photographic method (25). These findings by Smith et al. suggest that image analysis is a reliable method of determining plaque levels and gingival inflammation.

Furthermore, Cruz-Orcott et al. assess inter-examiner reliability between clinical examinations and photographic analysis of dental fluorosis (29). Cruz-Orcott et al. report that intra-rater reliability of photographic scoring of dental fluorosis was significantly higher when compared to scoring via clinical examinations (29). This suggests that using photographs may be a reliable method of diagnosing fluorosis.

Boye et al. compared diagnosing dental cavities using a full mouth examination versus intra-oral digital photographs in children aged five to ten (105). The results state that the intra-rater reliability had no statistically significant difference between visual examinations and photographic assessments indicating that photographs may be a reliable method of diagnosing cavities. These studies suggest that using photographs are a reliable method to diagnose various oral health conditions.

## *2.6 Conclusion*

Despite the MGI being used in many different clinical research settings since its conception, the quality of the MGI has not been thoroughly investigated since it was examined by Lobene et al. in 1989 (47). Furthermore, the reliability and validity of the MGI have not been examined when presented in an image-based format. After a comprehensive review of the evidence that determines an index to be reliable and valid, a large gap of knowledge became apparent regarding the MGI. This study will aim to determine the quality of the MGI when presented as a photographic research instrument to help alleviate ambiguity regarding validity and reliability.

# *Chapter 3: Methodology*

## *3.1 Study Design*

The purpose of this exploratory quantitative study was to determine the reliability and validity of the MGI when presented to reviewers in an image-based survey format. Intraoral photographs of subjects with a range of gingival health conditions were taken and used by dentists, dental hygienists, and non-clinical researchers to evaluate the gingival inflammation according to the MGI criteria. These photographs were uploaded to a secure web-based data collection tool and presented in a survey format to reviewers after MGI assignment instructions were provided. The data collected from the survey was used to determine reliability and validity of the index.

This study was reviewed and approved by the University of Alberta Research Ethics Board (Pro00075629).

## *3.2 Subject Recruitment and Participation*

Subjects for the study were recruited from the general dental hygiene clinic as well as the dental and dental hygiene student population at the School of Dentistry. Each subject was provided with a verbal and written summary of project and asked to sign an informed consent form before enrolling in the study.

For subjects recruited from the general dental hygiene patient clinic, dental hygiene students obtained and reviewed medical and dental history for each subject, performed extra- and intraoral examinations, and completed periodontal probing depth (PD) records. All clinical findings, including PD records, were reviewed and approved by

a dental hygiene clinical instructor. Subjects were selected for recruitment by one examiner (SH) based on inclusion and exclusion criteria by reviewing medical history and periodontal charts while in the clinic.

To add to the subject pool, dental and dental hygiene students were recruited by presenting the research question and the study requirements during a class lecture. The students all had active charts within the dental clinic as students were required to complete medical and dental histories as well as periodontal PD records on each other before working on the general public. The same examiner (SH) selected students based on the inclusion and exclusion criteria as determined by reviewing the medical history and periodontal charts. There was no compensation for subjects that agreed to participate.

The current study uses modified inclusion criteria developed by Syndergaard et al. (2014) that separated healthy subjects from subjects with gingivitis (106). The criteria were modified to only include subjects between the ages of 18 and 40 years of age to control for age-related periodontal disease. Syndergaard et al. included all subjects above 18 years of age (106), but findings from the Canadian Health Measures Survey indicate that 37.5% of adults 40-79 years of age have PD ≥5mm (107). In other words, people over the age of 40 have a higher likelihood of presenting with periodontitis which would confound the results. Subjects must have a minimum of 20 teeth and must have an active chart with the School of Dentistry.

Exclusion criteria include PD ≥5mm at any of the six sites of each tooth and clinical attachment loss (CAL) ≥2mm at any tooth as these measurements are classified as slight periodontal disease (108). Subjects were also excluded if there was presentation with any of the following conditions as they have been determined to increase the risk for

periodontal disease: auto-immune disorders (109), diabetes, cancer therapy or organ transplants (110, 111), pregnancy or lactation (112), use of antibiotics or immunosuppressant medication within the last month (113), removable prosthodontic or orthodontic appliances (114), the presence of an oral mucosal inflammatory condition (e.g., lichen planus, leukoplakia, and oral cancer) (115, 116), or previous diagnosis or treatment for periodontal disease.

Once a subject was enrolled in the study, a five-digit randomized, numerical code was generated and assigned to each subject. This code was used throughout the study to anonymize subjects and to ensure all documentation and photographs were linked to each subject.

To obtain photographs that provide a view of both arches, each subject inserted an intraoral cheek retractor (VWR International Co., Mississauga, ON) to position the lips and cheeks away from the teeth. Two anterior view photos were taken using the EyeSpecial C-II Camera on loan from SHOFU Dental Corporation. Two photos were taken to ensure that the resolution and quality of the photos were adequate. The photograph that had better resolution and a centred midline in the photo was chosen to use in the survey. The photos were taken without the use of the overhead dental light as the EyeSpecial C-II Camera is calibrated to provide even and consistent lighting, exposure and focus between subjects. Each photo was taken by the same member of the research team (SH). All photos were stored on an encrypted computer before being permanently erased off the memory card of the camera.

### *3.3 Reviewer Recruitment*

The reviewers for this study consisted of dentists, dental hygienists, and non-clinical researchers. Both dentist and dental hygienists were included as reviewers as they are both oral health professionals but have different focuses, areas of expertise, and education. These oral health professionals were included to examine the variability among MGI scores assigned by these reviewers. Non-clinical researchers were included since the MGI is used as a clinical tool in research studies.

A non-probability, convenience sample was used to email and invite reviewers to participate in this study. Emails were sent out to a group of dentists and dental hygienists through word-of-mouth. The non-clinical researchers were recruited via email from various departments at the University of Alberta. Emails were sent to five individuals initially and if one reviewer did not complete the survey, another email was sent until at least five surveys were completed. A convenience sample was justified for the exploratory purpose of this study to examine the assignment of MGI scores by different reviewers.

### *3.4 Sample Size*

The sample size for this study was informed through a power analysis using a mixed effects linear regression model. This model incorporates the average response of reviewers. As seen in Table 2, simulated power values are shown for 60, 65, 70, 75, and 80 subjects each reviewed by three through seven reviewers, using a statistical significance threshold of 0.05 and assuming an average difference in scores between

professions of 0.5. A significance threshold of 0.05 was chosen to indicate strong evidence of significance.

*Table 2: Power of the linear regression model with a given number of reviewers and subjects*

| | Subjects | | | | |
|---|---|---|---|---|---|
| **Reviewers of each job type** | **60** | **65** | **70** | **75** | **80** |
| 3 | 0.79 | 0.72 | 0.78 | 0.74 | 0.81 |
| 4 | 0.81 | 0.89 | 0.82 | 0.79 | 0.75 |
| 5 | 0.87 | 0.86 | 0.86 | 0.82 | 0.90 |
| 6 | 0.90 | 0.91 | 0.93 | 0.91 | 0.92 |
| 7 | 0.94 | 0.99 | 0.93 | 0.97 | 0.93 |

These simulations set the largest true difference in average MGI scores between professions to be 0.5. Additionally, we simulated random effects for reviewers with a standard deviation of 0.167 (0.5 divided by 3). These parameters were based entirely on speculation as no literature has examined the average differences in MGI scores between clinicians. To determine subject variability, we used a study by Carvajal et al. that examined 1650 individuals above the age of 18 who determined that gingivitis is present in 95.6% of individuals. Further, 22.5% of adults have mild gingivitis, 74% present with moderate gingivitis, and 3.6% have severe gingivitis (1). This study aimed to recruit at least 70 subjects, five each of dentists, dental hygienists, and non-clinical researchers to achieve a power value of 0.86.

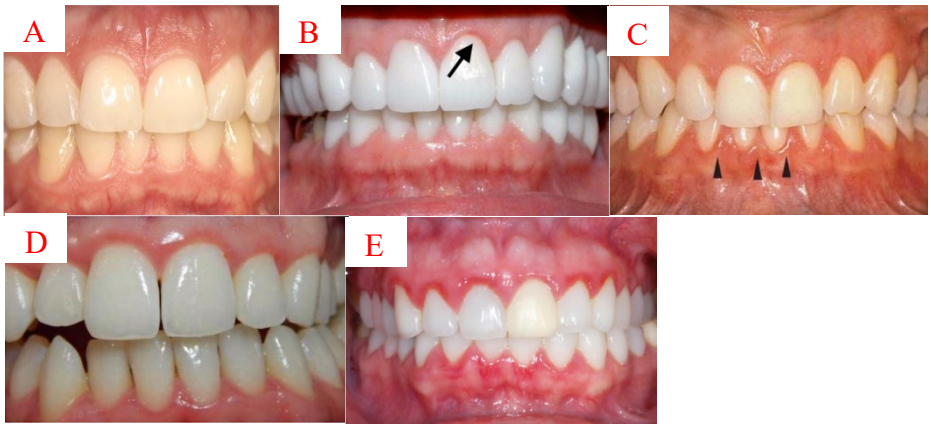
## *3.5 REDCap Survey*

A secure web-based data collection tool called Research Electronic Data Capture (REDCap) database was used for building and managing an online image-based survey using the intraoral photographs to present to the reviewers for MGI determination. The survey was created with support and advice from the Data Management and Informatics

of REDCap Support at the Women and Children's Health Research Institute at the University of Alberta.

Once a reviewer accepted the invitation to complete the survey, a link was emailed to each via the REDCap Survey website. This link was unique to each reviewer to ensure that the results would be saved if the reviewer wished to exit and return to complete the survey at a later time. The data was anonymized by REDCap; however, identification of the job type was indicated.

*Figure 4: Exemplar photos of each MGI category shown to reviewers*




*A: healthy; B: mild, localized; C: mild, generalized; D: moderate; E: severe*

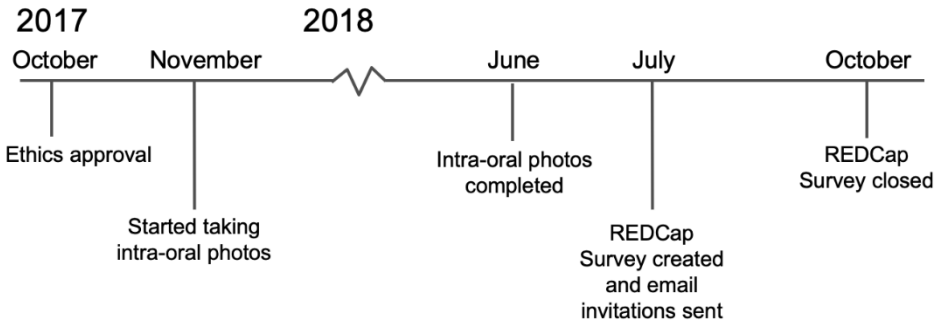
The introduction of the survey collected information about the reviewer including job type (dentist, registered dental hygienist, researcher); and the following information was asked of the clinicians: time in clinical practise, and school from which their clinical diploma/degree was obtained, as seen in Appendix A. Following the introduction, the survey provided a detailed explanation of the MGI with exemplar photos of each MGI category as seen in Figure 4 and Appendix B. This page was accessible throughout the study to ensure comprehension of the MGI and to assist in calibration among reviewers.

When the survey began, a total of 72 anterior view intraoral photos of the subject's gingiva were presented with one photo on each page of the survey. Reviewers were asked to evaluate each of the 72 photographs and determine the level of gingivitis present using the criteria defined by the MGI.

## *3.6 Timeline*

This study took place over a period of one year starting in October 2017 as seen in Figure 5. Once the REDCap survey was closed after receiving data from at least five reviewers from each job type, a statistician was hired to analyze the results of the survey.

*Figure 5: Timeline of this study*




## *3.7 Statistical Methods*

To model the relationship between a set of predictors (i.e. job types) and the probability of each level of an ordinal response (i.e. the assignment of MGI scores), a cumulative logistic mixed-effects regression model (CLMM) was used. A mixed-effects model includes both fixed effects and random effects. Fixed effects are unknown constants and are of primary interest as opposed to random effects that are assumed to be a random sample from a much larger set of effects (117). In this study, the CLMM was

chosen as it can account for fixed effects including job type and years of experience, and the random effects of subject and reviewer due to the non-independence in scores from the same reviewer or subject. Furthermore, the logistic model is designed to consider the ordered but non-continuous and more importantly, the non-interval type measure of ordinal response data (118), as well as taking into account more than two response levels. And finally, the CLMM was an appropriate statistical test to use as it allows multiple variables to be analyzed at the same time while reducing the effect of confounding factors (119).

Probability describes the likelihood of an event happening where odds are the probability of the event occurring divided by the probability of the event not occurring. An odds ratio can be written as:

$$\frac{odds\ of\ event\ in\ A\ group}{odds\ of\ event\ in\ B\ group}$$

Linear regression cannot be applied to an odds ratio due to the potential of negatively predicted values. In this study, this model uses cumulative probabilities up to a certain threshold, making the range of ordinal values binary at that threshold. The response of MGI scores = 1, 2, …, *j* where the ordering is natural. The associated probabilities are $\{\pi_1 + \cdots + \pi_j\}$, with a cumulative probability of a response less than or equal to *j* is (115, 116):

$$p(MGI\ score\ \leq j) = \pi_1 + \cdots + \pi_j$$

With the probability (p) of assigning the log-odds of an MGI value of *j*. A cumulative logit is then defined as:

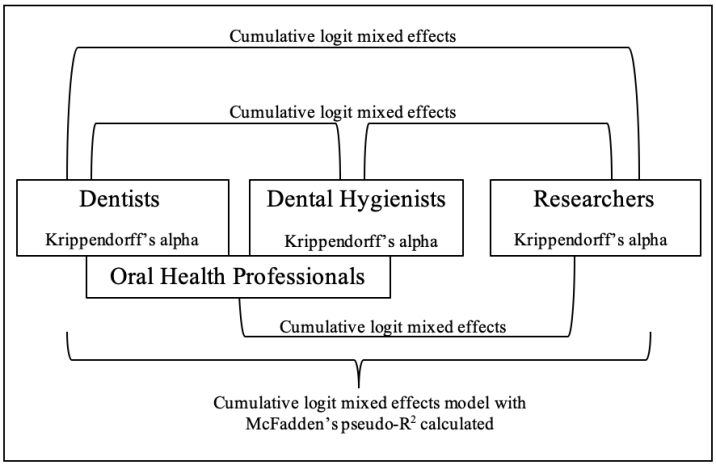
$$\log\left(\frac{p(MGI\ score\ \leq j)}{p(MGI\ score > j)}\right) = \log\left(\frac{p(MGI\ score\ \leq \mathrm{j})}{1 - p(MGI\ score\ \leq \mathrm{j})}\right) = \log\left(\frac{\pi_1 + \cdots + \pi_j}{\pi_{j+1} + \cdots + \pi_j}\right)$$

The equation describes the log-odds of the MGI score being less than or equal to *j*. This measures the probability of the response in category $\leq j$ and in a category $>j$. This model was also used to determine if the length of time in clinical practice of oral health professionals had an impact on MGI score assignment.

To make the mixed-effects regression model meet the model assumptions, scores of four (severe) were removed. The proportional odds assumption used in this model deals with the effect of model predictors on the log-odds of assigning a score of zero is the same as assigning a score of $\leq 1$, or $\leq 2$, or $\leq 3$. The assumption is that these effects are the same for each of these log-odds. When a score of four was included, the proportional odds assumption was found to be violated. To try and account for this, scores of three and four were merged; however, this still did not make the proportional odds a reasonable assumption. Of the 1224 observations from 72 subjects evaluated by 17 examiners, 53 (4.3%) of these observations were scores of four. Given so few observations, it was deemed reasonable to remove these scores entirely. To ensure that removing these scores did not affect the data, a likelihood ratio test was done to verify the proportional odds assumption, and there was no statistically significant difference between the model assuming and not assuming proportional odds. Effect estimates in the CLMM can be interpreted as additive differences in cumulative log-odds. If an effect estimate is equal to zero, there is no difference between the two groups being compared in the cumulative log-odds of assigning different MGI scores. These p-values were determined by Wald-type tests of significance for the model estimates in the CLMM assuming a normal distribution of the test statistic. It should also be noted that the comparison against oral health professionals and researchers was determined using a second, separate CLMM.

The benefit of the CLMM is that it accounts for baseline differences between reviewers and subjects and provides numerical estimates for the effect of job type on the cumulative log odds, allowing the comparison of one job type to another. Most importantly, the CLMM treats the ordinal natural of the MGI response properly. The CLMM was chosen to determine if one job type was more or less likely to assign different MGI scores. Figure 6 details the statistical methods used for comparisons within and between each group of reviewers.

*Figure 6: Visual representation of statistics applied within and across reviewers*




To determine the source of variability explained by the CLMM, McFadden's pseudo-$R^2$ was calculated from three regression models including a model that accounts for random effects individually, fixed effects individually, and a model accounting for random effects and the fixed effect of job type. The traditional $R^2$ is the square of the correlation between the model's predicted values compared to the actual values (120). The traditional $R^2$ has a correlation that ranges from negative one to one (120). However, a pseudo-$R^2$ was selected because the logistic regression models that were used cannot use negative values. When an $R^2$ value is squared, this square correlation now ranges

from zero to one and is known as a pseudo- $R^2$ (121). The term "pseudo" indicates a similar statistic to $R^2$ where higher values towards one will suggest a better fitting model (121). The more variability that is explained by the model, the better the data fits the model. There are a variety of types of pseudo-$R^2$ statistics that are commonly used for non-traditional logistic regression; however, the most interpretable type to use for this study was McFadden's pseudo-$R^2$. This is also known as the "likelihood-ratio index" where it compares the intercept-only model (the "null model") and the "model with no predictors" with the likelihood of the model we are considering (120). By calculating the pseudo-$R^2$ from multiple CLMMs, the usefulness of these covariates of job type and years of practice as predictors of cumulative log-odds could also be determined.

To address the consistency of the MGI score assignment within a given job type, Krippendorff's alpha statistic was calculated. Explicitly designed for ordinal data and its analogy to Cronbach's alpha, this was used as a measure of inter-rater reliability between the reviewers of each job type as well as across all 17 reviewers. Krippendorff's alpha was chosen over other agreement coefficients as it is applicable for measurement across several reviewers, it can be used for large sample sizes, and it is applicable for multiple scales of measurement, especially ordinal values (122). Krippendorff recommends to only rely on data with a reliability alpha value above 0.8 and to only draw tentative conclusions with a value between 0.667 and 0.8 (122). Furthermore, it is recommended that the reliability of the data be rejected when the confidence interval falls below the smallest acceptable reliability of 0.667. In this study, the confidence level was set to 95%, and a p-value of $\leq 0.05$ was considered statistically significant. Data are bootstrapped and

estimate follows a normal distribution. Krippendorff's alpha was chosen to determine the agreement within and among all job types.

Zumbo's ordinal alpha was used to measure internal consistency as this reliability coefficient has been shown to be more accurate than Cronbach's alpha for estimating reliability for binary and ordinal responses (73). The ordinal alpha incorporates a matrix of correlations among all items in an index by using a polychoric correlation matrix rather than the Pearson covariance matrix as used in Cronbach's alpha (73). An alpha value $\geq 0.7$ and $<0.8$ is said to have an acceptable correlation to establish adequate internal consistency, a value $\geq 0.8$ and $<0.9$ has good internal consistency, and a value $\geq 0.9$ indicates excellent internal consistency (71, 72). A 95% confidence boundary was set to establish the coefficient alpha. Internal consistency determines if the MGI truly measures the construct of interest, which in this study is gingivitis. Zumbo's ordinal alpha was the most appropriate statistical test to determine the internal consistency within this study.

A post-hoc logistic regression was used to compare CLMM results by ignoring the ordinal nature of the MGI scores and instead considering a binary response. A separate mixed logistic regression model was calculated with random effects accounting for random effect differences between subjects and reviewers. The scores of the MGI were split into "healthy" (scores of zero) and "inflammation" (scores of one, two, three, and four). It was thought that perhaps the scoring of the MGI might only be useful in distinguishing healthy gingiva versus inflammation. This allowed for the examination of how different job types classified types of inflammation compared to healthy gingiva. Based on the results of this logistic regression model, the MGI scores assigned were

separated into “healthy” (scores of zero) and “severe” (scores of four) to determine if one job type was more or less likely to distinguish healthy versus severely inflamed gingiva. A second, separate mixed logistic regression model was calculated with random effects accounting for random effect differences between subjects and reviewers.

Finally, a separate CLMM was used to examine reviewer fatigue throughout the length of the survey. The data were numbered from one to seventy-two based on their ordering in the dataset received by each reviewer. The observations were treated as continuous since the variables are measured along a continuum. These observations were added into the CLMM including fixed and random effects to determine if the reviewers were more or less likely to assign higher MGI scores as the survey progressed. Performing this statistical test can help determine if the length of the survey is appropriate.

# *Chapter 4: Results*

Seventy-two people (42 females and 30 males, aged 19 to 35 years; mean age: 24.1 ± 3.2 years) participated in this study, as seen in Table 3. No subjects reported tobacco smoking; however, one subject indicated marijuana smoking one to two times a week. Brushing habits ranged from once a day (5.6%), twice a day (86.1%) to more than twice a day (8.3%), and flossing ranged from never (4.2%) to greater than two times a day (54.2%).

*Table 3: Subject demographics by sex, smoking, brushing, and flossing habits; (mean ±SD)*

| Subject Characteristic | *n* | (%) | Mean Age |
|---|---|---|---|
| **Sex** | | | |
| Female | 42 | 58% | **23.4 (± 2.9)** |
| Male | 30 | 42% | **25.1 (± 3.4)** |
| Total | 72 | 100% | 24.1 (± 3.2) |
| **Smoking** | | | |
| Tobacco | 0 | 0% | |
| Marijuana | 1 | 1.4% | |
| **Brushing habits** | | | |
| <1x/day | 0 | 0% | |
| 1x/day | 4 | 5.6% | |
| 2x/day | 62 | 86.1% | |
| 2+ /day | 6 | 8.3% | |
| **Flossing habits** | | | |
| Never | 3 | 4.2% | |
| 1-2x/month | 2 | 2.7% | |
| 1-2x/week | 11 | 15.3% | |
| 3-4x/week | 14 | 19.4% | |
| Every second day | 3 | 4.2% | |
| Every day or more | 39 | 54.2% | |

*SD = standard deviation; values in bold indicate significant p-values (<0.05) as determined by a two-tail t-test: p=0.03 between male and female subjects*

A sample of clinicians and researchers were asked to assign an MGI score to each subject's gingival anterior photograph. Between August to October 2018, 26 reviewers were asked to complete the survey with a response rate of 65%. Of the 17 reviewers that agreed to participate in the MGI survey, five were researchers from various healthcare-

related departments at the University of Alberta. There were six dental hygienists and six dentists enrolled, further outlined in Table 4. It should be noted that researchers were effectively assigned a value of "zero years of clinical practice".

*Table 4: Reviewer demographics by job type, place of education, and years of practice*

| **Occupation** | **Faculty/Department at The University of Alberta** | |
|---|---|---|
| Researcher | Department of Medicine and Dentistry | |
| Researcher | Department of Medicine and Dentistry | |
| Researcher | School of Public Health | |
| Researcher | Department of Psychiatry | |
| Researcher | Faculty of Rehabilitation Medicine | |
| **Occupation** | **Place of Clinical Education** | **Years of Practice** |
| Dentist | University of British Columbia | 6 |
| Dentist | New York University | 7 |
| Dentist | University of Toronto | 11 |
| Dentist | University of Alberta | 17 |
| Dentist | University of Alberta | 20 |
| Dentist | University of Alberta | 26 |
| Dental Hygienist | University of British Columbia | 3 |
| Dental Hygienist | University of British Columbia | 5 |
| Dental Hygienist | University of British Columbia | 6 |
| Dental Hygienist | Regency Dental Hygiene Academy (ON) | 8 |
| Dental Hygienist | University of Alberta | 9 |
| Dental Hygienist | Cambrian College (ON) | 15 |

A cumulative logistic mixed-effects regression model (CLMM) was used to determine if one job type was more or less likely to assign different MGI scores. Table 5 illustrates the results of the CLMM with proportional odds assumption examining differences in MGI score cumulative log odds between different groups of reviewers. An estimate of 1.004 when comparing dentists and dental hygienists indicates that the estimate of the log-odds of a reviewer giving a subject a higher MGI score than any selected cut-off point is 1.004 times higher for hygienists than dentists. If the exponential value of 1.004 is calculated, this suggests that the odds of a hygienist giving a higher MGI score is 2.73 times higher than for a dentist. However, as the p-value is not

significant, it cannot be concluded from the model that hygienists and dentists differ in their score assignment distributions for the same subject.

*Table 5: Cumulative logit mixed effects model to determine whether differences exist between groups of reviewers and effect of years of practice*

| Comparison | Effect estimate | Standard error | p-value |
|---|---|---|---|
| Dentists and researchers | -0.769 | 0.815 | 0.345 |
| Hygienists and researchers | 0.234 | 0.618 | 0.705 |
| Hygienists and dentists | 1.004 | 0.577 | 0.082 |
| Oral health professionals and researchers | 0.156 | 0.669 | 0.816 |
| Test for effect of years of practice:* | 0.032 | 0.043 | 0.456 |

** indicates secondary model using regression coefficient*

Moreover, the model estimates that the cumulative MGI score log-odds is 0.769 lower for a dentist that a researcher. The exponential value suggests the log odds is 0.46 times lower for a dentist assigning a higher MGI score than a researcher; however, the high p-value indicates a high standard error and that it cannot be confidently stated that this effect is statistically different from zero. The results of this CLMM signify that job type does not make a difference in how MGI scores are assigned.

Furthermore, a similar test using the regression coefficient estimated in the CLMM determined the effect of years of practice on assigning MGI scores, resulting in an effect estimate of 0.032 and a p-value of 0.456. This can be interpreted as a 0.032 increase in cumulative log-odds for every one-year increase in years of clinical experience. Similar to the previous results, the p-value from the regression coefficient indicates that there is no association between years of practice of oral health professionals and the assignment of MGI scores.

To determine if the data appropriately fits the CLMM, McFadden's pseudo-$R^2$ was calculated from three regression models, as seen in Table 6. After including random effects (the effects accounting for baseline differences in subjects and reviewers) and fixed effects (job type and years of practice), a McFadden's pseudo-$R^2$ value of 0.260 was calculated.

*Table 6: Cumulative logit mixed effects model with McFadden's pseudo-R2 calculated*

| Model Type | McFadden's pseudo-$R^2$ |
|---|---|
| Random effects + fixed effects | 0.260 |
| Random effects + job type | 0.260 |
| Only random effects | 0.258 |

It was important to include both fixed and random effects in the full model as fixed effects are the variables of primary interest, where random effects control for the baseline differences in MGI scores and address the lack of independence from observations from the same reviewers or subjects. A model with only random intercepts for each subject and reviewer resulted in a McFadden's pseudo-$R^2$ value of 0.258. To examine fixed effects individually, a model with random effects and each separate fixed effect was applied. A model with random effects and job type had a McFadden's pseudo-$R^2$ value of 0.260 an increase from the random effects only model by 0.002. Further including the fixed effect of years of practice, the McFadden's pseudo-$R^2$ value slightly increased to produce a value of 0.260. This suggests that the full model, including random effects and fixed effects, is not a better fit than the intercept model with a value of 0.258 as these two values do not have a statistically significant difference. This further confirms that job type and years of experience are not important in determining subject scores after accounting for random differences between subjects and reviewers. The low pseudo-$R^2$ value of 0.260 for the final model suggests there are other factors that have not

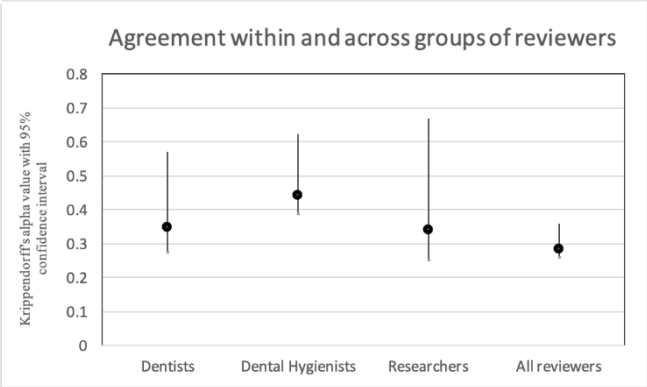
been considered in the model so far as this data was not available. Baseline differences between subjects and reviewers may include the differences in reviewer understanding of the MGI. For example, one reviewer may assign an MGI score of two for one subject where a second reviewer may assign a score of three to the same subject. This difference in score assignment may be due to the perceptions the reviewer has regarding the colour, consistency, and texture of the gingiva. For baseline differences in subjects, these could include variations in actual gingival health such as colour, inflammation, and presence of stippling.

A 95% confidence interval for Krippendorff's alpha was calculated to determine within-group reliability along with reliability across all reviewers. The alpha value calculated between all dentists was 0.347 (CI 0.278, 0.572), between the dental hygienists was 0.442 (CI 0.390, 0.625), and between the researchers was 0.342 (CI 0.255, 0.668). Examining the alpha value across all 17 reviewers to remove the effect of job type was calculated as 0.285 (CI 0.262, 0.360), as displayed in Figure 7. Examination of the data suggest that there is minimal agreement within each group of reviewers as no alpha value was above 0.442, and even less agreement across all reviewers, with an alpha value of 0.285. Moreover, the lower limit of all calculated confidence intervals does not fall above 0.667 and should be rejected, as recommended by Krippendorff (122). These data indicate no agreement across all 17 reviewers and among each individual group of reviewers.

Ordinal alpha was calculated to determine internal consistency. The correlation between items examines the correlation in the scores assigned by reviewers for each subject. The alpha value calculated across all 17 reviewers was calculated at 0.83 (95%

CI 0.67 to 0.87), indicating good internal consistency. This suggests a moderately high correlation between items, suggesting that reviewer scores tend to all be high together or low together for the same subject. However, given Krippendorff's alpha, it cannot be assumed that the scores given to each subject are the same. Good internal consistency suggests that the MGI measures what it purports to measure. In this study, good internal consistency assumes that the MGI is, in fact, measuring gingivitis.

*Figure 7: Reliability as measured by Krippendorff's alpha*




*Circle marker represents an alpha value with the upper and lower endpoints of the 95% confidence interval*

A post hoc analysis was performed to examine whether any of the reviewer job types have higher odds of assigning scores indicating health or inflammation. A mixed logistic regression model with random effects accounting for random effect differences between subjects and reviewers was used; however, scores were now considered binary. These results estimate the effect of job type on the log odds of classifying a subject with inflammation as -0.521 lower for dentists than researchers with a p-value of 0.577, an effect estimate of 0.596 between dental hygienists and researchers with a p-value of 0.407, and an effect estimate of -1.118 between dentists and dental hygienists with a p-value of 0.096. Consistent with the CLMM, the logistic regression was used to determine

that job type does not make a difference when classifying subjects with a healthy score or a score indicating inflammation.

A second mixed logistic regression model was calculated to determine if one reviewer job type has higher odds of assigning scores indicating health or severe inflammation. These results estimate the effect of job type on the log odds of classifying a subject with severe inflammation as 5.238 higher for hygienists than dentists with a p-value of 0.354, an effect estimate of 6.258 higher for dentists than researchers with a p-value of 0.317, and an effect estimate of 4.938 higher for hygienists than researchers with a p-value of 0.835. Similar to the results of the previous mixed logistic regression model, these data indicate that job type does not make a difference in assigning a healthy MGI score versus a score indicating severe inflammation.

A time trend analysis was performed using a CLMM to determine if reviewers were more or less likely to assign high MGI scores as the survey progressed. After accounting for random and fixed effects while assuming proportional odds, the model determined a statistically significant positive time effect of 0.02 (p-value = 0.0012). This suggests that as the images progressed throughout the survey, the cumulative log-odds of a reviewer assigning a higher MGI score increased by 0.02. However, as the reviewers were given the subject images in the exact same order, it cannot be definitively concluded whether this trend of increased likelihood of assigning a higher MGI score corresponds to reviewer fatigue or if there was an objective decrease in the oral health of the subjects in the order the photos were presented.

Overall, the results from these statistical tests indicate that the MGI is not reliable when presented as an image-based survey. Krippendorff states that unreliable data

reduces the chance for results to be valid (122). An instrument is determined to be valid if it measures what it is truly intended to measure (21, 63, 71, 80, 87, 123). However, after determining that the MGI is not a reliable index when presented as an image-based instrument, it cannot be suggested that the MGI is valid either. If there is no agreement in assigning MGI scores among oral health professionals, who should theoretically be able to assign the most accurate score to the true diagnosis, then the reliability and consistency of classifying gingivitis through the MGI is questionable. However, this study has only determined that the MGI is not reliable or valid when subjects are provided via an image-based survey.

# *Chapter 5: Discussion*

This quantitative exploratory study was designed to assess if job type impacts how MGI scores are assigned, the consistency between MGI score assignment within a given profession, and if the length of time in clinical practice impacts MGI score assignment. Reliability of the MGI to assess the presence of gingivitis was analyzed in this study through three different contributing factors: reviewer agreement as a measure of interrater reliability, the use of photographs to generate visual data from which the MGI was determined, and through statistical tests that contribute to a research definition of reliability.

## *5.1 Reviewer Agreement*

The results of the CLMM with proportional odds assumption determined that one group of reviewers is no more or less likely than the other groups of reviewers to assign different MGI scores, indicating that job type does not make a difference in how MGI scores are assigned. In addition, the results of the CLMM indicated no association between years of practice and assignment of MGI scores, suggesting that there is no difference in the length of practice time amongst oral health professionals and assigning MGI scores.

There are several possible reasons to explain why there was such poor agreement between reviewers. Much of the poor agreement may be attributed to a lack of examiner calibration. Ideally, a reviewer is calibrated against a gold standard diagnostic tool; however, after an extensive literature review, no such tool was found for the clinical

diagnosis of gingival diseases. In fact, the literature suggests that "*there is no gold standard or gauge for the clinical assessment of periodontal diseases*" (42). Hefti et al. also mention that in periodontal research settings, reviewers are often trained by very experienced clinical examiners to ensure there is a high degree of calibration amongst the peer group (42). Further, the term calibration is not considered appropriate to use since no gold standard exists, but instead, "reviewer agreement" determines how consistently reviewers assign various scores (42). In this study, efforts were made to provide a degree of agreement amongst reviewers. Reviewers were given the MGI scoring criteria with exemplar photos of each MGI category. The criteria and photos were available to all reviewers throughout the survey.

Despite these instructions, the results suggest little agreement among the reviewers. While this could be attributed to the inclusion of non-oral health trained professionals (non-clinical researchers), there was minimal agreement among each job type of reviewers, including oral health professionals. Despite an insignificant p-value, these data imply a trend towards the dentists and dental hygienists assigning different MGI scores. This suggests that a future study examining the agreement between an increased sample size of dentists and dental hygienists in assigning MGI scores could be warranted.

One possibility that may explain the variation in MGI scoring among reviewers may be the lack of familiarity of anatomical knowledge of the reviewers. The instructions provided to all of the reviewers did not define the anatomical margins of the different tissue types that exist in the oral cavity. It is speculated that some reviewers may have included parts of the alveolar mucosa in their assessment instead of focusing solely on the

free and attached gingiva. The oral health professionals should have been able to understand the different parts of the oral mucosa including the free gingiva, attached gingiva, and alveolar mucosa; however, some may not have been able to identify the margins separating the different tissue types on the images. In fact, Custers suggests in a medical education systematic review that after a one-year recall interval, one-third of the subject's anatomical knowledge is lost, further declining to below 50% the following year (124). Although this could be a possible explanation, dental education focuses on the anatomy of the oral health in great detail and it is unlikely that this knowledge would be lost so quickly after graduation.

The disagreement across oral health professionals in classifying the MGI was an unexpected result of the study. It was anticipated that a general agreement of MGI score application amongst the oral health professional group would provide a baseline that could be a benchmark that the researchers could have been compared against. The oral health professional group are specifically trained to examine and diagnose the gingiva based on the colour, consistency, and texture in everyday practice. However, Lanning et al. report inaccuracy and variability among dentists and dental hygienists when examining subjects presented in web-based cases (125). Currently, there is no gold standard in diagnosing gingival health (42); however, oral health practitioners often rely on bleeding on probing as the benchmark to indicate gingival health (126). The lack of alignment across these individuals may support that visual means to diagnose gingival health are insufficient. Baumgartner suggests that colour alone may be used to help identify the presence of gingivitis, but is not sufficient enough to diagnose the severity of inflammation (10). The literature suggests that there is disagreement over what the gold

standard of diagnosing gingivitis should be (42, 127). The results presented in this study align with previous suggestions (10, 38, 42, 127) that a component of instrumentation to determine if bleeding is present is critical to diagnosing gingival disease. A photo-based examination format would not be sufficient in the development of a gold standard to diagnose gingival health.

In this study, reviewers were required to make this assessment based on a series of images that were provided in the survey. It has been observed in other areas of clinical diagnostics that the visual identification of health-related issues can be challenging when based on review of images. For example, in a study by Manning et al., it was observed that complex visual images make it difficult for reviewers to discriminate between normal anatomical structures and abnormalities. In the study, reviewers were unable to correctly identify cancer lesions in 27.2% of postero-anterior chest radiographs, even when the features had been made visually obvious by the imaging process (128). The study found that "*even experienced reviewers may not always register visually obvious lesions"* (128). Similarly, Dallas et al. found that though reviewers correctly identified lesions in fundus photographs of the eyes, they had distinct discrepancies in identifying both the severity and borders of the lesion (129). These examples suggest that, when image review is used as a diagnostic tool, there may be two sources of variability among reviewers: 1) the ability to correctly identify abnormalities in patients when present, and 2) assessing the severity of the issue. In this study, it may follow that reviewers were: 1) unable to correctly identify inflammation in patients when present, and 2) if inflammation was correctly identified, following up with an appropriate scoring of the severity of inflammation.

Further disagreement among oral health professionals may be due to the differences in clinical decision making used by novice and experienced clinicians. Grant et al. suggest that there may be differences in the way that medical clinicians with differing years of experience diagnose various clinical features (130). The concept of clinical reasoning, also known as clinical problem solving or diagnostic problem solving, examines how individuals use problem solving skills and categorization processes to arrive at a diagnosis (131). Gruppen et al. explain that the process of clinical reasoning starts with combining information from the patient and/or situation with the clinician's prior knowledge to form an initial assessment (131). This assessment forms the foundation for the rest of the reasoning process and as such, a weak initial assessment is a barrier when formulating a diagnosis. Following the initial assessment, a rapid evaluation is performed at an unconscious level to determine if this assessment is appropriate and fits with the available information from the patient or situation. Categorization processes are then used to organize the information into whether it is considered relevant or irrelevant. Mandler et al. suggests that experts can easily ignore irrelevant information while novices tend to be more easily distracted by this information Contrary to what the literature suggests around different approaches to decision-making skills in novice and experienced clinicians, the results from this study indicate that it was unlikely the decision making processes played a role in the assignment of MGI scores. The results of the CLMM suggest no association between years of practice of oral health professionals and the assignment of MGI scores. The literature indicates that there are different processes used when it comes to decision making and that more experienced clinicians will formulate a diagnosis in a different manner than novice clinicians (130, 131).

Regardless of the mental processes that the clinician goes through to assign a score, whether novice or experienced, the results of this study show that these processes do not seem to affect how reviewers have assigned MGI scores.

Reviewer agreement is an important component of determining the validity and reliability of an index. This ensures that reproducible measurements can be obtained whether by the same reviewer at different time points or by different reviewers at the same time (42). Poor agreement between reviewers can result in inaccurate and unreliable data which can impact the conclusions drawn from a research study (132). The results of this study suggest poor consistency within each group of reviewers as well as across all 17 reviewers. This inconsistency raises the following questions: 1) were the instructions provided to the reviewers prior to the start of the survey adequate to achieve acceptable reviewer calibration?, 2) are the inconsistencies due to the challenges of using images for assessments?, and/or 3) is the MGI an unreliable index to use when classifying gingival diseases when presented in a photo-based format?

## *5.2 Photographic Assessments*

This study employs the use of photographs to present the gingiva of subjects with varying degrees of inflammation for reviewers to classify the levels of gingivitis. Practically, using a digital imaging system to assess the health of the gingiva is advantageous as it allows for an easy method of evaluation that can be performed by multiple reviewers at several different time points without the subject having to be present. Further, the risk of examiner bias is reduced as the subject in the photograph is

blinded to the reviewers as opposed to an in-person clinical examination. In addition, a permanent database can be created for future research studies.

However, a digital imaging system to assess the health of the gingiva can introduce some bias over in-person clinical examinations. When examining a subject in person, the reviewer has the ability to move positions to view the gingiva from multiple angles as well examine the lingual and posterior surfaces. These surfaces are challenging to photograph and typically, the anterior region is the most commonly photographed area to classify the health of the gingiva from a photograph (28, 29, 31, 133). Furthermore, when performing an in-person clinical examination, the reviewer is able to get a sense of the overall health of the subject including plaque and saliva levels that may not be visible in a photograph. Large amounts of plaque may encourage the reviewer to assume a poorer oral health status and may result in the reviewer assigning a more severe MGI score (134). Similarly, if the subject presents with xerostomia, the reviewer may assume that the lack of saliva is encouraging gingival inflammation and again, assign a higher MGI score (135).

A study by Smith et al. used an image analysis system where anterior intra-oral photographs were taken using a frame, lighting, and head-fixing apparatus where the digital camera was affixed and the subject's occlusal plane was aligned in a reproducible manner for each subject. The researchers used a plaque disclosing solution on the upper and lower incisors, followed by capturing an intraoral photograph, and measuring the plaque area on a computer software program. This was compared to in-person plaque scores measured by two reviewers using two separate plaque indices. Smith et al. determined the image analysis method was more sensitive in measuring plaque levels

when compared to in-person traditional plaque measurement indices (24). After validating the use of the image analysis technique, Smith et al. investigated the reliability of measuring gingival inflammation through the aforementioned photographic method (25). Smith et al. used Fleiss' coefficient of reliability to calculate a high level of interrater reliability when compared to clinical examinations and suggested that image analysis is a reliable method of determining gingival inflammation. However, using Krippendorff's alpha to analyze interrater reliability in the present study was found to be poor. Both Fleiss' coefficient and Krippendorff's alpha can be used to measure interrater reliability, however, Fleiss' coefficient is more appropriate for nominal data, where Krippendorff's is better suited for ordinal data (136). While a similar approach to generating digital images for assessment was used in this study, the poor reviewer agreement is inconsistent with what Smith et al. determined. It is uncertain whether photographic presentation of gingival inflammation is reliable when classifying gingivitis by use of the MGI, or if there are issues with the study design for this type of assessment and further research could be completed to determine the underlying causes.

There are technical considerations for digital image generation for use in assessing gingival health, including calibration for alignment, colour, and light. Smith et al. developed an apparatus that held the digital camera and allowed for the same angulation of the subject's head every time it was used. The images that were taken in this study were not adjusted to account for angulation errors, however, the gridlines on the LCD screen of the EyeSpecial C-II camera allowed for the proper alignment of the midline and occlusal plane in each photo. In the Smith et al. study, digital images of the gingiva were calibrated for colour against a common element. Colour is a determining

factor during the application of the MGI score (41). It was assumed during the study design that there was no need to additionally calibrate the colour of the photographs, as the EyeSpecial C-II Camera has an automatically calibrated flash that controls for lighting conditions from subject to subject. However, as the photos were not specifically calibrated against a common element there was no way to verify that the colour was completely calibrated from one image to the next. Martins et al. and Cochran et al. both suggest that images that have been taken using improper lighting or angulation may result in lip shadowing and can impact the appearance of the gingiva (22, 28). The use of a ring flash in this study was intended to avoid issues with lip shadowing; however, Cochran et al. postulate that the use of a ring flash may result in a greater area of the gingiva impacted by specular reflection (28). This may have caused the gingiva to have a reflective appearance, thereby making the identification of the colour or texture of the gingiva difficult to determine.

In addition to techniques used to generate a photograph, how the subject's gingiva is prepared in advance of the photograph (i.e. drying the gingiva) can also generate discrepancies when the images are subsequently used to diagnose oral health conditions. In three related studies using images and clinical examinations to diagnose fluorosis, there was disagreement about whether the use of photographs resulted in over- or under-diagnosis when compared to clinical exams. Martins et al., Wong et al., and Cruz-Orcutt et al. assessed examiner reliability by comparing clinical and photographic examinations to detect fluorosis in children (22, 26, 29). Both Wong et al. and Cruz-Orcutt et al. concluded that there was a significantly higher prevalence of diagnosing fluorosis using photographs than when compared to clinical exams. Note that for a fluorosis diagnosis, a

direct observation clinical exam with a dental mirror, light, and gauze to dry the teeth is the de facto gold standard (22, 28). Contrary to these results, Martins et al. found that clinical exams resulted in diagnosing a higher number of cases with fluorosis. Discrepancies between these studies may be due to the preparation of the subject when photographing the teeth. Martins et al. dried the teeth with gauze for the clinical examinations but let the teeth dry naturally to take the photographs (22). Wong et al. left the teeth wet while performing both the clinical examinations and while taking the photographs (26), while Cruz-Orcutt et al. reported minimal drying when performing the clinical exam, but dried the teeth prior to taking intraoral photographs (29). The diagnosis of fluorosis may differ based on the appearance of the teeth with a drier environment showing more detailed features and more moist environment masking some of the subtleties of fluorosis (27, 29). In this study, it is possible that the photographic presentation may have resulted in conflicting diagnoses of severity of gingival inflammation. The reviewers may have misclassified the severity of the gingival inflammation due to factors such as dryness of the gingiva in the images. No specific measures were designed in this study to dry the gingiva, however, no controls were implemented to ensure consistent levels of dryness. Consistently dried gingiva has been shown to help identify the texture of the attached gingiva as texture is a key identifier of healthy gingiva (137).

Despite the widespread use of photographs for diagnostic purposes (22-25, 27-30, 129, 133), there is varying evidence to indicate the efficacy when compared to clinical examinations (22, 26). The findings from this study indicate that the MGI is a challenging instrument to use when reviewers are presented with photographs of subjects with

varying degrees of gingivitis. Discrepancies between this study and previous studies that have had success with the use of digital images as a diagnostic tool may be due to design elements in this project. Previous studies that successfully used images for diagnosis typically only employed one to two reviewers to rate the subjects in either clinical settings or photographic presentations. This research study recruited 17 reviewers to assign MGI scores to 72 different subjects. A comprehensive literature search of past research studies that utilize intra-oral photographic analysis for the diagnosis of various dental conditions do not generally include more than seven examiners, and typically only use one or two examiners (22-31, 105, 133, 138, 139). Boyer et al. suggest that using a single reviewer increases the risk of subjective bias due to the individual's own perspective. This study included a much higher number of reviewers than is typically seen in previous studies with the assumption that including multiple reviewers in a study provides a higher degree of investigator triangulation and can enhance the conclusions stated (140, 141). Investigator triangulation involves recruiting two or more reviewers to provide several observations thereby increasing the validity of study findings (141). Despite the limitations with the instructions provided to reviewers and techniques used to generate a photograph, the number of reviewers used in this study should lead to a higher degree of methodological rigour. This suggests that it is possible that the MGI is an unreliable indicator to assess the presence of gingivitis when subjects are presented in an image-based survey.

## *5.3 Statistical Tests to Assess Reliability of an Indicator*

This study found that when reviewers assess the MGI based on an image-based survey, it cannot be determined to be reliable or valid. Reliability is composed of the following three concepts: interrater reliability as discussed above, internal consistency, and stability. In this study, interrater reliability and internal consistency were measured; however, stability was not measured as test-retest and intrarater reliability measures were not performed as this study examined the one-time application of the MGI in an image-based analysis. At this time, it is unknown what the consistency of the MGI over time is and this may impact the internal validity of the index (63).

Ordinal alpha was used to measure internal consistency. The reviewers' responses were determined to have high correlation between the scores given by reviewers for the same subject with an alpha value of 0.83. This suggests that the MGI is presumably measuring the same, single underlying factor (i.e. gingivitis); however, the low Krippendorff's alpha does not indicate reviewer agreement. Further, the literature states that internal consistency is sensitive to the number of items in an index (142), and in this study, there were over 1200 data points, suggesting further that the ordinal alpha value may have been affected by this large number. The high ordinal alpha suggests that MGI is actually measuring gingivitis, low agreement notwithstanding.

## *5.4 Validity of the MGI*

While reliability measures consistency and repeatability of an index, validity determines if the index truly yields what it is intended to measure (72, 79). Validity is evaluated by various approaches including criterion, content, and construct validity,

and/or by employing different frameworks including Messick's unified model or Kane's argument-based approach to validation (70, 80, 87, 94). These two frameworks together form a robust approach to assess validity and are used frequently by educational, psychological, and measurement practitioners (95-97).

The first aspect of Messick's framework was not helpful in this study to assess the validity of the MGI as an index. Assessments of criterion validity, content validity, and construct validity do not support the use of the MGI for diagnosing gingivitis from image-based surveys. As this study did not have a verified gingival diagnosis for each subject, the degree of criterion validity could not be assessed as this measure compares an index against a gold standard (21). Currently, the gold standard for diagnosing the presence of periodontitis is the use of periodontal probing to identify the presence or absence of bone loss; however, there is no agreed upon gold standard for the diagnosis of gingivitis (9, 127). In spite of this, bleeding on probing is considered to be the most objective sign of inflammation and is widely used in clinical settings for the diagnosis of gingival inflammation (38, 143, 144). However, as the MGI does not assess the health of the gingiva based on the presence of bleeding, criterion validity cannot be used.

Content validity is intended to assess that the construct of interest is actually the true construct being measured (21). In this study, the construct of interest is gingival inflammation. The MGI was validated as an index in 1989 in a study that included subjects with both Types I and II periodontal disease which included patients with periodontitis (47). While the inclusion of both subjects with gingivitis and periodontitis could give rise to the perception that the index lacks content validity (71), inflammation can be present in both types of subjects and it should be possible to apply the MGI in

both cases. Content validity can also be measured by internal consistency and, as discussed above, this study was found to have good internal consistency.

Construct validity is a measure that is established over time (71, 95). It involves an ongoing process of developing and testing predictions that contribute to understanding the construct of interest (92). Construct validity is challenging to apply in this study due to the requirement to gather observations and evidence over several years. ~~survey.~~ Applying the index to images to assess gingivitis may exclude relevant information that is present in the clinical setting, such as bleeding on probing (38). Continued discussion in the literature regarding the lack of a gold standard in assessing gingivitis continues to be a challenge to the construct validity of the MGI. This study may be the start of the inquiry to develop or refute construct validity of the MGI when used to evaluate the gingiva in a photo-based survey format.

The second aspect of Messick's framework, internal structure, includes interrater reliability and internal consistency (95). Krippendorff's alpha measured interrater reliability and was found to be poor between each group and among all 17 reviewers. Inadequate interrater reliability indicates that there is no consensus among the scores assigned by reviewers and the disagreement in scoring does not support the foundation for a valid and reliable index (145). Conversely, internal consistency was measured by the ordinal alpha and was found to be high (0.83). Scholtes et al. suggests that good internal consistency indicates that the items in the index are sufficiently correlated (71). For image-based analysis, the evidence from this study suggests that the MGI appropriately measures gingivitis as seen by scores that are consistently assigned within a

similar range for each subject; however, this does not indicate that the reviewers agree upon the scores assigned.

Messick's third source of evidence to establish/support validity examines relationships with other variables. This concept compares the MGI with other indices, such as the PMA, that also measure severity of gingival inflammation. The literature review completed for this thesis found no studies that examine interrater reliability or internal consistency of the MGI. However, there have been two studies compared the MGI to other gingival indices, both of which have been performed by Lobene et al. (41, 47). The authors report that the MGI is a valid index when it is compared against the GI in determining the severity of gingival inflammation (47). The GI uses a bleeding component to determine gingival inflammation (43), as bleeding on instrumentation is considered a reliable method of determining gingivitis (38, 143, 144). Lobene's et al. assessment that the MGI is a valid index contradict the results of this current study since Lobene et al. determined that the MGI is valid in classifying gingivitis. The disagreement between this study and Lobene's et al. studies, in regards to validity, could be due to the difference in the numbers of reviewers that were involved. Lobene et al. used a single examiner to assign an MGI during a clinical examination, while this study used 17 reviewers to assign an MGI using image-based surveys. Interrater reliability cannot be established for Lobene's et al. studies because there is only one examiner. Boyer suggests that "*survey-based research that relies on a single respondent may be biased because that respondent may potentially present a skewed or inaccurate view*" (105). By using 17 reviewers, the present study examined the consistency of reviewers in assigning MGI scores; however, interrater reliability was found to be poor indicating no agreement

among reviewers and as such, suggesting that the MGI may not be valid in classifying gingivitis.

The fourth source of validity evidence in Messick's framework is the response process which may be impacted by rater training (95). Proper training of the reviewers can affect the quality of the data received. While guidance was provided to define the ordinal values of the MGI and exemplar photos were given, it is not clear if the instructions were sufficient to influence the quality of data received. The response process cannot be assessed for the MGI based on the results of this study.

The fifth source of validity evidence in Messick's framework is the impact of consequences. This can be a difficult concept to measure as it examines the impact of the index, whether negative or positive, and to formulate a validity argument (95). Impact of consequences was not assessed in this study, nor have assessments of impacts of consequences been observed in the comprehensive literature review that was done. Lobene et al. list the following four positive impacts of consequences when using the MGI in a research setting: 1) the non-invasive nature of the index eliminates trauma to the soft tissues; 2) the MGI is logistically simpler since bleeding does not obscure the visual aspects of inflammation; 3) less variability in scores assigned due to elimination of the bleeding component; and 4) the MGI has greater sensitivity due to expanding the index to include localized and generalized variations of mild gingivitis. The use of a visual technique compared to an invasive one is suggested to be advantageous when applied in certain research settings such as clinical trials (47); however, it is unknown what the impact of consequences is in this research study. Using Messick's framework to assess the validity of the MGI it was shown that the lack of content validity, the challenge

of providing construct validity, and the poor interrater reliability indicate that the MGI does not appear to be an effective tool to classify gingivitis based on examining intraoral photographs.

Another common framework used to assess validity is Kane's argument-based approach. Kane suggests researchers gather evidence across four key inferences, when appropriate, within the validity argument. Kane labels these four inferences "scoring", "generalization", "extrapolation", and "implications" (80). This study assessed "scoring" as a form of validity in relation to the MGI. The "scoring" inference is similar to Messick's response process and can be related to rater training. Future studies could examine whether better reviewer training could result in improved reviewer agreement.

The "generalization" inference typically applies to newly developed indices as it examines if the index can be applied in a "real-world setting" (80). "Generalization" for indices is assessed by interviewing experts or reviewing previous literature. In this study, a review of previous literature was carried out to establish what researchers perceive as strengths or weaknesses of the application of the MGI.

The "extrapolation" inference examines whether the index can relate to a real-world setting though the MGI is not frequently used in private practice to assess gingivitis (41, 47). Based on the results of this study, it is not clear if the MGI is an effective tool to diagnose gingivitis when used in clinical research settings and extrapolation to private practice may not be valid when examined as intraoral photographs.

Finally, Kane's "implications" inference examines the decisions that researchers make after having applied an index and generated information. The information

generated by applying the MGI should allow practitioners to assess gingivitis in comparative studies. Clinical researchers will determine the efficacy of various interventions based on the application of the MGI (15, 16, 18, 34, 103, 104).

This study indicates that the use of the MGI could potentially lead to inconclusive results in comparative clinical studies. The results of this study and the high degree of variability between reviewers suggest it would be difficult to make informed decisions and take meaningful action based on the assessment generated using the MGI as an image-based survey. Despite the good internal consistency as determined by Zumbo's ordinal alpha suggesting that the MGI is, in fact, measuring gingivitis, the lack of robust support of criterion, content, and construct validity as well as from Kane and Messick's frameworks suggest that overall, the MGI is not a valid index when classifying gingival inflammation through a photo-based format.

### *5.5 Implications of using an Unreliable Index*

This study indicates that the MGI is an unreliable index with low validity for use in examining anterior photographs of varying levels of gingival inflammation. Using an unreliable index can have serious implications for research studies, as well as clinical practise. Marshall et al. examined 205 randomized controlled trials involving patients with schizophrenia and determined that studies were 36% more likely to report treatment that was superior than compared to the control treatment when using an unpublished index instead of ones with peer-reviewed evidence of reliability (37). An unpublished index may indicate that the results obtained from the index were insignificant or used in a

poor-quality trial. In Marshall's et al. study, the impact of using unreliable indices may result in patients with schizophrenia receiving inappropriate treatment.

In clinical studies, an unreliable gingival index may create false positives or negatives where a subject is incorrectly diagnosed to have or not have gingivitis and therefore the results of the study will be skewed. A recent study by Lynch et al. used one examiner to assess levels of gingival inflammation using the MGI before and after implementing an alcohol-free mouthwash rinse program compared to an alcohol-containing mouthwash (18). The results of this Lynch's et al. study suggest no difference between either mouthwash types when examining the appearance of the gingiva after the introduction of the rinses into daily oral health routines. However, given that the present study fails to demonstrate reliability and validity of the MGI, , it is unclear if the conclusions drawn by Lynch et al. can be substantiated as the levels of gingival inflammation described by the MGI (18). This can only be speculated, however, as the present study examined the health of the gingiva based on photographs, while Lynch et al. performed clinical examinations.

No studies have used the MGI to classify the gingival health using photographs at this time. It is important to assess the validity and reliability of the MGI for image-based surveys as this type of assessment is becoming increasingly common (146, 147). Telemedicine allows patients in remote areas the ability to have specialized medical care without having the difficulty of travel or burden of cost to get to the nearest medical centre (148). Torres-Pereira et al. and Totty et al. suggest photo-based consultations via a digital platform are a reliable method to diagnose oral lesions and follow-up with post-operative surgeries, respectively (146, 147). Based on the evidence collected in this

study, the MGI may not be a reliable index to use when classifying gingival inflammation based on a photo-based survey.

### *5.6 Sensitivity and Specificity of the MGI*

Sensitivity is the ability to detect positive cases without registering a false negative (53). Conversely, specificity refers to the extent that a test could detect negative cases without registering a false positive (53). Lobene et al. modified the GI to increase the sensitivity in the lower region of the index by adding an additional score to divide mild gingivitis into localized or generalized (41). However, the conclusions drawn from this study found poor agreement between reviewers indicating that this modification by Lobene may not, in fact, increase the sensitivity of the MGI.

A post-hoc analysis was performed to determine if splitting the MGI into a binary index to classify healthy gingiva from inflamed gingiva would increase the sensitivity of the index. The logistic model suggested that there was poor agreement among reviewers as no job type was more or less likely to classify a subject as having healthy or inflamed gingiva. However, as there was not a baseline value to compare the subjects score's against, it was unable to be determined if the MGI is useful in correctly identifying gingivitis when photographs are used to classify varying degrees of inflammation. These results suggest that the oral health professionals and non-clinical researchers were unable to detect significant changes in the gingiva to classify the subject as healthy or inflamed. Furthermore, a second mixed logistic regression model was calculated to determine that not one specific group of reviewers was more or less likely to assign scores indicating health or severe inflammation. These results suggest that reviewers were unable to detect

changes between healthy and severe inflammation indicating that the modification from Lobene et al. to increase the sensitivity of the index is not effective. Further, the classifications of the MGI may not be necessary as the results from this study determined that classifying gingivitis into different categories of inflammation is not able to be reliably determined.

## *5.7 Sources of Error*

A measure's validity can be affected by the measurement errors that may occur in research studies. These can be divided into two classes: systematic error and random error (93, 149). Systematic error occurs when a measure produces data that consistently differs from the true score and is typically seen as over- or under-estimated data (149). Random error occurs when factors affect data in a random way and cause deviations from the true score with no consistent pattern (149).

In this study, systematic error may have occurred due to improper calibration of the EyeSpecial C-II camera resulting in improper alignment or lighting conditions. However, as systematic error skews the data consistently in one direction, then all subjects photographed would be subject to this systematic error

Random error also decreases validity as the scores produced are inconsistent. Random error can result from a reviewer's mood that impacts their performance in assigning scores (93). Individual inconsistencies that may have occurred in this study could be due to carelessness, difficulties in understand the instructions, or challenges in accessing the survey due to computer unfamiliarity or internet access (93, 150). Bowling reports acquiescence bias can occur in electronic-based surveys as respondents tend to

select the answer closest to the question (150). To control for this, Bowling suggests occasionally rearranging the order of responses; however, in this study, this may result in response error if the MGI values are not placed in order but it is assumed they are ordered numerically (150). Furthermore, random error may have occurred in the colour calibration of different digital devices (150). The non-linear value, known as image gamma, can vary between different computer hardware and impacts the intensity of the pixels (151). Calibration to adjust the gamma value can be performed to ensure that all photographs are seen under similar conditions; however, this was not done in this research study. Moreover, reviewer fatigue may have played a role in this study. The literature defines reviewer fatigue as a situation where reviewers give less thoughtful answers due to the length of a survey or complexity of answering the survey questions (152). Reviewer fatigue cannot be examined in this study as the images were not administered randomly. Therefore, it cannot be determined if the reviewers were fatiguing near the end of the survey.

Statistical tests can account for some of the random error by adding variables into a regression equation to explain some of the variation in the model that cannot be explained by the existing variables (93, 149). In this study, this was addressed using the CLMM which included fixed effects (job type and years of practice) and random effects (baseline differences in subjects and reviewers).

## *5.8 Limitations*

There are a number of limitations to consider when interpreting the results of this research study. First, the reviewers were selected by convenience sampling. There may

be possible bias within the group of reviewers as this may not be a proper representation of the general population which could lead to difficulty in making generalized conclusions about the research (153).

Though care was taken to ensure photographs were taken similarly in terms of exposure and focus, the focal distance or angles were not strictly standardized. This could affect colour, light, and alignment which may impact the reviewer's ability to assess gingivitis. Smith et al. reports calibrating the images by placing a red articulating paper disc on the central incisor of each subject so the mean red pixel value of the disc can be compared against the mean red pixel value of the gingiva when analyzed digitally (25). As previously mentioned, the EyeSpecial C-II camera did not need to be calibrated due to the use of the ring flash and gridlines on the LED screen.

Stability, an important component of reliability, was not measured as the reviewers enrolled in this study did not participate in the survey a second time and scores were not obtained by a second set of reviewers. This does not allow for intrarater reliability or test-retest reliability to be measured, respectively. . Further, by evaluating test-retest reliability, a firmer conclusion can be made towards the reliability of the MGI as this can examine the repeatability of the MGI.. As neither of these measures were evaluated, it cannot be determined if the MGI is repeatable in an image-based survey format

## *5.9 Impact in Research and Clinical Practice*

Currently there is no agreed upon gold standard for the diagnosis of gingivitis, however, the most commonly used objective measure is bleeding upon probing (38, 154).

The MGI does not incorporate probing and Lobene et al. determined that a visual examination of the gingiva was as reliable as including a bleeding component, as when compared to the GI. However, the data gathered from this study suggest that visual examination of the gingiva when presented in a photo-based survey format is not adequate to reliably classify the health of the gingiva. As discussed, there may be differences when examining the gingiva in-person versus a photograph, however, this study supports the use of eliciting bleeding to diagnose gingival inflammation.

The results of this study suggest that the visual examination of the gingiva, including assessing for colour, contour, texture, and consistency changes, may not be critical in the diagnosis of gingivitis. Prior to 2017, the American Academy of Periodontology (AAP) diagnosed gingivitis by identifying the severity of the inflammation and whether this inflammation presented as localized or generalized (108). The severity of gingivitis included classifying the inflammation into mild, moderate, and severe, however, the AAP stated that these designations were based on subjective clinical assessments including the identification of bleeding, redness, texture and consistency changes (108). The most recent amendments to the AAP classification of gingivitis includes removal of the mild, moderate, and severe categories (155). The current version states, "*there is utility in defining the severity of gingivitis as a patient communication tool, but there is no objective clinical criteria for defining severity*" (155). Further, "*there is no robust evidence to clearly differentiate mild, moderate, and severe gingivitis, and definitions remain a matter of professional opinion"* (155). As such, the present study is supported by the current AAP modifications suggesting that gingivitis is challenging to differentiate between mild, moderate, and severe. Trombelli et al. provided a consensus

report for the current AAP classifications suggesting that the presence of gingival inflammation can be objectively and accurately assessed using a bleeding on probing score (154). Gingival bleeding can differentiate healthy from inflamed gingiva and can be further classified as localized or generalized (38, 154). As the MGI does not incorporate a bleeding component, this study did not use bleeding on probing to identify gingival inflammation. The results of this study suggest that solely using a visual examination to assess the health of the gingiva is not reliable and that using only a bleeding component may be useful in diagnosing gingival health.

## *5.10 Recommendations and Future Directions*

Based on the conclusions of this study, the evidence does not support the use of the MGI to evaluate the gingival health of subjects presented in an image-based survey format. In order to confidently conclude if the MGI is, in fact, an unreliable and invalid instrument to use with this method or if it is due to the design of the study, aspects of the experimental design should be modified.

In the future, if this study were to be performed again, it would be beneficial to enroll subjects from a broader portion of the population, to randomly recruit reviewers, to use a color calibration disc when taking digital photographs, and to examine test-retest reliability and intrarater reliability. Moreover, drying the gingiva to ensure similar saliva-free environments in all subjects would help with calibrating the photos from subject to subject. Furthermore, the objective clinical measure of bleeding upon probing should be recorded so that the subjects' MGI scores can be compared against it to provide evidence for criterion validity (21).

Creating a training program, such as a webinar, for the reviewers to participate in before taking the survey may assist with reviewer alignment and robustness of this study. Sadler et al. recommend providing a tutorial and quiz before the survey to test the knowledge of the reviewers (156). In the initial clinical study performed by Lobene et al., two reviewers were calibrated against an expert reviewer (41). If an assigned MGI score was not agreed upon, the subject was re-examined by all three reviewers until there was agreement on the MGI score and the extent of the inflammation (41).

The information provided by this work will contribute to future research when using the MGI in an image-based format to ensure adequate validity and reliability of the index. The conclusions of this study are supported by the current changes to the AAP classifications by removing the classifications of gingival health and relying on a bleeding component.

## *5.11 Conclusion*

Multiple research studies have used the MGI over the past three decades (15, 16, 18, 32, 34, 35, 101-104). The results of this present study add to the existing body of knowledge on the reliability and validity of using the MGI when presented in an image-based format. Furthermore, the conclusions presented in this study contribute to the understanding of the importance of including a bleeding component when assessing the health of the gingiva.

As there was no agreement within or between the groups of dentists and dental hygienists, these scores were not able to be established as the baseline value for each subject's photo. Because of this, the non-clinical researchers were not able to be

compared against the oral health professionals. Furthermore, when job type was ignored, there was still no agreement across all 17 reviewers, suggesting poor interrater reliability. Moreover, years of experience did not change the way oral health professionals assigned MGI scores and there was no difference in the way job types classified subjects between healthy and inflamed gingiva or between healthy and severely inflamed gingiva if the ordinal values of the MGI were reduced to binary values.

This appears to be the first study to examine the reliability and validity of the MGI when reviewers were presented photos of subjects with varying degrees of gingivitis for examination. The findings from this study determined that the MGI is not a reliable or valid instrument when used in this specific format and that reviewers do not agree with one another when classifying the severity of gingival inflammation. Based on this conclusion, using the visual aspects of the MGI may not be sufficient in future clinical studies and that a bleeding component should be included to accurately diagnose the health of the gingiva.

## *References*


1. Carvajal P, Gomez M, Gomes S, Costa R, Toledo A, Solanes F, et al. Prevalence, severity, and risk indicators of gingival inflammation in a multi-center study on South American adults: a cross sectional study. J Appl Oral Sci. 2016;24(5):524-34.

2. Diefenderfer K, Ahlf R, Simecek J, Levine M. Periodontal health status in a cohort of young US Navy personnel. J Public Health Dent. 2007;67:49-54.

3. Li Y, Lee S, Hujoel P, Su M, Zhang W, Kim J, et al. Prevalence and severity of gingivitis in American adults. Am J Dent. 2010;23(1):9-13.

4. Stamm J. Epidemiology of gingivitis. J Clin Periodontol. 1986;13:360-70.

5. Belstrom D, Damgaard C, Kononen E, Holmstrip P, Gursoy U. Salivary cytokine levels in early gingival inflammation. J Oral Microbiol. 2017;9(1):1364101.

6. Eberhard J, Grote K, Luchtefeld M, Heuer W, Scheuett H, Divchev D, et al. Experimental gingivitis induces systemic inflammatory markers in young healthy individuals: a single-subject interventional study. PLoS ONE. 2013;9(2):e55265.

7. Grellman A, Zanatta F. Diagnosis of gingivitis: state of the art. J Dent & Oral Disord. 2016;2(3):1017.

8. Mariotti A. Dental plaque-induced gingival diseases. Ann Periodontol. 1999;4(1):7-17.

9. Highfield J. Diagnosis and classification of periodontal disease. Aus Dent J. 2009;54(1):S11-26.

10. Baumgartner W, Weis R, Reyher J. The diagnostic value of redness in gingivitis. J Periodontol. 1966;37(4):294-7.

11. Crowley R. A method for evaluating consistency in diagnosis of gingivitis. Oral Surg Oral Med Oral Pathol Oral Radiol. 1955;8(11):1128-33.

12. Greene A. A study of the characteristics of stippling and its relation to gingival health. J Periodontol. 1962;33(2):176-82.

13. Muhlemann H, Son S. Gingival sulcus bleeding - a leading symptom in initial gingivitis. Helv Odontol Acta. 1971;15(2):107-13.

14. Ata-Ali J, Flichy-Fenandez A, Alegre-Domingo T, Ata-Ali F, Palacio J, Penarrocha-Diago M. Clinical, microbiological, and immunological aspects of healthy versus peri-implantitis tissue in full arch reconstruction patients: a prospective cross-sectional study. BMC Oral Health. 2015;15(43):1-10.

15. Herrera D, Escudero N, Perez L, Otheo M, Canete-Sanchez E, Perez T, et al. Clinical and microbiological effects of the use of a cetylpyridinium chloride dentifrice and mouth rinse in orthodontic patients: a 3-month randomized clinical trial. Eur J Orthod. 2018;40(5):465-74.

16. Lamster I, Alfano M, Seiger M, Gordon J. The effect of Listerine antiseptic on reduction of existing plaque and gingivitis. Clin Prev Dent. 1983;5:12-6.

17. Loesche W. Clinical and microbiological aspects of chemotherapeutic agents used according to the specific plaque hypothesis. J Dent Res. 1979;58(12):2404-12.

18. Lynch M, Cortelli S, McGuire J, Zhang J, Ricci-Nittel D, Mordas C, et al. The effects of essential oil mouthrinses with or without alcohol on plaque and gingivitis: a randomized controlled clinical study BMC Oral Health. 2018;18(6).

19. Zare D, Haerian A, Molla R, Vaziri F. Evaluation of the effects of diode (980Nm) laser on gingival inflammation after nonsurgical periodontal therapy. J Lasers Med Sci. 2014;5(1):27-31.

20. Russell A. Indices for recording periodontal disease. Working paper WHO/DH/33. Geneva to Periodontal Disease WHO Tech. 1960;Rep(Series No 207):1961 WHO Geneva.

21. Bannigan K, Watson R. Reliability and validity in a nutshell. JCN. 2009;18:3237-43.

22. Martins C, Chalub L, Lima-Arsati Y, Pordeus I, Paiva S. Agreement in the diagnosis of dental fluorosis in central incisors performed by a standardized photographic method and clinical examination. Cad Saude Publica. 2009;25(5):1017-24.

23. Moncada G, Silva F, Angel P, Oliveira O, Fresno M, Cisternas P, et al. Evaluation of dental restorations: a comparative study between clinical and digital photographic assessments. Oper Dent. 2014;39(2):e45-56.

24. Smith R, Brook A, Elcock C. The quantification of dental plaque using an image analysis system: reliability and validation. J Clin Periodontol. 2001;28(12):1158-62.

25. Smith R, Karmo M, Brook A, Lath D, Rawlinson A. GIngival inflammation assessment by image analysis: measurement and validation. Int J Dent Hygiene. 2008;6:137-42.

26. Wong H, McGrath C, Lo E, King N. Photographs as a means of assessing developmental defects of enamel. Community Dent Oral Epidemiol. 2005;33(6).

27. Golkari A, Sabokseir A, Pakshir H, Dean M, Sheiham A, Watt R. A comparison of photographic, replication and direct clinical examination methods for detecting developmental defects of enamel. BMC Oral Health. 2011;11(16).

28. Cochran J, Ketley C, Sanches L, Mamai-Homata E, Oila A, Arnadottir I, et al. A standardized photographic method for evaluating enamel opacities including fluorosis Community Dent Oral Epidemiol. 2004;32(S1):19-27.

29. Cruz-Orcutt N, Warren J, Broffitt B, Levy S, Weber-Gasparoni K. Examiner reliability of fluorosis scoring: a comparison of photographic and clinical examination findings. J Public Health Dent. 2012;72(2):172-5.

30. Elfrink M, Veerkamp J, Aartman I, Moll H, Ten Cate J. Validity of scoring caries and primary molar hypomineralization (DMH) on intraoral photographs. Eur Arch Paediatr Dent. 2009;10(1):5-10.

31. Ellis J, Seymour R, Robertson P, Butler T, Thomason J. Photographic scoring of gingival overgrowth. J Clin Periodontol. 2001;28:81-5.

32. Alan R, Marakoglu I, Haliloglu S. Peri-implant crevicular fluid levels of cathepsin-K, RANKL, and OPG around standard, short, and mini dental implants after prosthodontic loading. J Periodontol Implant Sci. 2015;45:169-77.

33. Hayes A, Krippendorff K. Answering the call for a standard reliability measure for coding data. CMM. 2007;1(1):77-89.

34. Mankodi S, Ross N, Mostler K. Clinical efficacy of listerine in inhibiting and reducing plaque and experimental gingivitis. J Clin Periodontol. 1987;14(5):285-8.

35. Vastardis S, Yukna R, Fidel P, Leigh J, Mercante D. Periodontal disease in HIV-positive individuals: association of periodontal indices with stages of HIV disease. J Periodontol. 2003;74(9):1336-41.

36. Mariotti A, Hefti A. Defining periodontal health. BMC Oral Health. 2015;15(1):S6.

37. Marshall M, Lockwood A, Bradley C, Adams C, Joy C, Fenton M. Unpublished rating scales: a major source of bias in randomised controlled trials of treatments for schizophrenia. Br J Psychiatry. 2000;176:249-52.

38. de Souza M, de Toledo B, Rapp G, Zuza E, Neto C, Mendes A. Reliability of bleeding and non-bleeding on probing to gingival histological features. J Int Acad Periodontol. 2003;5(3):71-6.

39. Engelberger T, Hefti A, Kallenberger A, Rateitschak K. Correlations among Papilla Bleeding Index, other clinical indices and historically determined inflammation of gingival papilla. J Clin Periodontol. 1983;10(6):579-89.

40. Greenstein G, Caton J, Polson A. Histologic characteristics associated with bleeding after probing and visual signs of inflammation. J Periodontol. 1981;52(8):420-5.

41. Lobene R, Weatherford T, Ross N, Lamm R, Menaker L. A modified gingival index for use in clinical trials. Clin Prev Dent. 1986;8(1):3-6.

42. Hefti A, Preshaw P. Examiner alignment and assessment in clinical periodontal research. Periodontol 2000. 2012;59:41-60.

43. Loe H, Silness J. Periodontal disease in pregnancy. I. Prevalence and severity. Acta Odontol Scand. 1963;21:533-51.

44. Loe H. The gingival index, the plaque index and the retention index systems. J Periodontol. 1967;6:610-6.

45. Greenstein G. The role of bleeding upon probing in the diagnosis of periodontal disease. J Periodontol. 1984;55(12):684-8.

46. Al Shayeb K, Turner W, Gillam D. Accuracy and reproducibility of probe forces during simulated periodontal pocket depth measurements. Saudi Dent J. 2014;26(2):50-5.

47. Lobene R, Mankodi S, Ciancio S, Lamm R, Charles C, Ross N. Correlations among gingival indices: a methodology study. J Periodontol. 1989;60(3):159-62.

48. Bentley C, Disney J. A comparison of partial and full mouth scoring of plaque and gingivitis in oral hygiene studies. J Clin Periodontol. 1995;22(2):131-35.

49. Ramfjord S. Indices for Prevalence and Incidence of Periodontal Disease. J Periodontol. 1959;1:51-9.

50. Massler M. The P-M-A index for the assessment of gingivitis. J Periodontol. 1967;38(6P2):592-8.

51. Alexander A, Leon A, Ribbons J, Morganstein S. An assessment of the inter- and intra-examiner agreement in scoring gingivitis clinically. J Periodont Res. 1971;6:146-51.

52. Klein S, Bohannan H, Anderson P, Disney J, Leone F. The 1st year of field activities in the National Preventive Dentistry Demonstration Program. Clin Prev Dent. 1979;8:3-6.

53. Rebelo M, de Queiroz A. Gingival indices: state of art. In: Panagakos F, Davies R, editors. Gingival Diseases: IntechOpen; 2011.

54. Boateng G, Neilands T, Frongillo E, Melgar-Quinonez H, Young S. Best practices for developing and validating scales for health, social, and behavioral research: a primer. Front Public Health. 2018;6(149).

55. Morgado F, Meireles J, Neves C, Amaral A, Ferreira M. Scale development: ten main limitations and recommendations to improve future research practices. Psicol Reflex Crit. 2017;30(3).

56. Panagiotakos D. Health measurement scales: methodological issues. Open Cardiovasc Med J. 2009;3:160-65.

57. Ainamo J, Bay I. Problems and proposals for recording gingivitis and plaque. Int Dent J. 1975;25(4):229-35.

58. Nowicki D, Vogel R, Melcer S, Deasy M. The gingival bleeding time index. J Periodontol. 1981;52(5):260-2.

59. DeVellis R. Scale development: theory and applications, fourth edition. Los Angeles: SAGE Publications; 2017.

60. Clark L, Watson D. Constructing validity: basic issues in objective scale development. Psychol Assess. 1995;7(3):309-19.

61. Messick S. Validity of psychological assessment: validation of inferences from persons' responses and performances as scientific inquiry into score meaning. Am Psychol. 1995;50:741-9.

62. Vinkers C, Tijdink J, Otte W. Use of positive and negative words in scientific PubMed abstracts between 1974 and 2014: retrospective analysis. BMJ. 2015;351:h6467.

63. Kimberlin C, Winterstein A. Validity and reliability of measurement instruments used in research. Am J Health-Syst Pharm. 2008;65:2276-84.

64. Hinkin T. A review of scale development practices in the study of organizations. J Management. 1995;21(5):967-88.

65. Ladhari R. Developing e-service quality scales: a literature review. J Retailing Consum Serv. 2010;17:464-77.

66. Netemeyer R, Bearden W, Sharma S. Scaling procedures: Issues and applications Thousand Oaks, CA: SAGE; 2003.

67. Bruton A, Conway J, Holgate S. Reliability: what is it and how is it measured? Physiotherapy. 2000;86(2):94-9.

68. Hernaez R. Reliability and agreement studies: a guide for clinical investigators. Gut. 2015;64(7):1018-27.

69. de Vet H, Terwee C, Knol D, Bouter L. When to use agreement versus reliability measures. J Clin Epidemiol. 2006;59:1033-9.

70. Streiner D, Norman G, Cairney J. Health measurement scales: a practical guide to their development and use. : Oxford University Press; 2015.

71. Scholtes V, Terwee C, Poolman R. What makes a measurement instrument valid and reliable? Injury. 2010:JINJ-4490.

72. Litwin M. How to measure survey reliability and validity. Thousand Oaks: SAGE Publications; 1995.

73. Gadermann A, Guhn M, Zumbo B. Estimating ordinal reliability for Likert-type and ordinal item response data: a conceptual, empirical, and practical guide. PARE. 2012;17(3):1-13.

74. McHugh M. Interrater reliability: the kappa statistic. Biochem Med. 2012;22(3):276-82.

75. Landis J, Koch G. The measurement of observer agreement for categorical data. Biometrics. 1977;33(1):159-74.

76. Liu J, Tang W, Chen G, Lu Y, Feng C, Tu X. Correlation and agreement: overview and clarification of competing concepts and measures. Shanghai Arch Psychiatry 2016;28(2):115-20.

77. Koo T, Li M. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med. 2016;15(2):155-63.

78. Weir J. Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. J Strength Cond Res. 2005;19(1):231-40.

79. Goodwin L, Goodwin W. Are validity and reliability "relevant" in qualitative evaluation research? Eval Health Prof. 1984;7(4):413-26.

80. Kane M. The argument-based approach to validation. School Psych Rev. 2013;42(4):448-57.

81. McGraw K, Wong S. Forming inferences about some intraclass correlation coefficients. Psychol Methods. 1996;1:30-46.

82. Shrout P, Fleiss J. Intraclass correlations: uses in assessing rater reliability. Psychol Bull. 1979;86:420-8.

83. Krippendorff K. Computing Krippendorff's alpha-reliability. University of Pensylvania; 2011.

84. Feng G. Mistakes and how to avoid mistakes in using intercoder reliability indices. Methodology. 2015;11(1):13-22.

85. Zhao X, Liu J, Deng K. Assumptions behind intercoder reliability indices. In: Salmon ICT, editor. Communication Yearbook 36. New York: Routledge2012. p. 418-80.

86. Cook D, Zendejas B, Hamstra S, Hatala R, Brydges R. What counts as validity evidence? Examples and prevalence in a systematic review of simulation-based assessment. Adv Health Sci Educ. 2013;19(2):233-50.

87. Messick S. Validity. Educational Testing Service. 1987;1987(2):i-208.

88. Cook D, Beckman T. Current concepts in validity and reliability for psychometric instruments: theory and application. Am J Med. 2006;119(2):166.e7-e16.

89. Kane M. Validation. In: Brennan RL, ed. Westport, CT: Praeger; 2006.

90. Streiner D. Starting at the beginning: an introduction to coefficient alpha and internal consistency. J Pers Assess. 2003;80(1):99-103.

91. Cook D, Kuper A, Hatala R, Ginsburg S. When assessment data are words: validity evidence for qualitative educational assessments Acad Med. 2016;91(10):1359-69.

92. McDowell I, Newell C. Measuring health: a guide to rating scales and questionnaires, 2nd edn. New York: Oxford University Press Inc; 1996.

93. McDowell I. Measuring health: a guide to rating scales and questionnaires, 3rd edn. New York: Oxford University Press Inc; 2006.

94. Cook D, Brydges R, Ginsburg S, Hatala R. A contemporary approach to validity arguments: a practical guide to Kane's framework. Med Educ. 2015;49(6):560-75.

95. Cook D, Hatala R. Validation of education assessments: a primer for simulation and beyond. Adv Simul. 2016;1(31):1-12.

96. American Educational Research Association, American Psychological Association, National Council on Measurement in Education. Validity. Washington, DC: Standards for educational and psychological testing; 2016.

97. Tavares W, Brydges R, Myre P, Prpic J, Turner L, Yelle R, et al. Applying Kane's validity framework to a simulation based assessment of clinical competence. Adv Health Sci Educ. 2018;23:323-38.

98. St-Onge C, Young M, Eva K, Hodges B. Validity: one word with a plurality of meanings. Adv Health Sci Educ. 2017;22:853-67.

99. Guyatt G, Deyo R, Charlson M, Levine M, Mitchell A. Responsiveness and validity in health status measurement: a clarification. J Clin Epidemiol. 1989;42(5):403-8.

100. Fayers P, Hays R. Assessing quality of life in clinical trials: methods and practice. 2nd edn. Oxford: Oxford University Press; 2005.

101. Barnett M, Gilman R, Charles C, Bartels L. Computer-based thermal imaging of human gingiva: preliminary investigation. J Periodontol. 1989;60(11):628-33.

102. Poklepovic T, Worthington H, Johnson T, Sambunjak D, Imai P, Clarkson J, et al. Interdental brushing for the prevention and control of periodontal diseases and dental caries in adults. Cochrane Database Syst Rev. 2013;18(12):CD009857.

103. Rai T, Karuna Y, Rao A, Nayak A, Natarajan S, Joseph R. Evaluation of the effectiveness of a custom-made toothbrush in maintaining oral hygiene and gingival health in cerebral palsy patients. Spec Care Dentist. 2018;38:367-72.

104. Ross N, Mankodi S, Mostler K, Charles C, Bartels L. Effect of rinsing time on antiplaque-antigingivitis efficacy of listerine. J Clin Periodontol. 1992;20:279-81.

105. Boye U, Willasey A, Walsh T, Tickle M, Pretty I. Comparison of an intra-oral photographic caries assessment with an established visual caries assessment method for use in dental epidemiological studies of children. Community Dent Oral Epidemiol. 2013;41(6):526-33.

106. Syndergaard B, Al-Sabbagh M, Kryscio R, Xi J, Ding X, Ebersole J, et al. Salivary biomarkers associated with gingivitis and response to therapy. J Periodontol. 2014;85(8):e295-e303.

107. Canada H. Report on the findings of the oral health component of the Canadian Health Measures Survey 2007-2009. Ottawa, Ontario; 2010.

108. Armitage G. Development of a classification system for periodontal diseases and conditions. Ann Periodontol. 1999;4(1):1-6.

109. Nair S, Faizuddin M, Dharmapalan J. Role of autoimmune responses in periodontal disease. Autoimmune Dis. 2014;2014(596824).

110. Hong C, Napenas J, Hodgson B, Stokman M, Mathers-Stauffer V. A systematic review of dental disease in patients undergoing cancer therapy. Support Care Cancer. 2010;18(8):1007-21.

111. Vasanthan A, Dallal N. Periodontal treatment considerations for cell transplant and organ transplant patients. Periodontol 2000. 2007;44:82-102.

112. Tettamanti L, Lauritano D, Nardone M, Gargari M, Silvestre-Rangil J, Gavoglio P, et al. Pregnancy and periodontal disease: does exist a two-way relationship? Oral Implantol. 2017;10(2):112-18.

113. Ciancio S. Medications' impact on oral health. JADA. 2004;135(10):1440-48.

114. Boke F, Gazioglu C, Akkaya S, Akkaya M. Relationship between orthodontic treatment and gingival health: a retrospective study. Eur J Dent. 2014;8(3):373-80.

115. Schifter M, Yeoh S-C, Coleman H, Georgiou A. Oral mucosal diseases: the inflammatory dermatoses. Aus Dent J. 2010;55(s1):23-38.

116. Scully C, Carrozzo M. Oral mucosal disease: lichen planus. Brit J Oral Max Surg. 2008;46(1):15-21.

117. Valentine J, Hedges L, Cooper H. The handbook of research synthesis and meta-analysis. 2nd ed. New York: Russell Sage Foundation; 2009.

118. Kosmidis I. Improved estimation in cumulative link models. JR Statist Soc B. 2013;776(1):169-96.

119. Sperandei S. Understanding logistic regression analysis. Biochem Med. 2014;24(1):12-8.

120. Freese J, Long J. Regression models for categorical dependent variables using stata College Station: Strata Press; 2006.

121. Laitila T. A pseudo-R2 measure for limited and qualitative dependent variable model. J Econmetric. 1993;56(3):341-55.

122. Krippendorff K. Content analysis: an introduction to its methodology. 3rd edn. Thousand Oaks, CA: SAGE Publications; 2013.

123. Hubley A, Zumbo B. Validity and the consequences of test interpretation and use. Soc Indic Res. 2011;103(2):219-30.

124. Custers E. Long-term retention of basic science knowledge: a review study. Adv Health Sci Educ. 2010;15(1):109-28.

125. Lanning S, Pelok S, Williams B, Richards P, Sarment D, Oh T-J, et al. Variation in periodontal diagnosis and treatment planning among clinical instructors. J Dent Ed. 2005;69(3):325-39.

126. de Souza P, de Toledo B, Rapp G, Zuza E, Neto C, Mendes A. Reliability of bleeding and non-bleeding on probing to gingival histological features. J Int Acad Periodontol. 2003;5(3):71-6.

127. Ranney R. Diagnosis of periodontal diseases. Adv Dent Res. 1991;5:21-36.

128. Manning D, Ethell S, Donovan T. Detection or decision errors? Missed lung cancer from the posteroanterior chest radiograph. Br J Radiol. 2004;77:231-5.

129. Dallas E, Clement R, Taylor D. Diagnosis from fundus photographs. Br J Ophthalmol. 2007;91(608-12).

130. Grant J, Marsden P. The structure of memorized knowledge in students and clinicians: an explanation for diagnostic expertise. Med Educ. 1987;21:92-8.

131. Gruppen L, Frohna A. Clinical Reasoning. In: Norman G, van der Vleuten C, Newble D, editors. International handbook of research in medical education. Dordrecht, Boston: Springer; 2002.

132. Hill E, Slate E, Weigand R, Grossi S, Salinas C. Study design for calibration of clinical examiners measuring periodontal parameters. J Periodontol. 2006;77:1129-41.

133. Kelly A, Antonio A, Maia L, Luiz R, Vianna R, Quintanilha L. Reliability assessment of a plaque scoring index using photographs Methods Inf Med. 2008;47.

134. Ohrn K, Sanz M. Prevention and therapeutic approaches to gingival inflammation. J Clin Periodontol. 2009;36:20-6.

135. Mizutani S, Ekuni D, Tomofuji T, Azuma T, Kataoka K, Yamane K, et al. Relationship between xerostomia and gingival condition in young adults. J Periodontol Res. 2015;50(1):74-9.

136. Zapf A, Castell S, Morawietz L, Karch A. Measuring inter-rater reliability for nominal data - which coefficients and confidence intervals are appropriate? BMC Med Res Methodol. 2016;16(93):1-10.

137. Orban B. Clinical and histologic study of the surface characteristics of the gingiva. Oral Surg Oral Med Oral Pathol Oral Radiol. 1948;1:827.

138. Mulic A, Tveit A, Wang N, Hove L, Espelid I, Skaare A. Reliability of two clinical scoring systems for dental erosive wear. Caries Res. 2010;44:294-99.

139. Erten H, Uctasli M, Akarslan Z, Uzun O, Baspinar E. The assessment of unaided visual examination, intraoral camera and operating microscope for the detection of occlusal caries lesions. Oper Dent. 2005;30(2):190-4.

140. Boyer K, Verma R. Multiple raters in survey-based operations management research: a review and tutorial. Prod Oper Manag. 2000;9(2):128-40.

141. Carter N, Bryant-Lukosius D, DiCenso A, Blythe J, Neville A. The use of triangulation in qualitative research. ONF. 2014;41(5):545-7.

142. Yang Y, Green S. Coefficient alpha: a reliability coefficient for the 21st century? J Psychoed Assess. 2011;29(4):377-92.

143. Lang N, Adler R, Joss A, Nyman S. Absence of bleeding on probing. An indicator of periodontal stability. J Clin Periodontol. 1990;17:714-21.

144. Newbrun E. Indices to measure gingival bleeding. J Periodontol. 1996;67(6):555-61.

145. Haj-Ali R, Feil P. Rater reliability: short- and long-term efects of calibration training. J Dent Ed. 2005;70(4):428-34.

146. Torres-Pereira C, Possebon R, Simoes A, Bortoluzzi M, Leao J, Giovanini A, et al. Email for distance diagnosis of oral diseases: a preliminary study of teledentistry. J Telemed Telecare. 2008;14(8):435-8.

147. Totty J, Harwood A, Wallace T, Smith G, Chetter I. Use of photograph-based telemedicine in postoperative wound assessment to diagnose or exclude surgical site infection. J Wound Care. 2018;27(3):128-35.

148. Sanchez Dils E, Lefebvre C, Abeyta K. Teledentistry in the United States: a new horizon of dental care. Int J Dent Hygiene. 2005;2(4):161-4.

149. Miller P. Measurement Error. In: Lavrakas P, editor. Encyclopedia of Survey Research Methods. Thousand Oaks, CA: SAGE Publications; 2008.

150. Bowling A. Mode of questionnaire administration can have serious effects on data quality. J Pub Health. 2005;27(3):281-91.

151. Cromey D. Avoiding twisted pixels: ethical guidelines for the appropriate use and manipulation of scientific digital images. Sci Eng Ethics. 2010;16(4):639-67.

152. O'Reilly-Shah V. Factors influencing healthcare provider respondent fatigue answering a globally administered in-app survey. PeerJ. 2017;5:e3785.

153. Bornstein M, Jager J, Putnick D. Sampling in developmental science: situations, shortcomings, solutions, and standards. Dev Rev. 2013;33(4):357-70.

154. Trombelli L, Farina R, Silva C, Tatakis D. Plaque-induced gingivitis: case definition and diagnostic considerations. J Periodontol. 2018;89(S1):S46-73.

155. Chapple I, Mealey B, Van Dyke T, Bartold P, Dommisch H, Eickholz P, et al. Periodontal health and gingival diseases and conditions on an intact and a reduced periodontium: consensus report of workgroup 1 of the 2017 world workshop on the classification of periodontal and peri-implant diseases and conditions. J Periodont. 2018;89:S74-84.

156. Sadler M, Yamamoto R, Khurana L, Dallabrida S. The impact of rater training on clinical outcomes assessment data: a literature review. Int J Clin Trials. 2017;4(3):101-10.

# *Appendix A: Participant Information Page on REDCap Survey*

*Confidential*



## Participant Information

Please answer the following questions

| | |
|---|---|
| Please provide your profession | ◯ Dentist<br>◯ Registered Dental Hygienist<br>◯ Researcher |
| How long have you been in practice (if dentist or hygienist)? Please provide a numeric answer | _____<br>(ex. 5 (for 5 years)) |
| Where did you receive your oral health education? | _____ |

# *Appendix B: Information Provided on REDCap Survey*

*Confidential*



## Modified Gingival Index

Please review the information below to understand how to complete the survey

---

The MGI is used to assess the prevalence and severity of gingivitis by visually examining the gingiva, without the use of instrumentation. You are asked to examine 72 photos and assign a score based on the following criteria:

0: healthy gingiva: absence of inflammation

1: mild inflammation: slight change in color and little change in texture (loss of stippling) but only a portion of margin

2: mild inflammation: same as above but involving entire gingival margin

3: moderate inflammation: glazing, redness, edema (swelling), hypertrophy of margin

4: severe inflammation: marked redness, edema, hypertrophy of margin; spontaneous bleeding; ulceration

Please ensure that the brightness on your screen is at the maximum.

Examples of each category are shown below:

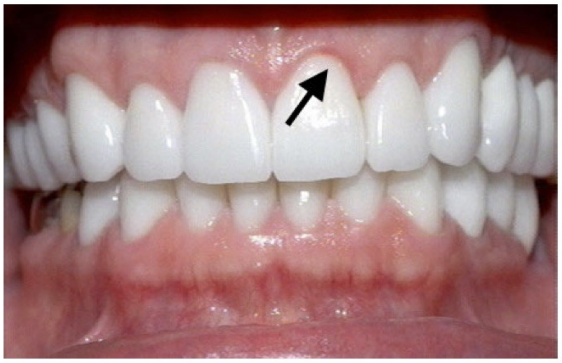
---

Healthy gingiva

0: absence of inflammation: light pink colour, stippling of gingiva, knife-shaped margins, pointed papilla

Mild inflammation

1: slight change in color and little change in texture but only a portion of margin



Mild inflammation

2: same as above but involving entire gingival margin

Moderate inflammation

3: glazing, redness, edema (swelling), hypertrophy of margin

Severe inflammation

4: marked redness, edema, hypertrophy of margin; spontaneous bleeding; ulceration