INTER-RATER RELIABLITY IN COMPREHENSIVE EXAMINATION SCORING: THE CASE FOR CONSISTENT AND COLLABORATIVE RATER TRAINING AND CALIBRATION

David Arron Saenz

February 15, 2023

ABSTRACT

There is a vast body of literature documenting the positive impacts that rater training and calibration sessions have on inter-rater reliability as research indicates several factors including frequency and timing play crucial roles towards ensuring inter-rater reliability. Additionally, increasing amounts research indicate possible links in rater-reliability and grade inflation pertaining to faculty status and length of employment. This research study analyzed the impact that consistent and collaborative training and rubric calibration, or the lack thereof, has on inter-rater reliability. Additionally, this study investigated whether faculty status and faculty length of employment has an influence on scoring reliability. Finally, participatory action research was planned to assist in the creation of a formal training policy and processed that could then be tested to explore the impact that such training would have on future examination rater-reliability. Five years of examination scoring data was utilized from a small Midwest region private graduate school. Intraclass correlation coefficient statistical analysis was employed to determine inter-rater reliability along with independent samples t-test to determine statistical significance the faculty groups. Mean scoring differences outputs were then tested utilizing a Likert-type scale to evaluate scoring gaps amongst faculty. The findings within this study indicate that inter-rater reliability is negatively impacted when no formal or consistent training. However, no significant differences between the mean scores were found based on faculty status nor between faculty length of employment. The hypothesis testing yielded support of the main hypothesis within this research study and the current literature in which inter-rater reliability is negatively impacted when no formal or consistent rater-training and rubric calibrations are performed for raters of examinations. The implications for practice resulting from this study display the need

for formal training policy and processes be created, and consistently completed to ensure inter-rater reliability will be positively impacted.

DEDICATION

It takes a village, so with that, I have several dedications and displays of gratitude to share for those that have supported me not only through the journey of completing this dissertation, but through the journey of life for which we all need shepherding and inspiration along the way.

First, I dedicate this to my wife and children who have been a source of support, motivation, perseverance, and strength throughout this entire experience. They have and always will be a beacon of inspiration for me both personally and professionally.

I also dedicate this to my mom, Eugenia Alvarez. Mom thank you for all that you have done throughout my life to support me. I know it was not easy and you made innumerable sacrifices along the way. Your support in the background, has always led me in the foreground.

An additional dedication to YiaYia and Papou for your unwavering support over the years and the pride you have shown towards me for undertaking and accomplishing this goal.

A special feeling of gratitude and dedication is also extended to my brothers and sisters and entire family along with my friends. You have all tirelessly motivated me and kept up my spirits in your own way for as far back as I can remember, and for that I am truly grateful.

Finally, to my Pop, A1C Richard A. Saenz. I lost you halfway through this journey, but your memory has served as the catalyst for completion. Nobody was prouder of me for having started this process and confidently knowing I would finish it than you were. I miss you every day, but I will miss you the most when I walk across the stage, look up to the rafters and not see you there for the first time in my scholarly life. You never truly forget a person you love and lose; you just learn to live without them. Unfortunately, the void it leaves is never filled, it just gets smaller. Regardless, we did it Pop.

ACKNOWLEDGMENTS

I would like to acknowledge and thank the following individuals who have helped complete my scholarly research and made this experience memorable and fun along the way. Dr. Jeffrey Bakken, for serving not only as my internship supervisor, but a mentor throughout. To my fellow EdD cohort of colleagues, from Introduction 1 to Introduction 17, you all played such an important role in my keeping my spirits high and keeping our focus on the finish line. I am thankful for having the opportunity to meet you and grateful knowing that we will remain friends long after as well. A thank you to those throughout my higher education career that gave me an opportunity to learn and grow into a transformational leader. Finally, to my supportive colleagues at the Institute that encouraged and inspired me. Your contributions will never be forgotten.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION 1

Introduction 1

The Comprehensive Examination 1

Components 1

Administration 1

Scoring 2

Inter-Rater Reliability, Training, and Rater Calibration 3

Statement of the Problem 3

Literature Review 3

Positive Impacts of Training and Calibration on Inter-rater Reliability 4

Calibration and Scoring 5

Faculty Status and Scoring 6

Solutions and Alternatives 7

Summary of Literature 8

Research Purpose 19

Definition of Key Terms 10

Significance of the Study 11

Organization of the Research Report 11

CHAPTER 2: LITERATURE REVIEW 13

Introduction 13

Summary of the Research Problem 13

Theoretical Framework 13

Positive Impacts of Training and Calibration on Inter-rater Reliability 14
Types of Training.....16
Opposing Views on Training.....18
Summary.....19
Calibration and Scoring 20
Timing and Scoring.....22
Summary.....24
Faculty Status and Scoring.....25
Grade Inflation.....28
Summary.....32
Solutions and Alternatives.....32
Summary of the Literature.....33
CHAPTER 3: RESEARCH METHODOLOGY AND METHODS 35
Introduction.....35
The Comprehensive Examination.....35
Components 35
Administration 35
Scoring.....36
Inter-Rater Reliability, Training, and Rater Calibration.....37
Similar Research on Training and Calibration.....37
Research Problem, Purpose, and Questions.....38
Research Problem 38
Research Purpose.....39

Research Question 39
Research Methodology 40
Quantitative Methodology 40
Action Research and Qualitative Methods 41
Research Design....42
Summary....42
Research Context 43
Setting 43
Participants....43
Participant Recruitment and Selection....44
Research Methods....44
Quantitative Data Collection and Analysis....44
Strategies....45
Timeline 45
Data Analysis....46
Hypothesis 1....47
Hypothesis 2....47
Hypothesis 3....48
Hypothesis 4....48
Procedures....49
Table 1 49
Qualitative Data Collection and Analysis....49
Researcher Positionality....50

CHAPTER 4: FINDINGS AND DISCUSSION 52
Introduction.....52
Findings.....53
Hypothesis 1.....53
Inter-rater reliability.....54
Independent Samples *t*-test.....55
Scoring Gaps Analysis.....55
Hypothesis 2.....57
Independent Samples *t*-test.....57
Hypothesis 3.....58
Inter-rater reliability.....58
Independent Samples *t*-test.....60
Scoring Gaps Analysis.....61
Hypothesis 4.....63
Inter-rater reliability.....65
Discussion.....65
Hypothesis 1.....65
Synthesis of Themes.....66
Connections to Literature.....67
Hypothesis 2.....68
Synthesis of Themes.....68
Connections to Literature.....68

Hypothesis 3.....69
Synthesis of Themes.....69
Connections to Literature.....70
Hypothesis 4.....71
Synthesis of Themes.....71
Connections to Literature.....71
Conclusion.....72
CHAPTER 5: CONCLUSIONS.....74
Introduction.....74
Research Purpose, Questions, and Answers.....75
Implications for Practice.....77
Diversifying Assessment.....77
Policy and Process Creation.....78
Suggestions of Future Research.....79
Addressing Rater Bias.....79
Timing of Training and Scoring.....80
Faculty Status of Raters.....82
Limitations.....82
Conclusion.....84
REFERENCES.....86

CHAPTER 1

Introduction

The Comprehensive Examination

At the Development Institute, a small, private graduate school in the Midwest region of the United States, a comprehensive examination is administered to students upon completion of all degree coursework requirements. The comprehensive exam has been administered for over 50 years and is viewed as the capstone to the students learning experience at the Institution. The examination is the highest stake component of the culminating requirements of the master's degree program. Students that successfully pass the comprehensive examination are eligible for degree conferral. Therefore, the stature of the comprehensive examination is matched only by the anxiety students feel as the day of the exam approaches.

Components

The comprehensive examination is a five-hour, written examination that covers major content areas within the curriculum. The exam is administered three times a year, once in late April, another in early June, and the final in early September. The examination questions are developed by representatives from the faculty assessment committee. Two faculty members, traditionally one tenured and one tenure-track faculty write the six-question examination near the beginning of the spring semester. Students are required to apply to sit for the examination.

Administration

The examination, pre-COVID, was traditionally administered in-person on a Saturday morning. Students are provided with desktop testing stations, paper, and pens for notes as well as earplugs as needed. Proctors are in each room to ensure examination integrity. Students are not allowed use of the internet and are only allowed to bring in beverages and snacks. Leaving

the testing station is allowed for restroom breaks. Upon completion, the student is instructed to leave all notes and testing material at their station. The proctor then removes the comprehensive examination from the desktop and saves it in a secure drive. Current state (during COVID), all students are administered the exam via a course in the Institutes' learning management system. The course is made available three days before the start of the exam and contains day of exam instructions, a writing template, and contact information. The exam is not synchronously live proctored, meaning proctors are not logged in viewing the students while they attempt the exam. Students are on an honor code in that they are instructed not to use any electronic devices nor the internet while attempting the exam. The exam is submitted through the grading center and the exam administrator pulls the exam, saves it a file that removes author data, to ensure blind grading. In both cases, students with approved accommodations are provided various supports such as extra time, multiple days for completion and/or specialized hardware to complete the exam.

Scoring

Upon completion of the examination, the exam administrator emails the comprehensive exam, along with the grading rubric, score sheet and deadline to complete their scoring to the designated faculty rater. Faculty will then send in their completed score sheets to be tabulated. A master scoring sheet, that contains the artificial student identification number that ties back to the real student identification is utilized to tabulate scores and flag for a possible third reader, as necessary. Once scoring is completed, students' scores are reported out to their assigned advisors as pass or fail. The advisor is then charged with contacting the student, via phone, and informing them of the results. If the student fails the examination, they will meet with their

advisor to create a strategy to prepare for the second attempt. Subsequent failed attempts require more in-depth advising and planning.

Inter-Rater Reliability, Training, and Rater Calibration

As previously mentioned, at no point, over the past four years and twelve comprehensive exam administrations has inter-rater reliability been assessed, nor formal or consistent rater training and calibration taken place. Talks of the need to do so have come up sporadically throughout that time. However, no action has taken place. As the institute's headcount continues to grow, so too does the amount of exams administration and ultimately, the need to ensure inter-rater reliability.

Statement of the Problem

There is a vast body of literature documenting the positive impacts that rater training and calibration sessions have on inter-rater reliability. Moreover, research indicates that the frequency and timing of training plays a crucial role in ensuring accuracy and equity in scoring. Further, there are increasing amounts research that have been completed regarding faculty status and its impact on grading and grade inflation. Despite the positive indices that highlight the impact of training and calibration, institutes of higher education often do not consistently train, assess, and calibrate faculty scorers.

Literature Review

This literature review serves to provide insight into the following areas of inter-rater reliability and calibration. First, an exploration of the research that points to positive effects of training and calibration as well as look into the types of training that have significant impact on inter-rater reliability, and finally a discussion on opposing viewpoints on training that is present in the literature. Second, a look at the increased literature that describes how not only the timing

of training impacts scoring, but also how the frequency and length of time scoring plays a role in shaping reliability. Third, a comprehensive review of the literature surrounding the impacts that faculty status has on scoring and grade inflation. Finally, a review of current proposed solutions to increase reliability and literature that offers up alternative approaches to this ever-growing field of research.

Positive Impacts of Training and Calibration on Inter-rater Reliability

There is a vast body of literature which has studied the positive impacts that rater training and calibration sessions have on inter-rater reliability. In a simple search utilizing keywords such as *inter-rater reliability* and *rater calibration*, one is assured to be provided with hundreds of journal articles. These journal articles, which span from higher education to business management, all share important similarities, those being, the importance of training and calibration and how they positively impact reliability in scoring and assessment.

In a study examining efficiency in assessment, researchers trained both students and faculty on how to rate essays. The results showed that the trained students rated as reliable as the faculty and the lack of rater-training is attributed to negative inter-rater reliability (Cole et al., 2012). Hung et al. (2012) focused their research on rater effects and the impact that tools utilized to train raters have on inter-rater reliability. The authors posit that reliability is impacted by focused training and moreover, raters that were not calibrated on the scoring rubric were shown to have less reliable scores. Crucially, PapaJohn (2002) found that the concept of concept mapping in training had an impact on reliability amongst more experienced raters. In a study completed by Rios et al (2017) in which the focus of the study was on the creation of a new assessment tool and how reliable computerized scoring would be in relation to human scoring,

they report that raters who went through comprehensive training and calibration before scoring, had displayed significant inter-rater reliability.

Unfortunately, despite the positive indices, institutes of higher education, and specifically the one being researched for this study, often do not consistently train, assess, and calibrate faculty scorers. As a result, inter-rater reliability has become a growing concern amongst the faculty and administrators.

Calibration and Scoring

The impacts of training, along with the types of training that are effective in ensuring higher rates of interrater reliability presented in this study have provided a setting for discussion. To further the discussion, a review of the literature specifically on calibration and actual scoring is presented. Research indicates that the timing and frequency of training plays a crucial role in ensuring accuracy and equity in scoring. Moreover, the timing of when scoring is completed by the rater has been researched and the impacts of how frequently and when scoring is completed and its effect on reliability assessed. Szafran (2017) found that not only was the timing of the training crucial, but the amount of rater collaboration in the training sessions showed an increase in inter-rater reliability. His study further found that the frequency and timing of when and how to integrate the third reader can impact inter-rater reliability.

Finn et al. (2018) undertook a very relatable study, as it pertains to gaps in scoring and non-structured timeframes for completing scoring. In their study, the authors research the impacts of time between scoring and scoring accuracy. The authors' data was derived from 350 raters of the Graduate Record Examination (GRE) general test. The results showed that gaps between scoring were found to decrease accuracy validity and reliability as well. However, the number of consecutive days scoring did not impact scoring validity or reliability. Further, the

authors found that the time spent scoring each response was a significant predictor in accuracy validity and reliability.

The Finn et al (2018) study was very useful and worthwhile as it provides a significant amount of data that supports how gaps between grading can negatively impact scoring. The policy of grading the comprehensive examination at the Institute is to set a deadline as to when raters are to submit their finalized grading sheets. One possible avenue for future research, upon conclusion of this study, would be to further delve into when the raters submit the grading sheets and gaps between submission. This could provide a critical data set to report out on. Some limitation exists as raters tend to either send their scores in bulk at one time or as they complete. Which, then can impact the reliability of the data is it pertains to when the scoring occurs. However, as the scores are logged on individual excel files, "last date saved" data could be extracted. Nonetheless, it will be interesting to explore.

Faculty Status and Scoring

There is significant research on the impact that faculty status has on scoring. Extensive studies highlight how faculty status (tenure, tenure-track, non-tenure) plays a role for those raters in scoring. Moreover, studies have researched grade inflation as it relates to faculty status as well as delving into the psychology as to why non-tenure and tenure-track faculty tend to either inflate grades or grade lower.

In one study, the researcher found a difference in grading/assessment between tenure and tenure track faculty. Where the latter grade more lenient then the former. Moreover, the authors posited that reciprocity in grades and instructor evaluation could be present, thus concluding a tenure-track professor be more lenient in grading with the expectation of receiving a better instructor evaluation from his/her students (Filetti et al., 2010).

That study lends itself well to the research completed by Kezim et al. (2005) in which they reviewed grade point averages over a 20-year period to investigate grade differences in mane GPA among adjunct, untenured, and tenured faculty. Their results showed adjunct faculty tended to grade higher than tenured and non-tenured track faculty. The authors posited adjuncts grade higher to score better on after-class faculty assessments completed by students. Deboer et al. (2007) completed a quantitative study utilizing survey research. The instructors were surveyed using an assessment tool. The focus of the study was to assess the impact of instructor attitude on grading behavior. The results supported the authors hypothesis that rater personality, based on the assessment, plays a part in grading. Significantly, the authors found that a higher need for approval resulted in lower grades. The authors posit that the lower grades and need for approval reflect attitudes of instructor's desire to attain approval from their fellow colleague instructors. Further, the author's data suggest that full professors assign significantly fewer higher grades than academic staff and the authors posit that full professors have less worry about job security therefore are more apt to grade harsher.

As the research displays, faculty status is found to have an impact on scoring. This is significant research and pertains well to this study as we have an ample amount of data as well as faculty statuses that can be tested again hypothesis on impacts of faculty status and comprehensive examination scoring. Further, the Institute has several faculty members that have been employed for 20-years or more. Those factors will certainly provide a wealth of data and discussion points.

Solutions and Alternatives

Throughout the review there has been several solutions posited as well as alternative foci as it relates to inter-rater reliability, training, and calibration. Rubric training in which

assessment is narrowed and focused was shown to increase inter-rater reliability (Attali, 2013). This research finding and several other trend towards not only to slimmed down rubrics, but also rater training, calibration, and rubric creation that is done in a more transparent and collaborative nature. Bamber (2015), while using student as participants in trainings, found that having a transparent and collaborative approach towards calibration increased inter-rater reliability. Further, the participants also felt an increased self-awareness of what was being assessed and positive feelings towards the assessment. This is an important study and one that could be utilized as an approach deployed by the Institution or studied further as it asks a question of whether students would perform better if they were fully informed, in detail, on how the rubrics are set and what is being assessed. Lee (1985) in her case study review of rater training suggested similar conclusions in that the training should be concise and transparent as possible to increase reliability in practices.

An alternative to rubric training is offered up by Bernardin and Buckley (1981) in which they argue that the focus of rater training should not be on the rubric and calibration, rather, it should be on the raters themselves, exploring their biases and tendency in grading and how to best calibrate against those. Herbert et al. (2014) asks a crucial question as part of their study, which is what at what expense does training and calibration have on the objectivity of the rater. Further, Papajohn (2002) posits that even after training a rater, one should note that raters may still not interpret and score in similar fashion, although, finding that experienced raters tend to rate more similar.

Summary of the Literature

The literature surrounding inter-rater reliability, training, calibration, solutions, and alternative approaches is robust. As such, inter-rater reliability continues to be a focus area for

research. With such in-depth research with various methods deployed, the one outcome most researchers have agreed upon is the importance that training plays on positively impacting inter-rater reliability. What remains uncertain is how institutes of higher education can assure training and calibration occurs, while faculty juggle their day-to-day tasks as teachers, advisors, mentors, and faculty administrators.

Research Purpose

This study will describe the impact that consistent training and calibration has on inter-rater reliability. Further, it will examine the relationship between faculty status and scoring as it pertains to inter-rater reliability. Five years of examination scoring data from 25 faculty members will be used from The Development Institute, a small Midwest region private graduate school.

The following research hypotheses will guide the study:

- **Hypothesis 1:** Is inter-rater reliability negatively impacted when no formal or consistent rater training is provided?
- **Hypothesis 2:** Will the scoring of tenured faculty be lower than tenure-track, clinical faculty, and adjunct faculty?
- **Hypothesis 3:** Will the scoring of recently hired faculty be more reliable than faculty that has been hired over 10 years?
- **Null Hypothesis:** Inter-rater reliability is high and consistent amongst all faculty.
- **Hypothesis 4:** (After action) Will inter-rater reliability significantly increase among trained raters regardless of faculty status.

Definition of Key Terms

Throughout this research study, key terms will be utilized to describe data components. The following is a list of key terms along with their definitions.

- **Adjunct faculty** is defined as faculty utilized to teach sections of courses. They are not considered full-time faculty and are not tenure eligible.
- **Artificial ID number** is defined as an identification number that has been created to mask the actual student identification number for blind scoring purposes.
- **Clinical faculty** is defined as faculty that are not tenure eligible.
- **Comp** is used to describe the Comprehensive Examination in short form.
- **Consistent rater training** is defined as formal training that has occurred at minimum once per year.
- **Formal rater training** is defined as a training delivered to faculty in-person/online in a structured and methodical format and setting. It is assumed to be required for all faculty raters.
- **High agreement** is defined as scoring between raters having averaged a range of 0-3 points of separation.
- **Inter-rater reliability** is defined as the degree at which scoring agrees among raters.
- **Low agreement** is defined as scoring between raters having averaged a range of 8-11.
- **Moderate agreement** is defined as scoring between raters having averaged a range of 4-7 points of separation.
- **No agreement** is defined as scoring between raters average 12 points or greater.
- **Not-recently hired faculty** is defined as faculty that has been employed with the Institute for 10+ years.

- **Rater Calibration or Calibration** is defined as an act, such as a training to ensure that a group of raters score consistently in utilizing a scoring rubric.
- **Recently hired faculty** is defined as faculty that has been employed with the Institute for 0-9 years**.**
- **Tenured faculty** is defined as an academic appointment without limit of time.
- **Tenured-track faculty** is defined as faculty who is eligible to be granted tenure at the Institute within a certain period of time.

Significance of the Study

Similar to the literature reviewed in this chapter, this study will continue to place the spotlight on the importance of training and calibration sessions. Inter-rater reliability is a key facet to ensuring scoring is fair and accurate, especially, in high stakes examinations. The significance of this study will go towards considerably impacting policy and process at The Development Institute as it relates to rater training and calibration for faculty raters that score the comprehensive examination at the Institute. I am confident that the Institute will ensure, moving forward, that rater training and calibration sessions occur at minimum once per year, ideally before and after the examinations. If this occurs, the vast amount of literature, including this action research, will surely improve inter-rater reliability moving forward. Further, I am confident that this study and its action can be utilized as a template, or a starting point, for similar size institutions and faculty that struggle with the demands of the day-to-day activity, while also trying to ensure exam scoring integrity.

Organization of the Research Report

The organization of the report will follow in a manner in relation to the process of participatory action research. First, a review of the current literature will be presented. Within

the literature, several studies will be utilized to highlight not only the impact that training and calibration has on inter-rater reliability, but also proposed actions and alternative viewpoints. Second methods, statistical analysis, and a review of the data followed by member checking via a focus group. Finally, action implementation, and assessment of the implementation, along with results, data analysis, and discussion. During the conclusion, main takeaways from the study will be discussed along with recommended next steps.

CHAPTER 2

Literature Review

Introduction

This literature review serves to provide insight into the following areas of inter-rater reliability and calibration. First, an exploration of the research that points to positive effects of training and calibration as well as a look into the types of training that have significant impact on inter-rater reliability, and finally a discussion on opposing viewpoints on training that is present in the literature. Second, a look at the increased literature that describes how not only the timing of training impacts scoring, but also how the frequency and length of time scoring plays a role in shaping reliability. Third, a comprehensive review of the literature surrounding the impacts that faculty status has on scoring and grade inflation. Finally, a review of current proposed solutions to increase reliability and literature that offers up alternative approaches to this ever-growing field of research.

Summary of the Research Problem

There is a vast body of literature documenting the positive impacts that rater training and calibration sessions have on inter-rater reliability. Moreover, research indicates that the frequency and timing of training plays a crucial role in ensuring accuracy and equity in scoring. Further, there are increasing amounts research that have been completed regarding faculty status and its impact on grading and grade inflation. Despite the positive indices that highlight the impact of training and calibration, institutes of higher education often do not consistently train, assess, and calibrate faculty scorers.

Theoretical Framework

Formulating a theoretical perspective while undertaking the scholarly literature has helped frame and guide the following research. Active learning, which is rooted in the theory of constructivism is being utilized to not only provide an overarching theme for the scholarly review, but also foment discussion on the use of training, calibration and collaboration on the rater and the anticipated actions that occur thereafter. Constructivism is a theory in education that asserts learners *construct* knowledge, actively as opposed to passively, by integrating new understandings with previous understandings ("Constructivism-Philosophy of Education," 2021). Active learning theory/Constructivism has been linked to Jean Piaget and his theories of cognitive development. Piaget's theory discusses how humans acquire, construct, and utilize knowledge ("Piaget's Theory of Cognitive Development," 2021). As applied to this study, active learning theory/constructivism holds that one should expect training and calibration sessions, if implemented effectively and utilized consistently, to positively impact rater scoring and grading and thus inter-rater reliability.

Positive Impacts of Training and Calibration on Inter-rater Reliability

There is a vast body of literature which has studied the positive impacts that rater training and calibration sessions have on inter-rater reliability. In a simple search utilizing keywords such as *inter-rater reliability* and *rater calibration*, one is assured to be provided with hundreds of journal articles. These journal articles, which span from higher education to business management, all share important similarities, those being, the importance of training and calibration and how they positively impact reliability in scoring and assessment.

In a study examining efficiency in assessment, researchers trained both students and faculty on how to rate essays. The results showed that the trained students rated as reliable as the faculty and the lack of rater-training is attributed to negative inter-rater reliability (Cole et al.,

2012). Further, Gardiner et al. (2020), in a longitudinal case study, assessed the impact that rubric-based training has on inter-rater reliability. The authors findings point to positive impacts and shared findings and results of previous researchers in the value of training and calibration, specifically, that calibration and rater training is a proven method in increasing inter-rater reliability. Gugiu et al. (2012) also performed their research utilizing rubrics as building blocks for training. Their quantitative study utilized 29 undergraduate students enrolled in an introductory political science course at a university located in the Midwest. The students were randomly assigned into groups and two raters were provided grading rubrics. Importantly, the rubrics created had clearly defined measurements and instructions on what to assess and how to grade. Upon analyzing the results of the grading, the researchers found significant reliability in scoring between the raters. The authors posit that the impact of a clear grading rubric along with a smaller number of raters positively impacted reliability. Additionally, Rios et al. (2017) utilizing the written assessment instrument HEIgthen WC sampled 985 examinees from a swath of exams delivered by a four-year institution of higher education. The authors studied three aspects of the examination, the first being quality and reliability of the assessment, the second, reviewed automated scoring versus human scoring outcomes, and the third a review of the scores on the HEIghten WC as compared to other examinee variables such as their SAT scores and GPA. The important finding in this research as it relates to reliability was that both the human and automated scoring displayed high rater reliability. Thus, the authors expect such assessments in writing completed by humans or automated have the ability to be reliable should the human raters go through professional scoring training and calibration, as was done with the raters in this study.

As mentioned, there is a vast amount of research that highlights the impact of training and calibration studies are proven to go a long way towards ensuring inter-rater reliability and equity in scoring and grading. In each of the studies cited, questions were posited in the literature on the types of training that are the most impactful on scoring and grading reliability. In the following section, several studies are offered, and their findings presented to display key findings on the types of trainings that researchers utilized that showed promising in ensuring inter-rater reliability.

Types of Training

Research has shown that training and calibration are proven to go a long way towards ensuring inter-rater reliability and equity in scoring and grading. However, questions arise in the literature on the types of training that are the most impactful on scoring and grading reliability. From rubric specific to focusing the training on the trainer, researchers have diligently tested and reported back on types of trainings and their influence in scoring and grading.

To introduce the varying research on training types, Hung et al. (2012) focused their research on rater effects and the impact that tools utilized to train raters have on inter-rater reliability. The authors posit that reliability is impacted by focused training and moreover, raters that were not calibrated on the scoring rubric were shown to have less reliable scores. From training on rubrics to training on concepts, PapaJohn (2002) undertook and found that the use of concept mapping in training had an impact on reliability amongst more experienced raters. The concept mapping design was utilized not only in training by calibration amongst raters with the goal of ensuring that raters were all conceptualizing what was being assessed and how it was being assessed. In an important study that utilizes training, along with collaboration and transparency in the grading methods, Bamber (2015) studied the impact of stakeholder

confidence of increased transparency in the examination assessment process. This mixed method study used survey research design and focus groups. 115 post-graduate students from a University in the United Kingdom participated in the study and completed the survey questionnaire. The focus of the study was to assess the benefits of involving students in the process of assessing scoring and grading and improve student's understanding on assessment and evaluation. The results showed that participants reported enhanced awareness of what is being assessed and increased positive feelings towards the assessment. Further, the participants felt enhanced personal awareness of their own practices as they relate to examination taking. Moreover, and perhaps the most telling is that the findings indicate increase rater reliability because of training and rater calibration. This study is quite useful as it supports a hypothesis that consistent training and calibration is needed to ensure inter-rater reliability. Further, the transparency and use of students is also a possible action that may be undertaken as it relates to updating rubrics. Finally, Lee (1985) offers an alternative perspective on rating. Lee's case study reviewed rating from the lens of employer to employee performance rating. The author studied cognitive operations of the raters, such how the raters observe, gather, and integrates rater training into their performance appraisal of the employees. Additionally, the author studied whether the raters were able to observe or not observe performance and how observing outputs impacted the raters. The author found that rater performance is impacted by several factors, such as rater attitudes, observational ability, and social contexts. Crucially however, the author posits that designing a rater training program, and implementing it successfully may improve rater accuracy and effectiveness.

As the research indicates, there are varying methods and approaches toward training that may impact inter-rater reliability. Moreover, the research trend tends to report that training and

calibration have positive impacts on scoring and grading. Whether that is trainer focused, rubric focused, or an enhanced collaborative approach, the consensus gleaned from the current literature heavily leans towards the positive

Opposing Views on Training

Whereas the vast amount of literature provides exceptional feedback on the positive impact of training and calibration as it pertains to inter-reliability scoring and grading, opposing viewpoints on training do exist. Researchers have questioned the impacts of training and calibration on rater perceptions and objectivity. In essence, at what cost is one removing the human from the rating.

One such viewpoint comes from a previously cited researcher, that while finding positive results in their study, also found areas of inconsistency and thus provides an alternative discussion. Papajohn (2002) performed his quantitative study at a large Midwest University and sampled approximately 500 examinees that attempted the Test of Spoken English examination. Those examinees results were rated by nine raters. The author posits the test scores are impacted by both the examinees actual performance on the exam and the raters' performance, as it relates to their interpretation and summary of the examinee's performance on the exam. The author employed rater training and concept mapping in the study to assess the impact of both training and mapping and how both impact rater-reliability. For concept mapping, the raters were trained on how to map and utilize the flowchart of concept mapping when rating the examinees. The author found after reviewing the concept maps of the raters, that the reliability amongst the raters and thus their maps, was lacking. Three of the raters were severely off in interrater reliability and the remaining six were labeled as needing retraining. The author correlates the reason for this to rater perceptions and individualism. Thus, the author posits that even though raters go

through training and calibration, there will still be the element of individual rater perceptions, and this will ultimately impact ratings. Finally, the author states that although there were differences, the idea of concept mapping as a rater tool is useful but needs to be employed thoughtfully and consistently to maximize the impact.

Another viewpoint comes from the research of Herbert et al. (2014). The authors qualitative study took place at a professional accounting institute and used a case study and focus group approach with a structured interview instrument that was utilized among nine "Markers". The focus of this study was to gauge acceptance of training on a standardized model of assessment and training on assessment. The authors found that training and calibration were welcomed, and the participants felt it would increase validity and reliability. Crucially, the authors posited questions in their study results that are worth exploring. Whereas training and calibration are seen to have positive impacts on rating, what is the cost of such calibration? Does it remove objectivity? The study is robust with data that supports standardized training and marking structure, however, the question remains, if standardization is key at what cost is the human element lost?

These two thoughtful viewpoints cast light into a shadow area that training, and calibration may present and ask pertinent questions on human objectivity. The case for training and calibration is one that is geared towards ensuring equity in scoring and grading as well as consistency. However, to what extent objectivity is acceptable is a question worth noting and importantly, worth exploring.

Summary

The positive impacts that rater training and calibration have on inter-rater reliability cannot be understated. Whether that training is focused on a rubric or the rater, the vast majority

of the literature amplifies the value of these important procedures and processes. Moreover, human perceptions, objectivity, and individualism must also be accounted for when assessing training and calibration. Unfortunately, despite the positive indices, institutes of higher education, and specifically the one being researched for this study, often do not consistently train, assess, and calibrate faculty scorers. As a result, inter-rater reliability has become a growing concern amongst the faculty and administrators.

Calibration and Scoring

The impacts of training, along with the types of training that are effective in ensuring higher rates of interrater reliability presented in this study have provided a setting for discussion. To further the discussion, a review of the literature specifically on calibration, and timing of such rater calibration, along with frequency and timing of raters scoring is presented. At present, research indicates that the timing and frequency of training plays a crucial role in ensuring accuracy and equity in scoring. Moreover, the timing of when scoring is completed by the rater has been researched and the impacts of how frequently and when scoring is completed and its effect on reliability assessed.

In a case study by Szafran (2017), a sample of 113 student written assignments were utilized in from a university's core curriculum course. The assignments were scored by two six member groups that formed scoring panels. If there was inter-reliability disagreement, a third rater was brought in. The focus of the study was to contextualize inter-rater reliability using a widely used assessment rubric and the impacts of a third grader and rater training. The author found that not only was the timing of the training crucial, but the amount of rater collaboration in the training sessions showed an increase in inter-rater reliability. His study further found that the frequency and timing of when and how to integrate the third reader can impact inter-rater

reliability. Additionally, Kayapinar (2014) in a mixed method study used a sample of 44 English Foreign Language (EFL) students and ten raters utilizing an assessment tool that rated the essay writing abilities of the students. The purpose of the researcher's study was to assess for variations or consistency in essay grading of the EFL students performed by the raters and further the discussion on rater reliability. The instruments for rating utilized by the raters were the general impression marking, easy criteria checklist, and essay assessment scale. What Kayapinar (2014) found was that utilizing the various instruments of rating always resulted in rating variations. Thus, the author posits, that there will always be rating variances regardless of instruments. However, and crucially, the research posits that calibrated raters, that are well-trained can have an impact on reliability as the calibration would lead to less ambiguity. This discussion around human objectivity is prevalent in scores of research articles, as is the impact that student assessments have on grading and scoring, which is another influencing factor that is discussed in detail within the literature review.

In a protocol paper, released by the Rhode Island Department of Elementary and Secondary Education (2021), to heighten awareness of the calibration process, "RIDE", points to the importance of such calibration as it pertains to scoring student work as well as display the impact that calibration on inter-rater reliability and scoring consistency. RIDE posits that the success of such calibration efforts and protocol implementation depends on the culture created amongst educators, specifically, the collaboration, focus, and reflective practice that goes into improving student learning, which the calibration process instills. Crucially, RIDE posits that, rubrics alone do not ensure consistency in scoring, rather, it is the calibration and training on the rubrics, and agreements that are met, that go further towards increasing reliability consistently. Furthermore, RIDE proposes that calibration is not just a tool for consistency in scoring but can

also serve as valuable professional development for educators as well as the opportunity for individual schools to gain insight into their curriculum and instruction practices. This paper, which reaches thousands of educators throughout the State of Rhode Island not only stresses the importance of calibration but also shines a light how such a practice, when employed consistently serves to develop educators while ensuring equity in grading.

As the research indicates, calibration is an important aspect that has shown to have a positive impact on inter-rater reliability. Moreover, when implemented in a timely and consistent manner, rater calibration also serves a method to provided professional development opportunities and expand collaboration efforts amongst raters, namely faculty. Calibration is not the only answer nor sole solution. A fully calibrated rater can inherently become inconsistent in grading or rating for many reasons. One such reason why a calibrated rater may still be inconsistent is the blocks of time they utilize in a day in scoring, or gaps in scoring. These factors, such as grading fatigue, have also been researched and are shown to significantly impact inter-rater reliability.

Timing of Scoring

The timing of scoring in the research sets out to explore the influence that length of time spent scoring and/or gaps in scoring has on inter-rater reliability. The literature presented displays a fascinating look at how timing influence's reliability and provides for actions that may be adopted to lessen the impact. Finn et al. (2018) undertook a very relatable study, as it pertains to gaps in scoring and non-structured timeframes for completing scoring. In their study, the authors research the impacts of time between scoring and scoring accuracy. The authors' data was derived from 350 raters of the Graduate Record Examination (GRE) general test. The results showed that gaps between scoring were found to decrease accuracy validity and reliability

as well. However, the number of consecutive days scoring did not impact scoring validity or reliability. Further, the authors found that the time spent scoring each response was a significant predictor in accuracy validity and reliability.

Wendler et al. (2019) sought to answer two questions, the first, does rater calibration impact scoring accuracy, and the second, does reducing the frequency of rater calibration impact scoring accuracy. The authors completed a comprehensive literature review on rating variability and reliability to help guide their hypotheses testing. Controlling for rater drift, also referred to inter-rater reliability was a major focus of their study along with calibration efforts. In their study, a total of 46 raters were utilized. The raters rating reliability scoring of the "GRE" essays were analyzed along with the length of time they took to score the essays. The raters were given five days to complete scoring, with six hours each day to score. Each day, there were between 124 and 134 essays provided to score. The researchers found that inter-rater reliability was remained steady amongst raters throughout days one to three, day four however was a turning point, the researchers reported significant reductions in reliability. The researchers posit that calibration sessions did have a positive impact on inter-rater reliability, however, grading fatigue appeared to play a role in the drop off of reliability. Further, upon completion of grading, the raters were sent a survey about their experience. Gleaned from this survey was data regarding rating experience. The researchers found that the more experienced raters had the most reliability in scoring. Whereas the researchers did not find an answer to frequency of calibration, they do posit that calibration has positive impacts as expressed in the data outcomes and findings.

Ling et al. (2014) performed a study on the impact of fatigue on raters. This quantitative study analyzed more than 14,000 audio responses to four Test of English as a Foreign Language

"TOEFL" examinations. The scores of 72 raters were utilized. The researchers sought to study the impact of fatigue that occurs amongst scoring for long periods of time over several days. The researchers' findings suggest that rater accuracy and reliability are more consistent when raters score in short shifts or sessions and there is greater rater productivity in shorter shifts as well. After the scoring was completed, the researched asked the raters to complete a survey. The survey results indicate high levels of fatigue for those raters scoring in longer shifts, especially as those shifts entered the afternoon hours. The authors posit that increase reliability and consistency, raters should be trained to complete scoring in shorter shifts, as an example scoring every two hours and then resting. Implementing this type of strategy, as the authors state, could reasonably lead to less human error and greater inter-rater reliability and scoring consistency.

Finally, Pownall et al. (2019), completed a mixed method study discusses potential influences in the grading process. As the researchers assert, these influences impact the rater and therefore the learning experience for the student, including grade inflation. A survey was deployed gauging cognitive influences on grading to 157 educators. The authors found that grading was impacted by several factors including previous grades to a student or exposure to the student and thus not following the scripted rubric. Further, the researchers found that grading fatigue was also another factor as displayed by inconsistent grading as more scoring occurred. The researchers suggest that raters break up the timing of their grading into smaller sessions and to ensure that the institution allows enough time for raters to sufficiently grade based on the full scope of the rubric, allowing for smaller period grading as well.

Summary

Calibration and the length of time of scoring, including scoring gaps, significantly impact inter-rater reliability. The studies presented advance the understanding that consistent and

collaborative training and calibration establish a solid foundation towards reducing scoring differences. Further, the studies highlight the importance of grading fatigue and present strategies to mitigate such fatigue.

At the Institute, the policy of grading the comprehensive examination is to set a deadline as to when raters are to submit their finalized grading sheets. The majority of grading sheets are submitted by faculty raters on the final deadline day. This leads one to believe that faculty may be doing "marathon" grading sessions to get the grading sheets in before the deadline. As the research indicates, grading fatigue does have negative impacts on inter-rater reliability. With that, one possible avenue for future research, upon conclusion of this study, would be to delve into when the raters submit the grading sheets and gaps between submission. This could provide a critical data set to report out on as part of additional action research. Some limitations do exist as raters tend to either send their scores in bulk at one time or as they complete. Which, then can impact the reliability of the data is it pertains to when the scoring occurs. However, as the scores are logged on individual excel files, "last date saved" data could be extracted. Nonetheless, it would be interesting to explore.

Faculty Status and Scoring

The literature asserts that there are multiple factors that impact inter-rater reliability. Thus far, we have traversed the inter-rater reliability landscape through the lens of training, types of training, calibration, viewpoints on human objectivity, timing of calibration and length of time spent scoring. Another factor that is presented in this literature review is the role that faculty status has on scoring, and grade inflation. There is significant research on the impact that faculty status has on scoring. Extensive studies highlight how faculty status (tenure, tenure-track, non-tenure) plays a role for those raters in scoring. Moreover, studies have researched grade inflation

as it relates to faculty status as well as delving into the psychology as to why non-tenure and tenure-track faculty tend to either inflate grades or grade lower.

In one study, the researcher found a difference in grading/assessment between tenure and tenure track faculty. Where the latter grade more lenient then the former. Moreover, the authors posited that reciprocity in grades and instructor evaluation could be present, thus concluding a tenure-track professor be more lenient in grading with the expectation of receiving a better instructor evaluation from his/her students (Filetti et al., 2010). Additionally, Johnson (2011) completed a study involving 2,008 freshmen at a research university and grading data derived from both full-time faculty and adjuncts, the researchers sought out to investigate correlations in grade inflation. The researcher utilized multiple statistical methods and data grouping to provide an extensive analysis of faculty grading. The researcher found that students that were taught by adjuncts received higher grades than those taught by full-time faculty. The author posits that whereas grading differences are apparent, there is a need to review why the grading differences occur. Much like several other researchers have found, student evaluations and adjuncts desire for higher evaluations must play a significant role in the grading differences.

Mcarthur (1999) completed a study took place in small eastern community college with average to low student enrollment. The purpose for this study was to assess the impact that adjunct faculty have on grading in relation to their full-time professor counterparts. Courses in the subject areas of humanities were utilized along with the course instructors that comprised of both full-time professors and adjunct faculty, six for the former and twelve of the latter. The course grades that were used in the study were quantified over three spring semesters and the class sizes varied from a maximum of 35 students. The researcher's intent was to argue against prevailing research that posits adjuncts bring standards down. Upon data analysis, the researcher

found an association that exists between grades and faculty status. Much like other research that has been performed, this study indicated that students are more likely to receive a grade of "A" from an adjunct than full-time professor. Further, the author suggests a possible reason for this, which furthers the literature on student evaluations impacting adjunct grading behavior. Moreover, in a study by Kirk et al. (2009) which analyzed the grades of 2,597 students in introductory finance courses at a business school, the researchers sought to test two hypotheses, the first, that grading by full-time faculty and adjunct faculty would not differ significantly and the second that student achievement in an introductory course taught by full-time faculty or an adjunct would not correlate to higher achievement in the subsequent course. The researchers found differences in grading and achievement in both cases as they relate to differences between full-time and adjunct faculty. Crucially, the researchers found that adjunct faculty had consistently assigned higher grades to students more so than full-time faculty. Moreover, the researchers found that students in the introductory course taught by full-time faculty performed significantly better in the subsequent course than those taught by adjuncts in the same course. In another stark finding, that peers through the lens of student retention or desired major, the researchers found that students that took the introductory course taught by adjuncts were less likely to choose that corresponding accounting major.

Furthermore, Moore et al. (1998) examined the relationship between faculty status and grading. The researchers utilized 417 introductory courses where were classified as "100 level" courses at a western university that has a student population close to 10,000. The researchers sampled courses from various majors and from instructors of all ranks and tested for GPA of the class after final grades. Further, the researchers analyzed 57 courses that were taught by professors, 38 by associate professors, 67 by assistant professors, 143 by instructors and 112 by

teaching assistants. The results of this study indicate overall that tenured instructors (of different rank) comparatively have lower GPA's results amongst each other. Crucially however, the comparison in GPA from tenured, regardless of rank and non-tenured faculty was dramatic. The GPA of tenured faculty was 2.54 were the GPA for non-tenured was 3.04. To which the authors report confidently non-tenured faculty give higher grades or grade inflate. A troubling hypothesis that the researchers posit is the assumption that non-tenured faculty give higher grades to "buy" better student evaluations, thus allowing for job security. The authors make note that institutions utilize student evaluations heavily when it comes to making judgements on teacher performance. The authors assert that administrators need to find alternate ways of evaluating non-tenured faculty should this trend continue. This assertion is a common theme within the literature as it implies student evaluations play a strong role in grading outcomes, specifically those grades provided by adjunct faculty.

Grade Inflation

Moore et al. (1998) provides an on-ramp study that provides a critical lens on grade inflation and the impact that student evaluations have on grading. Kezim et al. (2005) in which they reviewed grade point averages over a 20-year period to investigate grade differences in mane GPA among adjunct, untenured, and tenured faculty. Their results showed adjunct faculty tended to grade higher than tenured and non-tenured track faculty. The authors posited adjuncts grade higher to score better on after-class faculty assessments completed by students. Deboer et al. (2007) completed a quantitative study utilizing survey research. The instructors were surveyed using an assessment tool. The focus of the study was to assess the impact of instructor attitude on grading behavior. The results supported the authors hypothesis that rater personality, based on the assessment, plays a part in grading. Significantly, the authors found that a higher

need for approval resulted in lower grades. The authors posit that the lower grades and need for approval reflect attitudes of instructor's desire to attain approval from their fellow colleague instructors. Further, the author's data suggest that full professors assign significantly fewer higher grades than academic staff and the authors posit full-time professors have less worry about job security therefore are more apt to grade harsher.

Nikolakakos et al. (2012) research provided a comprehensive review of the cause of grade inflation. The researchers study utilized 235 students' final grades provided by 25 professors enrolled in a graduate degree program at a northeast college. The researchers' purposes were to first determine if the grade earned was perceived by both student and faculty as truly reflecting their academic achievement and second, determine the causes of any possible differences in perception between the final grade and academic achievement. The authors posit that grade inflation is a serious issue that distorts performance and has long term affects especially for students that are being instructed in the field of teaching. Further the authors suggest factors that lead to grade inflation, such as student evaluations of the instructor, faculty status, faculty personality, and lack of uniform grading practices. These, as the authors state, lead to grade inflation and ultimately a weaker work force. The results of their study indicate very alarming trends. First, the authors found that based on survey data, professors gave grades to students higher than what they earned as they feared poor student evaluations. Further, the authors found, in support of finding surrounding student evaluations, that 90 percent of students taught by adjuncts received a grade of "A". Crucially, faculty response to the survey also revealed that job security was a factor in grading and that giving grades lower than an "A" would result in student complaints. The authors further found differences in faculty rank when it came to opinions on course building. The faculty opined that if a course was created by a faculty

member without a terminal degree, then that course would inherently be less demanding and thus the outcomes would be inflated. Alternatively, the authors found that most students who completed the survey, felt they had earned the grade they received, with a small portion stating that they believed some of their peers were perhaps less deserving based on their rigor in the course. As a result of this study, the authors stress the important findings of faculty status and its impact on grade inflation.

In another comprehensive an important study Izienicki et al. (2019) not only discusses grade inflation, but also the use of extra credit as a tool for grading and gender. Their study utilized data derived from the national online extra survey of college-level sociology instructors. The data sets came from 100 four-year universities and 50 two-year colleges. A total of 978 faculty and 247 students comprised the sample data set for the researcher's survey. The researchers found that the use of extra credit was common. Further, they found women instructors were more likely to offer the opportunity for extra credit than their male counterparts. This study is significant not much from the standpoint of the use of extra credit, but more to why the use of extra credit was so common, especially among women. The researchers posit possible explanations being, student evaluations, credibility, and status. Further, the researchers posit because promotional opportunities and pay increases are tied to student evaluations, that some use extra credit opportunities to enhance the possibility of favorable student evaluations. Another important finding in their research pointed to years of experience or tenure and the use of extra credit. Those faculty with tenure or more years of experience tended to not provide extra credit, thus, a grading difference is more likely to occur between more experienced/tenured faculty and adjunct faculty.

Questions arise in the literature on how to combat grade inflation, and in one case, is grade inflation just a myth. Boretz (2004) undertook a case study seeking to demystify the notion of grade inflation, as it relates to faculty rank and grade inflation as well as the assertion that student evaluations play a key role in grade inflation. The author performed their research utilizing case study and literature review. In the study the author found several incomplete theories as to why the myth of grade inflation exists. Crucial to the authors argument is that most studies need to dive deeper into the research and find out truly if grade inflation is correlated to faculty status, student evaluations, or if the research is in essence, incomplete to the extent that one should not make broad stroke statements on findings. However, the author does acknowledge the existence of the possibility of grade inflation and certainly the data that might well indicate a correlation between faculty status. As a result, the author provides solutions to the myth, and steps in how to curtail the promulgation of it. First, as in the case of previous literature, a review of the use of student evaluations as a tool to assess faculty. Second, administrators accepting that student evaluations will differ and that adjuncts in most cases cannot control for some factors accounting for differences. Finally, and as the author points out most important, there is a need to review the content of the student evaluation, where it will allow the instructor to focus more on the goals of the teaching endeavor.

Further examples in how to combat grade inflation and inequality are provided by Korpan (2020). The researcher suggests several methods to lower grade inflation and grade inequality. First, ensure instructors are sufficiently cognizant of grading policies. Second, encourage instructors to provide rubrics to students on assignments. This helps ensures transparency on how the student's assignment will be evaluated and also assist increasing consistency and reliability in the grading process. Third, the researchers suggests that student

evaluations include both formative and summative assessment sections. Finally, analyzing the student evaluations and checking for validity and alignment with the learning objectives for the course.

Summary

As the research displays, faculty status is found to have an impact on scoring. This is significant research and pertains well to this study as we have an ample of amount of data as well as faculty statuses that can be tested again hypothesis on impacts of faculty status and comprehensive examination scoring. Further, the Institute has several faculty members that have been employed for 20-years or more. Those factors will certainly provide a wealth of data and discussion points.

Solutions and Alternatives

Throughout the review there has been several solutions posited as well as alternative foci as it relates to inter-rater reliability, training, and calibration. Rubric training in which assessment is narrowed and focused was shown to increase inter-rater reliability (Attali, 2013). This research finding and several other trend towards not only to slimmed down rubrics, but also rater training, calibration, and rubric creation that is done in a more transparent and collaborative nature. Bamber (2015), while using student as participants in trainings, found that having a transparent and collaborative approach towards calibration increased inter-rater reliability. Further, the participants also felt an increased self-awareness of what was being assessed and positive feelings towards the assessment. This is an important study and one that could be utilized as an approach deployed by the Institution or studied further as it asks a question of whether students would perform better if they were fully informed, in detail, on how the rubrics are set and what is being assessed. Lee (1985) in her case study review of rater training

suggested similar conclusions in that the training should be concise and transparent as possible to increase reliability in practices.

An alternative to rubric training is offered up by Bernardin and Buckley (1981) in which they argue that the focus of rater training should not be on the rubric and calibration, rather, it should be on the raters themselves, exploring their biases and tendency in grading and how to best calibrate against those. Herbert et al. (2014) asks a crucial question as part of their study, which is what at what expense does training and calibration have on the objectivity of the rater. Further, Papajohn (2002) posits that even after training a rater, one should note that raters may still not interpret and score in similar fashion, although, finding that experienced raters tend to rate more similar.

Summary of the Literature

The literature surrounding inter-rater reliability, training, calibration, solutions, and alternative approaches is robust. As such, inter-rater reliability continues to be a focus area for research. With such in-depth research with various methods deployed, the one outcome most researchers have agreed upon is the importance that training, and calibration play on positively impacting inter-rater reliability. Whereas there are researchers that find varying results on the impact of training and posit that human objectivity will always play a role regardless of training, those researchers also assert that training is important and needs to be part of a consistent and comprehensive plan for development.

Calibration training has proven to have a positive impact on reliability and thus a step towards ensuring grading equity. Additionally, burgeoning research indicates grading fatigue also impacts reliability, in a negative sense. As a result, researchers have provided guidance on

how to combat grading fatigue while ensuring reliability. That guidance in the form of action research can prove to be critical in the cases of institutions like the one being researched.

Furthermore, additional, and continuous research has also provided the field with areas improvement that need to be addressed. Concern circles around the impact that faculty status has grading. Cognitive research and need to achieve research have shown that adjuncts do indeed grade differently than their full-time faculty counterparts. This impact has then led to studies on grade inflation, crucially, the ever-growing body of research that indicates student evaluations of adjunct faculty plays a role in grading, in what one researcher referred to is a grade reciprocity.

What remains uncertain is how institutes of higher education can assure training and calibration occurs, while faculty juggle their day-to-day tasks as teachers, advisors, mentors, and faculty administrators. Additionally, with research pointing towards the need to assess the student evaluation of faculty, what actions can and should institutes of higher education take both short and long term, to address the notion of grade reciprocity in ensuring student outcomes are earned and grading is equitable.

CHAPTER 3

Research Methodology and Methods

Introduction

The Comprehensive Examination

At the Development Institute, a small, private graduate school in the Midwest region of the United States, a comprehensive examination is administered to students upon completion of all degree coursework requirements. The comprehensive exam has been administered for over 50 years and is viewed as the capstone to the students learning experience at the Institution. The examination is the highest stake component of the culminating requirements of the master's degree program. Students that successfully pass the comprehensive examination are eligible for degree conferral. Therefore, the stature of the comprehensive examination is matched only by the anxiety students feel as the day of the exam approaches.

Components

The comprehensive examination is a five-hour, written examination that covers major content areas within the curriculum. The exam is administered three times a year, once in late April, another in early June, and the final in early September. The examination questions are developed by representatives from the faculty assessment committee. Two faculty members, traditionally one tenured and one tenure-track faculty write the six-question examination near the beginning of the spring semester. Students are required to apply to sit for the examination.

Administration

The examination, pre-COVID, was traditionally administered in-person on a Saturday morning. Students are provided with desktop testing stations, paper, and pens for notes as well as earplugs as needed. Proctors are in each room to ensure examination integrity. Students are

not allowed use of the internet and are only allowed to bring in beverages and snacks. Leaving the testing station is allowed for restroom breaks. Upon completion, the student is instructed to leave all notes and testing material at their station. The proctor then removes the comprehensive examination from the desktop and saves it in a secure drive. Current state (during COVID), all students are administered the exam via a course in the Institutes' learning management system. The course is made available three days before the start of the exam and contains day of exam instructions, a writing template, and contact information. The exam is not synchronously live proctored, meaning proctors are not logged in viewing the students while they attempt the exam. Students are on an honor code in that they are instructed not to use any electronic devices nor the internet while attempting the exam. The exam is submitted through the grading center and the exam administrator pulls the exam, saves it a file that removes author data, to ensure blind grading. In both cases, students with approved accommodations are provided various supports such as extra time, multiple days for completion and/or specialized hardware to complete the exam.

Scoring

Upon completion of the examination, the exam administrator emails the comprehensive exam, along with the grading rubric, score sheet and deadline to complete their scoring to the designated faculty rater. Faculty will then send in their completed score sheets to be tabulated. A master scoring sheet, that contains the artificial student identification number that ties back to the real student identification is utilized to tabulate scores and flag for a possible third reader, as necessary. Similar strategies were employed by researchers that also set out to study inter-rater reliability. In Szafran (2017), a sample of 113 student written assignments were utilized and the assignments were scored by two six member groups that formed scoring panels. If there was

inter-reliability disagreement, a third rater was brought in. The focus of the study was to contextualize inter-rater reliability using a widely used assessment rubric and the impacts of a third grader and rater training.

Once scoring is completed, students' scores are reported out to their assigned advisors as pass or fail. The advisor is then charged with contacting the student, via phone, and informing them of the results. If the student fails the examination, they will meet with their advisor to create a strategy to prepare for the second attempt. Subsequent failed attempts require more in-depth advising and planning.

Inter-Rater Reliability, Training, and Rater Calibration

As previously mentioned, at no point, over the past four years and twelve comprehensive exam administrations has inter-rater reliability been assessed, nor formal or consistent rater training and calibration taken place. Talks of the need to do so have come up sporadically throughout that time. However, no action has taken place. As the institute's headcount continues to grow, so too does the amount of exams administration and ultimately, the need to ensure inter-rater reliability.

Similar Research on Training and Calibration

The existing research and literature suggest a positive correlation between training and calibration and increases on inter-rater reliability. Gardiner et al. (2020), in a longitudinal case study, assessed the impact that rubric-based training has on inter-rater reliability. The authors findings point to positive impacts and shared findings and results of previous researchers in the value of training and calibration, specifically, that calibration and rater training is a proven method in increasing inter-rater reliability. Additionally, Gugiu et al. (2012) performed their research utilizing rubrics as building blocks for training. Their quantitative study utilized 29

undergraduate students enrolled in an introductory political science course at a university located in the Midwest. The students were randomly assigned into groups and two raters were provided grading rubrics. Importantly, the rubrics created had clearly defined measurements and instructions on what to assess and how to grade. Upon analyzing the results of the grading, the researchers found significant reliability in scoring between the raters. The authors posit that the impact of a clear grading rubric along with a smaller number of raters positively impacted reliability.

Research Problem, Purpose, and Questions

Research Problem

There is a vast body of literature documenting the positive impacts that rater training and calibration sessions have on inter-rater reliability. In a study examining efficiency in assessment, researchers trained both students and faculty on how to rate essays. The results showed that the trained students rated as reliable as the faculty and the lack of rater-training is attributed to negative inter-rater reliability (Cole et al., 2012).

Moreover, research indicates that the frequency and timing of training plays a crucial role in ensuring accuracy and equity in scoring. Finn et al. (2018) researched the impacts of time between scoring and scoring accuracy. The results showed that gaps between scoring were found to decrease accuracy validity and reliability as well. Additionally, the authors found that the time spent scoring each response was a significant predictor in accuracy validity and reliability.

Furthermore, there are increasing amounts research that have been completed regarding faculty status and its impact on grading and grade inflation. Deboer et al. (2007) assessed the impact of instructor attitude on grading behavior. The results supported the authors hypothesis

that rater personality, based on the assessment, plays a part in grading. Significantly, the authors found that a higher need for approval resulted in lower grades.

Unfortunately, despite the positive indices that highlight the impact of training and calibration, institutes of higher education often do not consistently train, assess, and calibrate faculty scorers. Additionally, although there is an overwhelming amount of research that indicates the impact that faculty status, and student evaluations have on grading, actions to mitigate these impacts are not prevalent in the literature.

Research Purpose

This study will describe the impact that consistent training and calibration has on inter-rater reliability. Further, it will examine the relationship between faculty status and scoring as it pertains to inter-rater reliability.

Research Question

Does the lack of rater-training and calibration and faculty status explain the relationship between faculty scoring grading reliability on the comprehensive examination?

The following research hypotheses will guide the study:

- **Hypothesis 1:** Is inter-rater reliability negatively impacted when no formal or consistent rater training is provided?
- **Hypothesis 2:** Will the scoring of tenured faculty be lower than tenure-track, clinical faculty, and adjunct faculty?
- **Hypothesis 3:** Will the scoring of recently hired faculty be more reliable than faculty that has been hired over ten years.
- **Null Hypothesis:** Inter-rater reliability is high and consistent amongst all faculty.

- **Hypothesis 4:** (After action) Will inter-rater reliability significantly increase among trained scores regardless of faculty status?

Research Methodology

Quantitative Methodology

Recent studies have guided the research and methodology for this research study. Filetti et al. (2010) completed a quantitative method study utilizing derived data analysis. Data was derived from 595 student's grades for 33 sections of an English course taught by 12 professors were utilized. Students and professor data was collected from a small liberal arts University. The focus of the study was to assess whether tenure affects grading/assessment. Additionally, Kezim et al. (2005) employed a quantitative study and used data derived analysis. Grade Point Averages "GPAs" data was derived from a small undergraduate college. GPAs were analyzed over a 20-year period. The focus of the study was to investigate grade differences in mean GPA among adjunct, untenured, and tenured faculty.

This study draws on quantitative research methods to analyze and interpret comprehensive examination scoring. The data utilized was derived from 16 comprehensive examinations that have taken place over the past 5 years and scored by 25 faculty members. The data sets included the resulting scores of 397 examinations and a total of 794 unique exam scores provided by faculty. Scoring on the comprehensive examination ranges from 0 to 40, with 0 being the lowest possible score and 40 being the highest possible that can be given be each individual rater. The student examination attempts are deidentified of all personal data. The 25 faculty members are identified in the research by a random faculty identification number, faculty status, and date of hire. The deidentification of both student data and faculty data was completed prior to the me having received the raw data. Deidentification was performed to ensure

anonymity and provide an additional check against bias towards student or faculty due to my research positionality. Inter-rater reliability and hypothesis testing is performed utilizing two methods. The first, method is in the form of descriptive and inferential statistics. The second method will be the utilization of a Likert-type scale. See Table 1 for scale. As the data set was derived from existing data, no additional collection methods were employed.

Action Research & Qualitative Methods

Active learning, which is rooted in the theory of constructivism is being utilized to frame the process undertaking in employing action research within this study. Constructivism is a theory in education that asserts learners *construct* knowledge actively, as opposed to passively, by integrating new understandings with previous understandings (Creswell & Creswell, 2018). Active learning theory/Constructivism has been linked to Jean Piaget and his theories of cognitive development. Piaget's theory discusses how humans acquire, construct, and utilize knowledge ("Piaget's Theory of Cognitive Development," 2021).

As applied to this study, active learning theory/constructivism holds that one should expect training and calibration sessions, if implemented effectively and utilized consistently, to positively impact rater scoring and grading and thus inter-rater reliability. In a study by, Hung et al. (2012) on rater effects and the impact that tools utilized to train raters have on inter-rater reliability. The authors posit that reliability is impacted by focused training and moreover, raters that were not calibrated on the scoring rubric were shown to have less reliable scores.

The action research component of this study was employed after data analysis and interpretation of faculty scores. For this study, a voluntary focus group solicitation email was sent out to all faculty to discuss the results of the findings and discuss actionable next steps. Next steps conceptualized were to include updating comprehensive examination training and

calibration policy and processes as well as the use of the concept mapping. PapaJohn (2002) undertook and found that the use of concept mapping in training had an impact on reliability amongst more experienced raters. The concept mapping design process was conceptualized to not only train and calibrate raters but also with the goal of ensuring that raters were all conceptualizing what was being assessed and how it was being assessed.

Research Design

The goals of this study were to evaluate inter-rater reliability scoring amongst faculty on the comprehensive examination and, after data analysis, create an actionable plan in collaboration with faculty that integrates the findings towards meaningful change in training and calibration. To accomplish these goals, hypotheses were established that linked previous research findings, as identified through the literature review, and expands upon those as they pertain to inter-rater reliability and the impacts of faculty status and length of employment on scoring reliability. As mentioned, deidentified student scoring attempts were created and profiles for each faculty member for each exam iteration. The profile consists of a unique identification number, faculty status, and length of employment at the time of scoring. Theses profiles assist in categorization of scoring and faculty status in testing against the hypotheses. Additionally, I created a reliability Likert type scale to further define and determine inter-rater reliability, see Table 1 for an example of the "Comprehensive Exam Reader Scoring Agreement" as it pertains to the Institute which assists with the disaggregation of data year by year. After data analysis and hypotheses testing, the results are shared by presenting these findings to faculty in the form of a focus group. The focus group would then formulate an actionable plan to address training and calibration amongst faculty scorers to ensure great inter-rater reliability moving forward.

Summary

Seeking to answer the research question, does the lack of rater-training and calibration and faculty status explain the relationship between faculty scoring grading reliability on the comprehensive examination, I have utilized quantitative methods making use of derived data analysis as well as action research employing focus group methods. The data from a small midwestern private graduate school, was derived from 16 comprehensive examinations that have taken place over the past 5 years 2017-2021 and scored by 25 faculty members. The data sets included the resulting scores of 397 examinations and a total of 794 unique exam scores provided by faculty. Quantitative methods were employed to test for significant differences and evaluate inter-rater reliability. Additionally, those methods will allow for the analysis of correlations against the hypotheses. The results of this study will be tremendously useful as a large portion of the research is dedicated to analyzing data about tenured, tenure-track, non-tenure/adjunct faculty scoring on the comprehensive examination. That analysis will guide the action research portion of the study with the goal of the Institute adjusting policy and processes to ensure increased inter-rater reliability.

Research Context

Setting

This research study took place at a small Midwest region private graduate school. The "Institute", herein, is in an urban setting and consists of one of campus. The Institute focuses its curriculum on developmental theory and has a total student population under 1,000. Graduates of the Institute go on to hold positions in the fields of social services, education, non-profits, and clinical settings.

Participants

The participants of this study consisted of 25 faculty members of varying status, tenured faculty, tenure-track faculty, clinical faculty, and adjunct faculty as well very lengths of time of employment at the Institute. The scoring results of the comprehensive examination attempts of 397 students over the span of five years were utilized from the 25 faculty members.

Participant Recruitment and Selection

As the data set was derived from existing data, no recruitment and selection were required for the data analysis portion of this study. Participant recruitment for the voluntary focus group and data presentation was utilized. An email was sent to all faculty providing an overview and abstract of the research performed and intended outcomes of the focus group study which includes a discussion on the findings, dialogue around the comprehensive examination as a tool of graduate student assessment, and sentiment regarding next steps in formulating policy and processes around rater training and rubric calibration moving forward. Faculty that agreed to participate were provided with discussion prompts on how to approach training and calibration of the comprehensive examination and along with the current rubric that is provided to faculty upon prompting to score an examination.

Research Methods

Quantitative Data Collection and Analysis

The data utilized was derived from 16 comprehensive examinations that have taken place over the past 5 years and scored by a total of 25 faculty members. Data collection of the comprehensive examination attempts and subsequent was performed from established records for the calendar years of 2017, 2018, 2019, 2020, and 2021. Faculty data collection and subsequent profile creation was created utilizing established onboarding documentation provided by the Chief Human Resources Officer. Prior to the data being sent to me, deidentification of all data

was performed to ensure anonymity and provide an additional check against bias towards student or faculty due to my research positionality. Upon receipt of the data, coding and data alignment was undertaken to ensure that descriptive and inferential statistical analysis could be performed within SPSS and test thus complete hypotheses testing.

Strategies

The student examination attempts were deidentified of all personal data. The 25 faculty members are identified in the research by a random faculty identification number, faculty status, and date of hire. Scoring on the comprehensive examination ranges from 0 to 40, with 0 being the lowest possible score and 40 being the highest possible that can be given be each individual rater. The deidentification of both student data and faculty data was completed prior to the me having received the raw data. The student examination attempts were deidentified of all personal data. Faculty profiles of deidentified data sets were created to utilizing a random faculty identification number, faculty status, and date of hire. Upon receipt of the data, coding and data alignment was undertaken to ensure that descriptive and inferential statistical analysis could be performed within SPSS and thus complete hypotheses testing.

Timeline

Data extraction, coding, and initial analysis within SPSS was undertaken throughout the months of November 2021 through April 2022. The timeline aligned with pertinent course objectives and progression through the program. Full data analysis in the forms of descriptive statistics along with inferential statistics, including, “independent samples *t*-tests” and “intra-class correlation coefficient tests”, as well as the Likert-scale “Comprehensive Exam Reader Scoring Agreement” analysis of data occurred through the months of February 2022 to April 2022. Initial write-up and discussion of the results was completed 30 days after the completion

of data analysis and hypotheses testing. The write-up was completed so that it could be utilized for the qualitative action research component of the study. The commencement of the action research component took place in late April 2022. An email solicitation to meet and discuss the preliminary findings in the form of a formal presentation that enacted a meeting of the faculty assessment committee in early May 2022. During that time, the committee was presented with the research data, findings, and next steps in the hopes of testing hypothesis 4. Final recommendation of policy and process updates were drafted but not delivered to the Institute's executive leadership and faculty assessment committee nor was formal rubric training and rater calibration during that time. The goal of this timeframe and portion of the research was to complete the training and calibration session before the June 2022 comprehensive examination. Unfortunately, neither occurred as the faculty assessment committee did not move forward with the recommendations presented during the May 2022 committee meeting where hypotheses testing results 1, 2, and 3 were presented. Additionally, in early June 2022 I was informed that my position at the Institute was being eliminated, and I subsequently resigned in August 2022. Prior to my resignation I was afforded the opportunity to test the June 2022 comprehensive examination on its own with the benefit of previous years data.

Data Analysis

Data analysis was performed utilizing SPSS with descriptive, frequencies, and inferential statistical analysis measures being employed. The Likert-type scale was utilized once mean testing outputs were established. Intraclass correlation coefficient statistical analysis was utilized to determine inter-rater reliability overall between Readers (1) and Readers (2) of the comprehensive examination from 2017 to 2021. Intraclass correlation coefficient statistical analysis was also utilized, to determine reliability amongst faculty readers that were employed

from one to nine years, and faculty readers that were employed ten years or greater. Further, hypothesis testing, utilizing inferential statistics by the way of *t*-tests were performed to evaluate if significant differences between faculty scores exist overall and to test for significance amongst the scores of faculty with different statuses. Data source analysis plan in relation to individual hypotheses testing is described below.

Hypothesis 1

Is inter-rater reliability negatively impacted when no formal or consistent rater training is provided? The scores of faculty Readers 1 (R1) and faculty Readers 2 (R2) were analyzed via SPSS for frequency and descriptive analysis. Intraclass correlation coefficient statistical analysis was utilized to determine inter-rater reliability overall as well. Additionally, an independent samples *t*-test was utilized to determine statistical significance between the groups. The mean scoring differences outputs were then tested utilizing the Likert-type "Comprehensive Exam Reader Scoring Agreement" to complete a scoring gap analysis. The null hypothesis for this data set is inter-rater reliability is not significantly impacted when no formal or consistent rater is provided.

Hypothesis 2

Will the scoring of tenured faculty be lower than tenure-track, clinical faculty, and adjunct faculty? The scores of Readers 1 (R1) and Readers 2 (R2) were analyzed via SPSS for overall mean scores differences, individual mean scoring, and overall mean paired scoring along with the readers faculty status at the time of scoring. An independent samples *t*-test was utilized to determine statistical significance between the groups. The null hypothesis for this data set is faculty status does not significantly impact scoring.

Hypothesis 3

Will the scoring of recently hired faculty be more reliable than faculty that has been hired over 10 years? The scores of Readers 1 (R1) and Readers 2 (R2) were analyzed via SPSS for overall mean scores differences, individual mean scoring, and overall mean paired scoring along with the readers length of employment at the time of scoring. An independent samples *t*-test was utilized to determine statistical significance between the groups. To analyze correlations in scoring amongst faculty length of employment an intraclass correlation test was run. The mean scoring differences outputs were then tested utilizing the Likert-type "Comprehensive Exam Reader Scoring Agreement" to complete a scoring gap analysis. The null hypothesis for this data set is years of employment does not significantly impact scoring.

Hypothesis 4

Will inter-rater reliability significantly increase among trained scores regardless of faculty status? After the data presentation to the assessment committee was completed, the faculty was recommended to complete a formalized rubric training and calibration. The goal of this training and calibration session was to not only test hypothesis 4, but also begin to strategize and create a policy and process for consistent training and calibration moving forward. As the assessment committee did not complete the training and calibration session prior to the June 2022 examination as recommended, hypotheses 4 was unable not be officially tested. However, to further the discussion on the impacts of rater training and calibration, the June 2022 comprehensive examination data was analyzed on its own. The scores of Readers 1 (R1) and Readers 2 (R2) were analyzed via SPSS for overall mean scores differences, individual mean scoring, and overall mean paired scoring. Intraclass correlation coefficient statistical analysis was utilized to determine inter-rater reliability overall as well

Procedures

Upon receipt of the data, coding and data alignment was undertaken to ensure that descriptive and inferential statistical analysis could be performed within SPSS and thus complete hypotheses testing. The coding consisted of the student identification number/record, corresponding score, reader 1 and reader 2 random faculty identification number, faculty status, and date of hire. There were multiple sets of data created to assist with coding and alignment and ultimately analysis. One set contained all examination attempts and scores, and faculty reader profiles, regardless of the year that the examination has taken place. This was data set was labeled "all years". Additional sets of data were created by year, example "2017", "2018", and so on. The reasoning that the "all years" data set was created was to provide increased reliability testing for the overall study while performing descriptive and inferential statistical analysis. The individual data sets were created not only for statistical analysis as mentioned with the "all years" data set, but also to disaggregate the data and identify trends in scoring on a more granular level and progression, or regression of inter-rater reliability scoring through 2017-2021.

Table 1

Comprehensive Examination Reader Scoring Agreement	
Scoring Differential Range Among Readers	**Labels**
0-3	High Agreement
4-7	Moderate Agreement
8-11	Low Agreement
12 >	No Agreement

Qualitative Data Collection and Analysis

The participatory action research phase of this study commenced upon conclusion of the initial quantitative data analysis. As mentioned within the quantitative analysis section, I completed the initial data analysis as expected and moved forward with the intended participatory action research phase where I present the data and findings to faculty. The goal of

this phase was two-fold, first to gauge the faculty's responses to the data presented while the discussing the comprehensive examination as an assessment tool, and second, to work collaboratively on an action plan towards creation of policy and process as it relates to training, review, and assessment of the comprehensive examination rubric.

Researcher Positionality

As a first-generation college graduate, I understand the importance of higher education and the impacts that attaining a degree have on one both personally and professionally. A large population of students that attend the Institute are also first-generation students. These students seek out graduate degrees that are focused in the areas social and behavioral services. These focuses then turn into careers that support the most vulnerable in our society. The learning and work the students put in towards such self-less roles cannot be understated.

As such, it is a pleasure for me to have the opportunity to serve these students in the capacity that I do. As a student services professional at the Institute, I am charged with assisting students throughout their entire graduate school career. Whether that is providing clarity about the programs offered, or providing guidance on Institutional processes, I serve these students as I was served during my undergrad and graduate careers. Further, as the proctor for the comprehensive examination, I take great pride in shepherding students and faculty through the process of application, to administration, to grading, and finally, degree conferral. That framework provided, I perform this research through the lens of a student, colleague, and administrator. The search for answers to ensure equity, development, and evolution is what further fuels the research I have undertaken.

Based on my positionality, which combines the roles of student affairs professional, and proctor, I am obliged to acknowledge the inherent bias I have as the researcher, one is which is

explicit to student success and equity. Additionally, my role as the chief records officer within the Institute in processing satisfactory academic progress and degree conferral must also be provided to the reader. With those acknowledgements, I draw back on the efforts made throughout this methods portion of the dissertation to ensure that those biases have been addressed from a research perspective. It is ultimately my goal that the combination of data and action will provide a comprehensive look at not only the importance inter-rater reliability and calibration, but also how collaboration efforts such as this one being undertaken can lead to actionable results the positively impact policy and process in ensuring that the Institute is student-ready.

CHAPTER 4

Findings and Discussion

Introduction

The comprehensive exam has been administered for over 50 years and is viewed as the capstone to the students learning experience at the Development Institute. The comprehensive examination is high stakes in that students must pass it to complete their graduate studies and therefore be conferred their degree. I have served as the proctor and processor scores for the comprehensive exam since April 2017. Throughout this time, I have noticed that, overall, there appeared to swaths of differences in scoring between readers of the examination. These differences were sometimes more prominent than others and seemed to vary based on the pairing of the readers and the status and years at Erikson. As I began thinking about an action research dissertation project, I thought about how several faculty members over this period would mention that there was a pressing need for training on the approach to assessing the exam as well as the need for calibration on the grading rubric. Those conversations stemmed from an analysis that was completed focusing on training during the institutional reaccreditation site visits that occurred in 2019.

There is existing research and literature that suggest positive correlations between training and rubric calibration and increases in inter-rater reliability. In a study examining efficiency in assessment, researchers trained both students and faculty on how to rate essays. The results showed that the trained students rated as reliable as the faculty and the lack of rater-training is attributed to negative inter-rater reliability (Cole et al., 2012). Additionally, Hung et al. (2012) focused their research on rater effects and the impact that tools utilized to train raters

have on inter-rater reliability. The authors posit that reliability is impacted by focused training and moreover, raters that were not calibrated on the scoring rubric were shown to have less reliable scores.

There is a vast amount of research into the impact that faculty status may have on grading. Research completed by Filetti et al. (2010) found differences in grading/assessment between tenured and tenure track faculty. As previously mentioned, in serving as the proctor and recorder of scores for the comprehensive examination, I have seen scoring gaps in the examination results that appeared to have low inter-rater reliability. Although anecdotal, the gaps in scoring were apparent enough to spark my interest into reviewing faculty status. Furthermore, there is research on the impact of newly hired faculty as compared to longer employed faculty as it relates to grading, both from a grade inflation to grade deflation standpoint. Deboer et al. (2007) found a need for approval in therefore grade inflation for newer faculty.

As a result of my experience in proctoring the examination and viewing scoring differences, albeit anecdotally, along with the research into training, calibration, faculty and employment scoring/grading impacts, my chosen research centers on the impact of training and calibration in scoring as well as researching the role that faculty status and years employed may have on grading and inter-rater reliability.

Findings

Hypothesis 1

The primary hypothesis for this research study was as follows: Is inter-rater reliability negatively impacted when no formal or consistent rater training is provided? The null hypothesis appears as follows: H_o1: Inter-rater reliability is not significantly impacted when no formal or

consistent rater training is provided. Total faculty reader scores from 2017 to 2021 were (*N* = 794) which were split evenly between faculty reader 1 (*N* = 397) and faculty reader 2 (*N* = 397). Faculty reader 1 scoring and faculty reader 2 scoring on average were relatively similar with faculty reader 1 scores at (*M* =28.65, SD = 5.589) and faculty reader 2 scores at (M = 28.55, SD = 5.756).

Table 1

		Reader 1 Score	Reader 2 Score
N	Valid	397	397
	Missing	2	2
Mean		28.65	28.55
Median		29.00	29.00
Mode		31	32
Std. Deviation		5.589	5.756
Range		28	27
Minimum		12	13
Maximum		40	40

Inter-rater reliability

To test the hypothesis that inter-rater reliability is negatively impacted when no formal or consistent rater training is provided (faculty reader 1 score *N* = 397, *M* = 28.65, *SD* = 5.58) (faculty reader 2 score *N* = 397 *M* = 28.55, *SD* = 5.75) an intra-class correlation coefficient test was performed. A moderately low degree of reliability was found between faculty readers 1 and faculty readers 2 scores. Intraclass correlation was run on a two-way mixed model as the faculty readers are not fixed and set for absolute agreement between the scores. The average measure of ICC was .602 with a 95% confidence interval from .515 to .673 ($F(396,396) = 2.507$, $p<.001$).The inter-rater reliability analysis produced results that support not accepting the null hypothesis.

Table 2

Intraclass Correlation Coefficient

	Intraclass Correlation[b]	95% Confidence Interval		F Test with True Value 0		
		Lower Bound	Upper Bound	Value	df1	df2
Single Measures	.430[a]	.347	.507	2.507	396	396
Average Measures	.602[c]	.515	.673	2.507	396	396

Intraclass Correlation Coefficient

	F Test with ..
	Sig
Single Measures	<.001
Average Measures	<.001

Independent Samples t-test

To test for significance in the hypothesis that inter-rater reliability is negatively impacted when no formal or consistent rater training is provided (faculty reader 1 score N =397, M = 28.65, SD = 5.58) (faculty reader 2 score N = 397, M = 28.55, SD = 5.75) an independent samples t-test was performed. As a result of the independent samples t-test, the null hypothesis is accepted as $t(792) = .244$, p =.807. The results indicate no significant difference between the mean scores of faculty readers 1 and faculty readers 2.

Scoring Gaps Analysis

As part of an overall curiosity into the scoring differences I adjusted the data within a master spreadsheet that housed the scoring readers for faculty readers 1 and faculty readers 2. A third column was then created that would provide the scoring differences between each faculty reader. This data was analyzed utilizing descriptive measures with the intent of learning more about the overall scoring gaps as well as year to year scoring gaps with overall 2017-2021 (N = 397, M = 4.00, SD = 3.69, range 0-26), for the year 2017 (N = 73, M = 4.73, SD = 4.12, range 0-18), for the year 2018 (N = 86, M = 5.01, SD = 4.06, range 0-18), for the year 2019 (N = 82, M =

4.78, SD = 3.26, range 0-15), for the year 2020 (N = 73, M = 4.60, SD = 3.28, range 0-12), and finally for the year 2021 (N = 83, M = 4.83, SD = 3.70, range 0-20).

Table 3

Statistics

		Scoring difference 2017	Scoring difference 2018	Scoring difference 2019	Scoring difference 2020	Scoring difference 2021
N	Valid	73	86	82	73	83
	Missing	13	0	4	13	3
Mean		4.73	5.01	4.78	4.60	4.83
Median		4.00	4.00	5.00	4.00	4.00
Mode		2[a]	2	5	2	1
Std. Deviation		4.127	4.060	3.262	3.282	3.708
Range		26	18	15	12	20
Minimum		0	0	0	0	0
Maximum		26	18	15	12	20

I created an instrument to test for agreement on scoring ranges overall and year to year. For overall agreement 2017-2021, the mean places the readers scoring ranges in the moderate agreement category. For all years, although the scoring ranges fluctuate slightly, but not significantly, the readers scoring ranges are all in the moderate agreement category. Although this instrument provided a moderate agreement, I was still curious about the ranges and the percentage of ranges per year and overall, that fell in the low agreement and no agreement areas (ranges 8 >). The results of this analysis yielded the following. For the year 2017 (*N* = 14 16.3% of paired scores had scoring ranges from 8 or greater. For the year 2018 (*N* = 20) 23.5% of paired scores had scoring ranges from 8 or greater. For the year 2019 (*N* = 20) 21%. For the year 2020 (*N* = 13) 15.1% of paired scores had scoring ranges from 8 or greater. For the year 2021, (*N* = 16) 18.8%. The year 2017 had the highest single differential range amongst readers, that being 26 followed by 2021 with a single high differential range amongst readers, that being 20.

Comprehensive Examination Reader Scoring Agreement	
Scoring Differential Range Among Readers	**Labels**
0-3	High Agreement
4-7	Moderate Agreement
8-11	Low Agreement
12 >	No Agreement

Hypothesis 2

The secondary hypothesis for this research study was as follows: Will the scoring of tenured faculty be lower than tenure-track, clinical faculty, and adjunct faculty? The null hypothesis appears as follows: H_o2: Faculty status does not significantly impact scoring. Total tenured faculty reader scores from 2017 to 2021 were (N = 267). Total "all other faculty" reader scores from 2017 to 2021 were (N = 527). Tenured faculty reader scores were on average slightly higher than "all other faculty" reader scores with tenured faculty reader scores at (M =28.73, SD = 6.23) and "all other faculty" reader scores at (M = 28.63, SD = 5.36).

Independent Samples t-test

To test for significance in the hypothesis that the scoring of tenured faculty would be lower than tenure-track, clinical faculty, and adjunct faculty (tenured faculty reader score N =267, M = 28.73, SD = 6.23) (All other faculty reader score N = 260, M = 28.59, SD = 5.24) an independent samples t-test was performed. As a result of the independent samples t-test, the null hypothesis is accepted as t (525) = .290, p =.772. The results indicate no significant difference between the mean scores of tenured faculty readers and all other faculty readers.

Table 4

Group Statistics

	Group	N	Mean	Std. Deviation	Std. Error Mean
Exam Scores	Tentured Faculty	267	28.73	6.233	.381
	All other faculty	260	28.59	5.241	.325

		t-test for Equality of Means		
		Significance		Mean Difference
		One-Sided p	Two-Sided p	
Exam Scores	Equal variances assumed	.386	.772	.146
	Equal variances not assumed	.386	.772	.146

Hypothesis 3

The tertiary hypothesis for this research study was as follows: Will the scoring of recently hired faculty be more reliable than faculty that has been hired over 10 years? The null hypothesis appears as follows: H_o3: Years of employment does not significantly impact reliability in scoring amongst recently hired faculty and faculty hired over 10 years. Total faculty reader scores with 0-9 years of employment (N = 250). Total faculty reader scores with 10 years or more of employment (N = 158). Faculty reader scores with 0-9 years of employment (M =28.77, SD = 5.10) were on average slightly lower than faculty readers with 10 years or more of employment (M =28.87, SD = 6.25).

Inter-rater reliability

To test the hypothesis that the scoring of recently hired faculty will be more reliable than faculty that has been hired over 10 years an intra-class correlation coefficient test was performed amongst the groups of readers. The approach to this was to pair readers 1 with readers 2 with their respective category. To perform this task, I filtered all reader 1 scores that had faculty employment from 1-9 years, with their paired reader 2 that had faculty employment from 1-9 years. This same process was performed for reader 1 and reader for faculty with 10 years or more years of employment. As a result, (faculty readers (1) with 0-9 years employment scores N = 79, M = 29.20, SD = 5.42); (faculty readers (2) with 0-9 years employment scores N = 79, M =

28.33, *SD* = 4.75); (faculty readers (1) with 10 year or greater scores *N* = 125 *M* = 29.01, *SD* = 6.31); (faculty readers (2) with 10 year or greater scores *N* = 125 *M* = 28.74, *SD* = 6.21)

For faculty readers scores with employment from 0-9 years, a moderate degree of reliability was found between faculty readers 1 and faculty readers 2 scores. Intraclass correlation was run on a two-way mixed model as the faculty readers are not fixed and set for absolute agreement between the scores. The average measure of ICC was .621 with a 95% confidence interval from .411 to .757 (F(78,78) = 2.66, p<.001).

Table 5

Scale Statistics

Mean	Variance	Std. Deviation	N of Items
57.53	75.662	8.698	2

Intraclass Correlation Coefficient

	Intraclass Correlation[b]	95% Confidence Interval		F Test with True Value 0		
		Lower Bound	Upper Bound	Value	df1	df2
Single Measures	.451[a]	.258	.609	2.665	78	78
Average Measures	.621[c]	.411	.757	2.665	78	78

Intraclass Correlation Coefficient

	F Test with ..
	Sig
Single Measures	<.001
Average Measures	<.001

For faculty readers scores with employment from 10 year or more, a moderate degree of reliability was found between faculty readers 1 and faculty readers 2 scores. Intraclass correlation was run on a two-way mixed model as the faculty readers are not fixed and set for absolute agreement between the scores. The average measure of ICC was .620 with a 95% confidence interval from .458 to .733 (F(124,124) = 2.62, p<.001).

Table 6

Scale Statistics

Mean	Variance	Std. Deviation	N of Items
57.74	113.547	10.656	2

Intraclass Correlation Coefficient

	Intraclass Correlation[b]	95% Confidence Interval		F Test with True Value 0		
		Lower Bound	Upper Bound	Value	df1	df2
Single Measures	.449[a]	.297	.579	2.622	124	124
Average Measures	.620[c]	.458	.733	2.622	124	124

Intraclass Correlation Coefficient

	F Test with ..
	Sig
Single Measures	<.001
Average Measures	<.001

Based on the inter-rater reliability analysis results comparing both groups, the null hypothesis is neither accepted nor rejected.

Independent Samples t-test

To test for significance in the hypothesis that that the scoring of recently hired faculty will be more reliable than faculty that has been hired over 10 years (faculty reader with 0-9 years of employment score N =158, M = 28.77, SD = 5.10); (faculty reader with 10 years or more of employment score N = 250, M = 28.87, SD = 6.25) an independent samples t-test was performed. As a result of the independent samples t-test, the null hypothesis is accepted as $t(406) = -.179$, p =.858. The results indicate no significant difference between the mean scores of faculty readers with 1-9 of years employment and faculty readers with 10 or more years of employment.

Table 7

Group Statistics

	Group	N	Mean	Std. Deviation	Std. Error Mean
Exam Scores	Faculty Years of Employment 0-9	158	28.77	5.103	.406
	Faculty Years of Employment 10+	250	28.87	6.251	.395

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means	
		F	Sig.	t	df
Exam Scores	Equal variances assumed	8.473	.004	-.179	406
	Equal variances not assumed			-.187	380.319

Independent Samples Test

		t-test for Equality of Means		
		Significance		Mean Difference
		One-Sided p	Two-Sided p	
Exam Scores	Equal variances assumed	.429	.858	-.106
	Equal variances not assumed	.426	.851	-.106

Scoring Gaps Analysis

As performed with Hypothesis 1, as part of an overall curiosity into the scoring differences I adjusted the data within a master spreadsheet that housed the scoring readers for faculty readers with 0-9 years of employment and faculty readers with 10 or more years of employment. I then created a third column that would provide the scoring differences between the groups faculty reader. I analyzed this data utilizing descriptive measures with the intent of learning more about the overall scoring gaps as well as year to year scoring gaps amongst the groups.

For faculty readers scores with 0-9 years of employment, overall 2017-2021 (N = 79, M = 4.22, SD = 3.34, range 0-15), for the year 2017 (N = 10, M = 3.96, SD = 3.57, range 0-10), for the year 2018 (N = 23, M = 3.96, SD = 3.57, range 0-15), for the year 2019 (N = 13, M = 5.54, SD = 3.41, range 0-11), for the year 2020 (N = 11, M = 4.36, SD = 3.41, range 0-11), and finally for the year 2021 (N = 22, M = 43.77, SD = 3.71, range 0-12).

Table 8

Statistics

		Scoring Differences 2017	Scoring Differences 2018	Scoring Differences 2019	Scoring Differences 2020	Scoring Differences 2021
N	Valid	10	23	13	11	22
	Missing	13	0	10	12	1
Mean		3.90	3.96	5.54	4.36	3.77
Median		3.50	3.00	5.00	3.00	3.00
Mode		0[a]	1[a]	5	2[a]	2
Std. Deviation		3.178	3.574	3.526	3.414	3.116
Minimum		0	0	1	0	0
Maximum		10	15	12	11	12

For faculty readers scores with 10 or more years of employment, overall 2017-2021 (*N* = 125, *M* = 4.93, *SD* = 4.34, range 0-26), for the year 2017 (*N* = 28, *M* = 5.54, *SD* = 5.27, range 0-26), for the year 2018 (*N* = 22, *M* = 5.68, *SD* = 5.11, range 0-18), for the year 2019 (*N* = 29, *M* = 4.31, *SD* = 2.71, range 0-11), for the year 2020 (*N* = 23, *M* = 3.61, *SD* = 3.10, range 1-12), and finally, for the year 2021 (*N* = 23, *M* = 5.57, *SD* = 4.90, range 0-20).

Table 9

Statistics

		Scoring Differences 2017	Scoring Differences 2018	Scoring Differences 2019	Scoring Differences 2020	Scoring Differences 2021
N	Valid	28	22	29	23	23
	Missing	1	7	0	6	6
Mean		5.54	5.68	4.31	3.61	5.57
Median		5.00	4.00	4.00	2.00	4.00
Mode		2	1	5	1	1[a]
Std. Deviation		5.274	5.177	2.714	3.100	4.907
Range		26	18	11	11	20
Minimum		0	0	0	1	0
Maximum		26	18	11	12	20

I once again utilized the instrument created to test for agreement on scoring ranges overall and year to year. For overall agreement 2017-2021, the mean places both groups of readers scoring ranges in the moderate agreement category. For all years, although the scoring

ranges fluctuate slightly, but not significantly, both groups scoring ranges are all in the moderate agreement category.

Comprehensive Examination Reader Scoring Agreement	
Scoring Differential Range Among Readers	**Labels**
0-3	High Agreement
4-7	Moderate Agreement
8-11	Low Agreement
12 >	No Agreement

As with hypothesis 1, although this instrument provided a moderate agreement output for both groups, I was still curious about the ranges and the percentage of ranges per year and overall, that fell in the low agreement and no agreement areas (ranges 8 >). The results of this analysis yielded the following for the faculty readers with 0-9 years of employment. For the year 2017 (*N* = 1) 4.3% of paired scores had scoring ranges from 8 or greater. For the year 2018 (*N* = 4) 17.3% of paired scores had scoring ranges from 8 or greater. For the year 2019 (*N* = 4) 17.3%. For the year 2020 (*N* = 1) 4.3% of paired scores had scoring ranges from 8 or greater. For the year 2021, (*N* =2) 8.6%.

The results of this analysis yielded the following for the faculty readers with 10 or more years of employment. For the year 2017 (*N* = 7) 23.9% of paired scores had scoring ranges from 8 or greater. For the year 2018 (*N* = 6) 20.4% of paired scores had scoring ranges from 8 or greater. For the year 2019 (*N* = 3) 10.2%. For the year 2020 (*N* = 3) 10.2% of paired scores had scoring ranges from 8 or greater. For the year 2021, (*N* =6) 20.5%.

As you can see based on the data provided there are fluctuation in ranges from both groups, however the faculty readers with 0-9 years of employment have an overall smaller gap in scoring amount as well as a lower range of scoring gaps as compared to the faculty readers with 10 or more years of employment.

Comprehensive Examination Reader Scoring Agreement	
Scoring Differential Range Among Readers	**Labels**
0-3	High Agreement
4-7	Moderate Agreement
8-11	Low Agreement
12 >	No Agreement

Hypothesis 4

The "action research" hypothesis for this research study was as follows; Will inter-rater reliability significantly increase among trained scores regardless of faculty status? The null hypothesis appears as follows: H_o4: Inter-rater reliability is not significantly impacted when no formal or consistent rater training is provided.

Hypothesis 4 was created with the hopes of testing for the impacts of what a formal rater training and calibration would have on inter-rater reliability. Unfortunately, as previously mentioned, the Institute's assessment committee did not act on the findings nor request to hold a formal rubric training and calibration prior to the June 2022 comprehensive examination. The results provided are for the June 2022 comprehensive examination only. To reiterate, no rubric calibration or training was provided to the faculty prior to rating the June 2022 examination. Total faculty reader scores were (N = 64) which were split evenly between faculty reader 1 (N = 32) and faculty reader 2 (N = 32). Faculty reader 1 scoring and faculty reader 2 scoring on average were relatively similar with faculty reader 1 scores at (M =29.75, SD = 5.442) and faculty reader 2 scores at (M = 30.63, SD = 5.975).

Table 10

	Mean	Std. Deviation	N
Reader 1 Score	29.75	5.442	32
Reader 2 scores	30.69	5.975	32

Inter-rater reliability

As mentioned, because there was no formal rater training and rubric calibration, I was unable to test hypothesis 4 as intended. However, continuing the research on the importance of completing rater training and rubric calibration can be gleaned from the data analysis performed. To test inter-rater reliability in the absence of formal or consistent rater training and calibration (faculty reader 1 score $N = 32$, $M = 29.75$, $SD = 5.44$) (faculty reader 2 score $N = 32$, $M = 30.63$, $SD = 5.97$) an intra-class correlation coefficient test was performed solely on the June 2022 examination results. A poor degree of reliability was found between faculty readers 1 and faculty readers 2 scores. Intraclass correlation was run on a two-way mixed model as the faculty readers are not fixed and set for absolute agreement between the scores. The average measure of ICC was .127 with a 95% confidence interval from -.815 to .577 ($F(31,31) = 1.143$, $p<.356$). The inter-rater reliability analysis produced results that support not accepting the null hypothesis.

Table 11

Intraclass Correlation Coefficient

	Intraclass Correlation[b]	95% Confidence Interval		F Test with True Value 0		
		Lower Bound	Upper Bound	Value	df1	df2
Single Measures	.068[a]	-.290	.405	1.143	31	31
Average Measures	.127	-.815	.577	1.143	31	31

Intraclass Correlation Coefficient

	F Test with ..
	Sig
Single Measures	.356
Average Measures	.356

Discussion

Hypothesis 1: Is inter-rater reliability negatively impacted when no formal or consistent rater training is provided?

Synthesis of Themes

There were three very important themes that came out of the data analysis. The first important theme, which I believe is vital to this study is that overall inter-rater rater reliability is impacted when there is no consistent rater training and rubric calibration present. The results of the intra-class correlation have the overall reliability at a moderate level. Although the results of the intra-class correlation did not reveal a high level of reliability, what I find very reassuring is that this core group of faculty, although not having had consistent training and calibration, and some not having any training or calibrations at all, did an overall good job in this category.

The second important theme came as a result of the independent samples *t*-test. The data analysis showed no significant differences between faculty reader 1 scores and faculty reader 2 scores. We can see by the means as well as the results of the intra-class correlation that readers, although not having been consistently trained and calibrated, were still able to statistically not have significant differences in scoring. This is an important finding as it does demonstrate that while agreements are not absolute or at a high level, the faculty would not be starting from a level of agreement or significance that would be alarming or would possibly need intense intervention to correct the agreement rates.

The third important theme came from the results of scoring gap analysis. Whereas this data was not completed through a statistical measurement, the descriptive statistics that were born out of the data displayed some concerns in the ranges of scoring gaps between faculty readers 1 and faculty readers 2. Crucially, some scoring difference ranges were as high as 26 points. Additionally, scoring ranges from 8 points or greater accounted anywhere from 16% to 23% of scores within a given years examination. Although the overall means were close and

agreement was moderate, these scoring range differences and the percentages that occur can impact a student's ability to pass the examination or require a third reader.

Connections to the Literature

As several studies show, there are positive impacts on inter-rater reliability when training and calibration occur. Cole et al (2012) provided this comprehensive look into the impact of training when the researchers trained both students and faculty on essay rating and performed rubric calibration. Individuals that were trained in this study showed positive inter-rater reliability as a result. Additionally, Rios et al. (2017) had raters go through training and calibration within that study and once again the results displayed significant inter-rater reliability.

As discussed previously, the faculty at the Institute had an overall moderate level of reliability. This interpretation of intra-class correlation results was provided utilizing results interpretation of Koo and Lee (2016). There are two leading research studies that provide interpretation of results from an intra-class correlation analysis. I utilized the results interpretation of Koo and Lee (2016) as I felt their research and rationale of interpretation was grounded in quantitative data analysis, which this hypothesis and the subsequent results were as well. An alternative interpretation of intra-class correlation results is provided by Cicchetii (1994). I declined to use those interpretations as they are typically used for mixed-methods and qualitative intra-class correlation analysis.

As the proctor of the comprehensive examination, I felt that the overall results of hypothesis 1 testing were positive. Although I felt it appropriate to not accept the null, in general, these results are a great start at looking at how to impact policy and process to get the reliability up from moderate to high, which is not too far away. There are several tactics that

have been developed within this field of research that can employed such as the timing of rubric calibration and training (Szafran, 2017) to the approaches to scoring such as the study by Finn et al. (2018) in their research exploring scoring and rest periods over time.

Hypothesis 2: Will the scoring of tenured faculty be lower than tenure-track, clinical faculty, and adjunct faculty?

Synthesis of Themes

As provided within the results, the null hypothesis was accepted as no significant difference between faculty reader scores and all other faculty readers scores was found when an independent samples *t*-test analysis was completed. Of importance was the means for both reader groups were nearly the same. Thus, once again we see as in hypothesis 1, the faculty, even when paired differently to test against status and scoring, came out similar in average scoring. This theme of the Institute faculty, on average having moderate agreement is again very reassuring and provides a solid foundation that one can assume with the implementation of training and calibration on a consistent basis will be able to have a more significant impact.

Connections to the Literature

In accepting the null hypothesis, this was counter to some of the research utilized in this study. That research along with my experience in proctoring and score retrieval from faculty was the catalyst to for the creation of hypothesis 2. For example, Filetti et al. (2010) found in their study that tenure track professors tend to be more lenient in grading. Additionally, Johnson, (2010) addresses the impact of grades amongst faculty and found that tenure-track and adjunct faculty are more likely to provide higher grades than that of tenured faculty. In the case of this study, the results of the independent samples *t*-test suggest no significant differences between

tenured faculty reader scores and all other faculty readers scores. In fact, tenured faculty on average provided higher grades on the comprehensive examination then their other faculty peers.

Hypothesis 3: Will the scoring of recently hired faculty be more reliable than faculty that has been hired over 10 years?

Synthesis of Themes

As in the case of hypothesis 1, there were three very important themes that came out of the data analysis for hypothesis 3. The first theme, which is the theme throughout this study in addressing inter-rater reliability was that reliability within these two groups analyzed is impacted when there is no consistent rater training and rubric present. The results of the intra-class correlation have the overall reliability at a moderate level, and slightly higher than in hypothesis 1 where there was no grouping and was simply testing readers 1 against readers 2. The results of the intra-class correlation did not reveal a high level of reliability, but again an assuring theme is that this core group of faculty, overall had a moderate level of reliability despite the lack of consistent training and calibration present.

The result of the independent samples *t*-test revealed the second theme as it pertains to hypothesis 3. The data analysis showed no significant differences between faculty reader scores with 0-9 years of employment and faculty reader scores with 10 or more years of employment. In these cases, we see via the means as well as the results of the intra-class correlation that readers, although not having been consistently trained and calibrated, were still able to statistically not have significant differences in scoring. This an important finding serves to reinforce that while agreements are not absolute or at a high level, the faculty would not be starting from a level of agreement or significance that would cause consternation or been seen as an impossible leap to increase reliability.

As we so often do in academia and research specifically, we find ourselves jumping into the "rabbit hole" in exploration for explanations or just general curiosity. As such the third important theme came from the results of scoring gap analysis. As noted previously, this data was not completed through a statistical measurement, however, the descriptive statistics that were created provided some concerns in the ranges of scoring gaps between faculty reader scores with 0-9 years of employment and those of faculty readers scores with 10 or more years of employment. Amongst these groups, the scoring range gaps were overall smaller on average for faculty scores with 0-9 years of employment as compared to those with 10 or more years of employment. Overall percentages from 8 points or greater were also lower for the 0-9 employment group as compared to the 10 or more years of employment group. That said, this is by no means scientific exploration, however with that, and based on the results, an assumption could be made that faculty with 10 or more years of employment could benefit from rubric calibration to assist in lowering the scoring gaps observed.

Connections to the Literature

This continued theme of adjunct and newly hired faculty grade differential was where hypothesis 3 was born through. Research taken part by Kezim et al. (2005) on grade point averages over a 20-year span found that adjunct faculty tended to grade higher than tenured faculty. After reading through the literature, I found this theme of grade inflation, new hire faculty and adjunct faculty continuously appearing through my review. Several researchers found themes in grade inflation or grade differentials. Chief among those themes were faculty status but also job security. Nikolakakos et al. (2012) found that 60% of faculty respondents to a survey regarding factors of grade inflation cited job security as a reason contributing to higher percentages of grades. Moreover, Izienciki and Setchfield (2019) found that instructors with less

teaching experience were more likely to provide options for extra credit to students. These connections of experience and grades, or newly hired faculty as compared to longer employed faculty give credence to additional research in this field as well as methods to address grading more consistently regardless of employment length.

Hypothesis 4: (After action) Will inter-rater reliability significantly increase among trained scores regardless of faculty status?

Synthesis of Themes

As previously mentioned, I was unable to test for hypothesis 4 in its truest sense. The lack of decision by the assessment committee to hold a formal rater training and rubric calibration session stymied my efforts to test how those initiatives would have impacted inter-rater reliability, testing for significance, and scoring gap analysis. However, in solely utilizing the June 2022 comprehensive examination data and testing that data for inter-rater reliability, the main theme of this research was evident, that overall inter-rater rater reliability is impacted when there is no consistent rater training and rubric present. The results of the intra-class correlation run only on the June 2022 examination results speak volumes in that they produced an overall reliability at a poor level. The results of this singular intra-class correlation reveal the continued impact of not having consistent training and calibration has in inter-rater reliability.

Connections to the Literature

As discussed previously, the faculty at the Institute had an overall poor level of reliability as they pertain solely to the June 2022. Once again, this interpretation of intra-class correlation results was provided utilizing results interpretation of Koo and Lee (2016). I utilized the results interpretation of Koo and Lee (2016) as I felt their research and rationale of interpretation was grounded in quantitative data analysis, which this hypothesis and the subsequent results were as

well. An alternative interpretation of intra-class correlation results is provided by Cicchetii (1994). I declined to use those interpretations as they are typically used for mixed-methods and qualitative intra-class correlation analysis. Notwithstanding its singular lens, the results of the June 2022 comprehensive should not be glossed over and should further show what current literature has found, that inter-rater reliability is impacted when there is no formal training and calibration completed. The timing of the training and calibration is key as noted by Szafarn (2017) as well as Gugiu et al. (2012) where their research reflects positive impacts of consistent training and calibration.

Conclusion

The results of the hypotheses testing were in some respects not surprising. Being privy to the scores over the past five years and anecdotally noticing themes helped me formulate the hypotheses along with the assistance of the existing literature. I knew that while there were scoring gap swings prevalent in every iteration of the comprehensive examination throughout the years, that it was not as dramatic as it may appear. However, I did feel that the readers scores could be more reliable. Having completed the statistical analysis and the results showing that overall, excluding the June 2022 examination results, the inter-rater reliability is moderate, brought a sense of relief in some respects as it provided positive aspects of this journey and a sense of being not too far off from achieving high reliability. Although the test for significance yielded acceptance of the null hypotheses, the results of the scoring gaps analysis yielded areas of opportunity to improve. The opportunity to review and analysis this data and report out the results are positives as we lead towards improvement and policy driven process moving forward as it pertains to rater training and rubric calibration.

As I set out on this journey, my experience at the Institute as a proctor and colleague helped guide me toward the direction of this research. The results of the data analysis have provided a clear picture of opportunity and action for both myself and my faculty colleagues and administrators. Our charges as higher education professionals are to be reliable and consistent. The pathway to reliability and consistency is not straight and requires work, dedication, and buy-in, while always striving to see things through the lens of the student. Knowing that, the results of this research and the subsequent actions taken to improve upon what is already a solid base will undoubtedly serve to remind students and colleagues that the Institute is dedicated to its mission of knowledge, practice, and reflection.

CHAPTER 5

Conclusions

Introduction

The purpose of this research study was to explore the impact that consistent and collaborative training and rubric calibration, or the lack thereof, has on inter-rater reliability. Furthermore, this study explored the influences that faculty status has on scoring outcomes/equity. Additionally, this study also investigated the affects that length of employment has on scoring and inter-rater reliability. The apex goal of this research study was to utilize the data in an intentional manner to create a formal strategy towards codifying a comprehensive examination training and calibration policy and process led by the Institute's assessment and faculty committees.

Hypothesis testing within this study found that inter-rater reliability is indeed negatively impacted when no formal or consistent rater training and rubric calibration occurs, nor the existence of such training and calibration. Additional hypothesis testing into the impact of faculty status and scoring differences found no significant differences. Finally, hypothesis testing on connections between inter-rater reliability and length of employment found no significant differences in scoring. Unfortunately, the fourth and final hypothesis was unable to be tested due to inaction within the Institute's assessment and faculty committees to move forward with training and rubric calibration upon being briefed on the initial data findings.

The dialogue contained within this chapter rounds out the research by once again focusing on the research purpose and questions along with clarity that can was gleaned via the data analysis. Implications for practice will be discussed in the true spirit of action research along

with suggestions for future research. Finally, research limitations are presented to wholistically round out my own, and the Institute's guiding pillars, of knowledge, practice, and reflection.

Research Purpose, Questions, and Answers

The purpose of this research study was to describe the impact that consistent training and calibration has on inter-rater reliability. Additionally, this study went on to examine the relationship between faculty status and scoring as it pertains to inter-rater reliability and if faculty length of employment impacts rater scoring. Five years of examination scoring data from 25 faculty members was used from The Development Institute, a small Midwest region private graduate school.

Hypothesis 1 asks the main question of this research study, that being if inter-rater reliability negatively impacted when no formal or consistent rater training is provided? As previous literature attests, inter-rater reliability is negatively impacted when no formal or consistent rater training and/or rubric calibration is completed as does this research study. The results of the intra-class correlation provided produced overall reliability at a moderate level. Crucially, although the results of the intra-class correlation did not reveal a high level of reliability, the independent samples *t*-test showed no significant difference between faculty reader 1 scores and faculty reader 2 scores. As previously mentioned, these are important finding as it does demonstrate that while agreements are not absolute or at a high level, the faculty would not be starting from a significantly lower level of agreement.

Hypothesis 2 continues the work of previous literature in looking into grading differences, specifically if the scoring of tenured faculty would be lower than tenure-track, clinical faculty, and adjunct faculty? Although the null hypothesis was accepted as no significant difference between faculty reader scores and all other faculty readers scores was

found when an independent samples *t*-test analysis was completed, there are several themes in the current literature that suggest differences in grading based on faculty status occur, otherwise referred to as grade inflation. I address this in more detail in the suggestions for future research section later in the chapter as I feel there is an opportunity to delve deeper into this hypothesis utilizing the data that was provided for this research study.

Hypothesis 3, which asks if the scoring of recently hired faculty be more reliable than faculty that has been hired over 10 years sought to address questions of inter-rater reliability and grade inflation in connection to literature themes that centered around tenured and non-tenured faculty scoring differences. Much like the results of hypothesis 1, the results of the intra-class correlation produced reliability at a moderate level, only slightly higher than the results produced in hypothesis 1. The result of the independent samples *t*-test showed no significant differences between faculty reader scores with 0-9 years of employment and faculty reader scores with 10 or more years of employment. I feel it is important to reiterate the importance of these results in that while agreements were not absolute or at a high level, one would assume that with even a minimal amount of training performed over time consistently, that these results have the potential to significantly increase given the lack of training over the past 5 years.

Hypothesis 4 was meant to serve as the crux of the action research within this dissertation as it would have included a thorough training and rubric calibration one week prior to the next comprehensive examination. Unfortunately, the assessment committee did not act, even after being presented with findings. Therefore, hypothesis 4 was unable to be tested. Considering this, I continued the research and completed testing the June comprehensive examination data set. As expected, a low degree of reliability was found between faculty readers 1 and faculty readers 2 scores after intra-class correlation coefficient testing was completed.

The results of hypothesis testing yield obvious actions points for the Institute to undertake and support the findings both within this research study and the literature in which inter-rater reliability is negatively impacted when no formal or consistent rater-training and rubric calibrations are performed for raters of examinations.

Implications for Practice

Data informed decision making is a hallmark of higher education professionals. So much so that assessment committees work tirelessly in the pursuit of data to guide decision makers towards continuous improvement. With assessment data in-hand, higher education professionals that can impact internal policy and processes must do so confidently utilizing the data provided. However, creation or change must be done so in a collaborative nature as policy and process created in a silo often suffer from a failure to launch.

Diversifying Assessment

In higher education, when we think of assessment, one may naturally think about student centric outcomes and how institutes of higher education are assessing for student success. Hundley et al. (2019) posits that institutes of higher education should utilize assessment in response to the shifting needs of students that are taking place. This is historically in-tune with traditional assessment; however, holding this singular focus is outdated and not inclusive. The implications for practice of this research study widens the lens of assessment to not only focus on student success, but on faculty success as well.

As the student-body diversifies, so too does the faculty-body and in response, assessment must also diversify. As a higher education professional with almost 20 years of experience, I can attest that student success does not occur in a vacuum, it takes a village to ensure that an institute of higher education is truly student-ready (McNair et al., 2016). As institutes respond to the

changing definition of what student success looks like, they must also reflect on how faculty-success is viewed as well. The implications of this research study lends itself to the possibility of moving the needle of faculty assessment away from, in some cases, solely grades and student satisfaction survey results to a more holistic view that encourages training, professional and development initiatives that impact the student experience. In doing so, adding in assessment for inter-rater reliability, and increasing reliability standards can provide for a more equitable and inclusive pathway as part of an overall faculty assessment/review process.

Policy and Process Creation

The penultimate impact for practice resulting in this research study is for institutes of higher education, including the Development Institute, to create a formal process that includes rater-training and rubric calibration. As the data has shown, training is crucial and significantly impacts inter-rater reliability. These results suggest that should a formal training policy and process be created, and consistently performed, inter-rater reliability will be positively impacted, and in the case of the Development Institute, no doubt increase from current level of reliability.

Training is not the only implication for practice. Rubric calibration is also a part of a comprehensive policy and process plan that can manifest out of this research study. Intentional care must be taken when creating, calibrating, and assessing rubrics. Reddy & Andrade (2010) assert that the validity of rubrics and their usage center around clarity and appropriateness of language of which raters' calibration occurs. The true value of utilizing a rubric, along with ensuring appropriate rater training and calibration on a rubric is to provide a set of standards for which to assess (Young et al., 2016). Setting a standard not only assists with assessment, but it also creates a pathway towards equity initiatives in mitigating latent or inherent bias in the scoring process.

Much like rethinking assessment, the creation of a formal policy and process for training raters and calibrating a rubric must be completed in a collaborative manner. It is imperative that buy-in from faculty is present and their voices are heard and respective. Additionally, members of an assessment committee, student, and academic services, and crucially, current students and alums are suggested to be included when creating policy and process. Finally, transparency of this plan is paramount. There should be no ambiguity as to what the policy and process is, who the owners are, and when assessing the assessors takes place. The integrity of an institute of higher education can easily be determined but the transparency at which is presents its policies and its consistency in practice.

Suggestions for Future Research

There are several avenues to explore when looking at future research. The literature, which has been presented within this study, along with the hypothesis testing that occurred, suggest that rubric training and calibration impact inter-rater reliability. However, as we continue to explore the impacts, specifically calibration, there are studies that suggest regardless of training and calibration, raters may still impact reliability based on their tendencies to be either too lenient, or too severe when scoring.

Addressing Rater Bias

Tymms and Higgins (2018) boldly suggest in their research that raters go through a pre-appointment procedure paneling process to scale their sense of leniency/severity along with statistical adjustments to the scores their provide based on data. Whereas I agree that there is a need to continue data analysis and reflective assessment on scoring, I do worry that pre-screening raters and making scoring adjustments opens the possibility of further scrutiny of results. Such scrutiny can manifest in the form of what may be seen as “cherry picking” raters

which can lead to questions of inequitable practices, which rater-calibration in and of itself attempts to mitigate. Further, such actions could be seen as "curving" but from the rater standpoint and therefore not providing a true picture of the achievement, or lack thereof, of the student.

As we reflect and continue to implement contemporary learning designs, I do feel that there is a valid case for review of raters and tendencies. The question of rater autonomy comes up frequently throughout the literature as it pertains to rater-calibration. Reddy and Andrade (2010) noted resistant tendencies of instructors towards the use of rubrics. These resistant tendencies can be an attestation of the concern of a loss of autonomy in the scoring process. Although these tendencies exist, if our focus is on reliability of raters, future research is warranted on rater bias and therefore needed to ensure that all submissions are viewed as equitable as possible. Additionally, further research on meta-cognition strategies and integration of such strategies would be an avenue of future research. Strategies that can implemented in the training and calibration sessions such as activating prior knowledge, planning for the task of rating and ultimately structured reflection, both individually and as a group of raters can be assessed in future research.

Timing of Training and Scoring

Several studies that have been cited within this research study address the impacts of training and rubric calibration on inter-rater reliability. Several research studies also provide insight into the timing of rater training along with rater feedback. Szafran, 2017 and Finn et al. 2018 studies address these strategies within their research studies, respectively. As mentioned throughout this study, there has been a lack of training for raters and rubric calibration in general within the Institute around the comprehensive examination. Knowing this and based on the

results of hypothesis testing, I suggest further research into the timing of such training both within the Institute and for future researchers. Once a policy for training and calibration is in place, it will be interesting to assess when the training occurred and if inter-rater reliability is impacted. For example, are their reliability differences if training and calibration occurs 1 month, 1 week, or 1 day prior to the commencement of scoring? Furthering this research will be critical from an Institute standpoint so that a rigorously tested policy and process can be put in place that is data driven and informed.

Based on my experience as the chief proctor for the comprehensive examination, which entails not only proctoring, but managing the entire process comprehensive examination administration including the recording of grades, I have had the opportunity to observe the flow of scored exams filter into the examination repository. At present, there is guidance or suggestions provided to the raters as to length of time spent on scoring the examination. For example, some raters score the exams in bulk, in what I refer to as marathon sessions, where some raters complete one or two a day through the span of a week. Wendler et al. (2019) and Ling et al. (2014) both discuss scoring fatigue and its impacts on inter-rater reliability in their respective research studies. Based on their finding's, scoring fatigue does play a role in impacting inter-rater reliability. Therefore, a suggestion for not only future research, but policy suggestions as well, is to continue studying the impacts of scoring fatigue. The hopes of future research in this area is to provide guidance as part of a comprehensive training and policy program for raters which include suggestions on length of time spent scoring to ensure consistent scoring reliability in addressing scoring fatigue.

Faculty Status of Raters

As mentioned, hypothesis 2, will the scoring of tenured faculty be lower than tenure-track, clinical faculty, and adjunct faculty, found the null hypothesis to be accepted as the results indicated no significant difference between mean scores of tenured faculty and all other faculty. That said, I firmly believe that additional research is needed in this area. To further this research study, I would suggest adjusting the employment statuses utilized and breaking them down individually, not by large groupings as performed in this study, and assess for results. For example, assess tenured faculty against clinical faculty, and clinical faculty against adjunct faculty. Although speculative, this could lead to findings that warrant additional research towards the impact that faculty statuses have on scoring.

Overall, I recommend the future researchers continue to investigate these impacts and perform deeper dives, especially in full-time and part-time faculty. Fedler's study (1989) found that grades submitted by full-time and part-time instructors indicated that part-time faculty graded higher than their full-time colleagues. In addition to Felder, research that was cited within this paper leads one to believe that there are differences in scoring that occurs amongst differences in faculty status and therefore warrants either replication of this research with faculty breakdown recommendations listed above

Limitations

This research study encompassed five years of comprehensive examination data to assess the impact of training and calibration has on inter-rater reliability. Although comprehensive in nature and focused on rater-training and rubric calibration, there are certainly multiple factors present when discussing the nature of inter-rater reliability. Factors such as the length of time a rater performs the task of scoring (scoring fatigue), to the timing of when training is delivered,

can all play a role on how inter-rater reliability is affected. The following is presented to prompt collegial discussions on the nature of limitations observed at the time of this study while also opening an opportunity to utilize these limitations in future research.

This research study data set included scores for 397 examinations and a total of 794 unique exam scores rated by 25 faculty members over a five-year span. A limitation of this study is the results did not include scores of a faculty third read for exams that required a third reader. I specifically chose to exclude third reads to ensure that hypothesis testing remained consistent between two raters, example, faculty reader 1 and faculty reader 2. Third reader scores were not significant across all exam iterations throughout the five years of data. From 2017-2021 there was a total of 33 examinations that required a third reader, peaking in 2018 where in total, 15 examinations required a third read and bottoming out in 2019 where only 3 examinations required a third read.

A second limitation is that the research design included a tool, "Comprehensive Examination Reader Scoring Agreement", that was created specifically for this research study. This tool had not been previously created or utilized to test for rater agreement or reliability. Although this tool was presented to the Institute IRB and Institute faculty assessment committee as part of the research proposal and design planning, it again had not been tested prior to completing this research. The hypothesis testing completed was performed on quantitative data and statistical analysis while the "scoring gap analysis" portion, utilizing statistical analysis for gaps in scoring for reader 1 and reader 2, relied on the "comprehensive examination readers scoring agreement" tool to make assumptions. Based on this, there is a limitation in validity of the "scoring gap analysis". It is suggested that should the Institute decide to include this tool in

the training and assessment policy, that faculty vote on the ranges presented within the tool to assess for agreement in the future.

A third limitation is that the research design excluded the use of individual question scores and knowledge competency scores. The comprehensive examination score sheet breaks down scoring via question by question. Based off the individual scores, a final composite score is then generated. It was that final composite score for both readers 1 and 2 that was utilized for this research study. Additionally, raters are asked, on the score sheet, to provide knowledge competency scores. These scores have no bearing on the final examination score but are provided with the intention that the student's advisor, upon releasing the score to the student and pass/fail status, go over the examination question by question and address the knowledge competency scores as well as they pertain to mastery of subjects.

Conclusion

This chapter concluded that the absence of consistent rater training and rubric calibration, negatively impacts inter-rater reliability. The findings provided implications for practice that would not only affect the Development Institute but offer the possibility of positively affecting other institutes of higher education that choose to replicate and assess their own approach to consistent and collaborative rater training and rubric calibration. In review of the state limitations and with future research in mind, key tenets manifested in writing this chapter, those being opportunity and adaptation. The Institute has the opportunity to adapt their approach to training and calibration and create a standard in which policy and process are created in a collaborative and intentional spirit.

Serving as the chief proctor for the comprehensive examination over the past five years at the Institute has always provided a sense of closure to me. My primary role at the Institute,

within enrollment management, afforded me the opportunity to attract new students that sought to make an impact in the field of human services, while also allowing me to play a pivotal part in the retention of students that were currently enrolled. Proctoring the comprehensive examination for future graduates marked the end of the road in their journey, and mine, in which the that journey often stretched the span of three to four years. I embarked on this research study with the hopes of impacting policy and securing a more formal and consistent process of training and rubric calibration. I end this study, and my tenure at the Institute, confidently knowing that I served its students pursuant of its mission and institutional pillars through this research study by applying knowledge into practice and reflecting on the impact of those actions towards the betterment of self and those around us.

REFERENCES

Attali, Y., Lewis, W., & Steier, M. (2013). Scoring with the computer: Alternative procedures for improving the reliability of holistic essay scoring. *Language Testing*, *30*(1), 125–141.

Bamber, M. (2015). The Impact on stakeholder confidence of increased transparency in the examination assessment process. *Assessment & Evaluation in Higher Education*, *40*(4), 471–487.

Bernardin, H., & Buckley, M. (1981). Strategies in rater training. *The Academy of Management Review*, 6(2), 205-212.

Boretz, E. (2004). Grade inflation and the myth of student consumerism. *College Teaching*, 52(2), 42.

Cicchetti DV (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*. 6 (4): 284–290. doi:10.1037/1040-3590.6.4.284

Cole, T. L., Cochran, L. F., Troboy, L. K., & Roach, D. W. (2012). Efficiency in assessment: Can trained student interns rate essays as well as faculty members? *International Journal for the Scholarship of Teaching and Learning*, *6*(2).

Constructivism-Philosophy of Education. (2021, May 20). In *Wikipedia*. https://en.wikipedia.org/wiki/Constructivism_(philosophy_of_education)

Creswell, J. W., & Creswell, J. D. (2018). *Research design: Qualitative, quantitative, and mixed methods approaches* (5th ed.). Sage Publications.

Deboer, B. V., Anderson, D. M., & Elfessi, A. M. (2007). Grading styles and instructor attitudes. *College Teaching*, *55*(2), 57–64.

Fedler, F. (1989). Adjunct profs grade higher than faculty at three schools. *Journalism Educator*,

44(2), 32-37.

Filetti, J., Wright, M., & King, W. M. (2010). Grades and ranking: When tenure affects assessment. *Practical Assessment, Research & Evaluation*, *15*(14).

Finn, B., Wendler, C., Ricker-Pedley, K. L., & Arslan, B. (2018). Does the time between scoring sessions impact scoring accuracy? An evaluation of constructed-response essay responses on the "GRE"® General Test. Research Report. ETS RR-18-31. *ETS Research Report Series*.

Gardiner, L. R., Kim, D.-G., & Helms, M. M. (2020). Key recommendations for improving AoL assessments: A longitudinal analysis of rater bias and reliability in embedded rubric-based measurements. *Journal of Education for Business*, 95(4), 227–233.

Gugiu, M. R., Gugiu, P. C., & Baldus, R. (2012). Utilizing generalizability theory to investigate the reliability of the grades assigned to undergraduate research papers. *Journal of MultiDisciplinary Evaluation*, 8(19), 26–40.

Herbert, I. P., Joyce, J., & Hassall, T. (2014). Assessment in higher education: The potential for a community of practice to improve inter-marker reliability. *Accounting Education*, *23*(6), 542–561.

Hundley, S. P., Kahn, S., & Banta, T. W. (2019). *Trends in assessment: Ideas, opportunities, and issues for higher education*. Stylus Publishing LLC.

Hung, S.-P., Chen, P.-H., & Chen, H.-C. (2012). Improving creativity performance assessment: A rater effect examination with many facet rasch model. *Creativity Research Journal*, *24*(4), 345–357.

Izienicki, H., & Setchfield, S. (2019). Extra credit in the sociology classroom. *Teaching Sociology*, 47(1), 32–42.

Johnson, I. Y. (2011). Contingent instructors and student outcomes: An artifact or a fact? *Research in Higher Education*, 52(8), 761–785.

Kayapinar, U. (2014). Measuring essay assessment: Intra-Rater and inter-rater reliability. *Eurasian Journal of Educational Research*, *57*, 113–135.

Kirk, F., & Spector, C. A. (2009). A comparison of the achievement of students taught by full-time versus part-time adjunct faculty in business courses. *Academy of Educational Leadership* Journal, 13(2), 73-81.

Korpan, C., PhD. (2020). Grade Inflation and Inequality. Retrieved May 10, 2021, from https://www.uvic.ca/learningandteaching/assets/docs/d---grade-inflation-and-inequality.pdf

Kezim, B., Pariseau, S. E., & Quinn, F. (2005). Is grade inflation related to faculty status? *Journal of Education for Business*, *80*(6), 358.

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine*, 15(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Lee C. (1985). Increasing performance appraisal effectiveness: Matching task types, appraisal process, and rater training. *Academy of Management Review*, *10*(2), 322–331.

Ling, G., Mollaun, P., & Xi, X. (2014). A study on the impact of fatigue on human raters when scoring speaking responses. *Language Testing*, 31(4), 479–499.

Mcarthur, R. C. (1999). A comparison of grading patterns between full- and part-time humanities faculty: A preliminary study. *Community College Review*, 27(3), 65–76.

McNair, T., Albertine, S., Cooper, M., McDonald, N., Major, T. (2016). *Becoming a student-ready college: A new culture of leadership for student success*. Jossey-Bass.

Miles, M. B., Huberman, A. M., & Saldana, J. (2020). Qualitative data analysis: A methods sourcebook. Sage Publications.

Moore, M., & Trahan, R. (1998). Tenure status and grading practices. *Sociological Perspectives*, 41(4), 775-781.

Nikolakakos, E., Reeves, J. L., & Shuch, S. (2012). An examination of the causes of grade inflation in a teacher education program and implications for practice. *College and University*, 87(3), 2–13.

Papajohn, D. (2002). Concept mapping for rater training. *TESOL Quarterly*, *36*(2), 219-233.

Piaget's Theory of Cognitive Development. (2021, May 20). In *Wikipedia*. https://en.wikipedia.org/wiki/Piaget%27s_theory_of_cognitive_development

Pownall, I., & Kennedy, V. (2019). Cognitive influences shaping grade decision-making. *Quality Assurance in Education: An International Perspective*, 27(2), 166–178.

Reddy, Y. M., & Andrade, H. (2010). A review of rubric use in higher education. *Assessment & Evaluation in Higher Education*, 35(4), 435–448.

Rios, J. A., Sparks, J. R., Zhang, M., & Liu, O. L. (2017). Development and validation of the written communication assessment of the "HEIghten"® Outcomes Assessment Suite. Research Report. ETS RR-17-53. *ETS Research Report Series*.

Szafran, R. F. (2017). The Miscalculation of interrater reliability: A case study involving the AAC&U VALUE rubrics. *Practical Assessment, Research & Evaluation*, *22*(11).

The Rhode Island Department of Elementary and Secondary Education. (2021). *Calibration Protocol for Scoring Student Work.*

https://www.ride.ri.gov/Portals/0/Uploads/Documents/Teachers-and-Administrators-Excellent-Educators/Educator-Evaluation/Online-Modules/Calibration_Protocol_for_Scoring_Student_Work.pdf

Tymms, P., & Higgins, S. (2018). Judging research papers for research excellence. *Studies in Higher Education*, 43(9), 1548–1560.

Young, K., James, K., & Noy, S. (2016). Exploration of a reflective practice rubric. *Asia-Pacific Journal of Cooperative Education*, 17(2), 135–147.

Wendler, C., Glazer, N., & Cline, F. (2019). Examining the calibration process for raters of the "GRE"® General test. ETS GRE® Board Research Report. GRE®-19-01. Research Report Series. ETS RR-19-09. *ETS Research Report Series*.