A SYSTEMATIC APPROACH TO INTERRATER RELIABILITY IN EARLY CHILDHOOD TEACHER PERFORMANCE EVALUATIONS

by

Bryndle Laine Bottoms

A dissertation submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in
Educational Research, Measurement and Evaluation

Charlotte

2022

Approved by:

____
Dr. Richard Lambert

____
Dr. Carl Westine

____
Dr. Rebecca Shore

____
Dr. Kelly Anderson

ABSTRACT

BRYNDLE LAINE BOTTOMS. A Systematic Approach to Interrater Reliability in Early Childhood Teacher Performance Evaluations
(Under the direction of DR. RICHARD G LAMBERT)

Teacher evaluations are routinely conducted across the United States for licensure and professional development supports. However, there is limited research on the interrater reliability of these evaluation assessment systems, despite federal recommendations (Graham et al., 2012). This research explores the systematic approach to interrater reliability utilized by the Early Educator Support (EES) Office in North Carolina. The EES Office supports the Birth-through-Kindergarten (B-K) teacher licensure of over 900 early educators in both private and public sectors. The evaluators employed undergo extensive trainings and hold a B-K license themselves. As part of the training, the evaluators undergo an interrater reliability activity that requires them to act as raters who rate ten fictitious teacher profiles, using the North Carolina Teacher Evaluation Process (NCTEP) rubric. This research aimed to understand the rater response process. In this study, three Many Facets Rasch Models are used to understand rater patterns of strictness, leniency and potential bias based on the race of teacher profile. Additionally, two of the models are compared to understand the extent that these rater response patterns are exhibited in their real caseloads of actual early educators. In conclusion, the group of raters do show evidence of strictness, leniency, and bias, however it is mostly exhibited by a small number of individual raters. It is possible to use the results to inform the professional growth of these evaluators, so that all early educators served by the EES Office receive valid, fair, and reliable teacher evaluations. Furthermore, it depicts a systematic approach to interrater reliability that could be used by other evaluation and assessment systems across the country.

DEDICATION

First, for my incredible mother, Elaine, who has sacrificed her entire existence to ensure my success. For my aunts- Phyllis & Joyce, who both shaped me to be a woman of humble, Godly character. For my godson, Sullivan Everett, who has brought immeasurable joy to our lives. For my nieces, Eleanor Grace, and Rosemary Lorraine, may you both learn to be brave, strong, and fearless through my influence. And finally, for my incredible army of friends—Ashley & Carrie, Sarah, Necie, Samantha, Ashley, and Richard— who kept pushing me forward, when I had given up on myself.

ACKNOWLEDGEMENTS

In this life, we are not meant to journey alone but instead, work together. Many people have walked along this journey beside me or supported this journey from afar. First, I am thankful to the University of North Carolina at Charlotte GASP funding and the Department of Educational Leadership for seeing my dedication to this work, and for pouring into me over the past four years. I am thankful for the many opportunities that the Center for Educational Measurement and Evaluation and the Early Educator Support Office have made available to strengthen my research. I want to thank Dr. Rich Lambert for your incredible leadership, support and guidance. I believe it was divine intervention that you came into my life. I am eternally grateful for the opportunity to have worked beside you and learn from your expertise.

Secondly, I thank my mentor, Dr. Bradley Witzel. You sparked a fire within me for learning, research, and math. You saw something in me all those years ago and have continued to cultivate those passions to help me arrive at this level. It is because of your leadership and mentorship that I was successful in this journey.

I thank my committee members Drs. Carl Westine, Rebecca Shore, and Kelly Anderson, for encouraging and believing in me. Each of you had a very distinct and powerful impact on my time in my program. I thank my fellow ERME cohort members for the constant laughs, support, and dedication to each other.

Finally, I would not have made it through the chaos without my CEME colleague, Scott. You have taught me ways to be a better researcher, a better teacher, and a better human. Thank you for showing me how to have balance and boundaries, and how to laugh through the bumps in the road.

TABLE OF CONTENTS

LIST OF TABLES

Chapter I: Introduction

Teacher evaluations are utilized by principals across the country to inform teacher practice, provide formative feedback, and maintain teacher licensure requirements. However, these practices are inconsistent across the states because local education agencies typically determine how teacher evaluations are implemented in various school districts. Consequently, the role of implementing evaluation systems is often left to principals. This is only one responsibility in a vast continuum of principal responsibilities. Furthermore, evaluation systems can vary from state to state, district to district and even school to school.

One of the foundational reports investigating inconsistencies within teacher evaluation found that when a teacher evaluation system fails to provide a full picture of each individual teacher's quality, it is failing to meet the individual needs of the teacher to improve student learning (Weisberg et al., 2009). It has also been shown that student learning is directly related to teacher effectiveness, and that teacher effectiveness can grow when involved in capacity building (Stoll et al., 2006). This capacity building includes the organizational structure of a school or community. There have been many policies enacted by local education agencies in attempts to strengthen effectiveness to improve student learning outcomes (Vescio et al., 2008). School leaders are tasked with providing support and growth for the continued development of teachers. Therefore, the quality and reliability of teacher evaluation processes is critical for continuing improvement of student growth.

Statement of the Problem

One failure of teacher evaluation systems is the treatment of teachers as replaceable parts, an approach termed "the widget effect" by Weisberg et al. (2009). This phenomenon describes the finding from the study that many school districts across the country assume similar classroom effectiveness for all teachers, which means most teachers are rated at a satisfactory performance level with little differentiated feedback that can lead to improved practice. This practice views teachers as interchangeable “widgets” instead of individual professional experts in their field, who deserve a valid, fair, and reliable evaluation that can be used to better inform their individual teaching practice.

Additionally, Weisberg et al. (2009) recommend four considerations to design an evaluation system that better supports educators. First, all teachers deserve a fair, accurate, and credible evaluation that can differentiate teachers based on effectiveness. Secondly, evaluators should be highly trained, focused, and held accountable for their work. While for many districts and states these evaluators are administrators, a system which uses external evaluators could be held to an even higher standard due to the focused training needed to successfully implement an evaluation. Third, evaluation systems should have consequences, meaning low performance evaluation ratings result in additional supports and high-performance evaluation ratings result in advancement. Finally, evaluation systems should provide a space for dialog between teachers and evaluators to discuss openly their progress, areas of growth and professional needs (Weisberg et al., 2009).

One of these evaluation systems is the North Carolina Teacher Evaluation Process (NCTEP), which every K-12 public school educator in North Carolina (NC) takes part in, regardless of content area or grade level. These educators are also all evaluated on the same 30-item Likert scale rubric, called the NCTEP rubric. For most educators in the state, these evaluation services are provided by school administration, where approximately over 100,000 educators receive evaluation services each year (NCDHHS, n.d.).

In the area of early childhood education, many of these teachers and classrooms are not found on traditional school campuses with a principal to evaluate them. For these teachers, the licensure and evaluation procedures must be carried out by external evaluators. In the state of NC, over 900 early childhood educators, licensed Birth-through-Kindergarten (B-K) require mentor, licensure and evaluation services that are not provided by their principals or site administrators. Principals are not expected to be content experts in all areas, nor does a typical administration preparation program include the evaluation of preschool teachers, classrooms, or curriculum.

These services are provided by the Early Educator Support (EES) Office which maintains an evaluation assessment system that supports early educators who need B-K licensure support. In the 2020-2021 academic year, the EES Office conducted over 2,350 individual evaluations, because each B-K teacher receives anywhere from 2-4 evaluations per year. It is essential that these early childhood educators are provided with the appropriate resources to successfully maintain licensure as well as professional development growth. It may be possible to increase the chances that these educators are receiving a fair, reliable and valid evaluation through an advanced statistical model that explores rater behaviors to determine areas of strictness, leniency or bias.

One of the main difficulties in analyzing interrater reliability for teacher evaluation is the need for all the evaluators to identify, understand and consistently rate the same event at the same time in the classrooms. It would be an unrealistic expectation to ask all the mentors and evaluators to travel to the same selection of model classrooms, to observe, at the exact same time to ensure they are each seeing the same behaviors, and then make consistent markings on the NCTEP rubric regarding that educator's teaching quality. Instead to increase interrater reliability, experts in leadership took a selection of artifacts and created ten fictious profiles of classroom settings with photos, videos, and detailed descriptions. These fictious profiles were then marked or rated by each member of the leadership team to determine consistent NCTEP rubric ratings (Lambert, Moore et al., 2021). Research has not been conducted to further understanding of how highly qualified, licensed external evaluators use the NCTEP rubric. The NCTEP standards and elements can be found in Appendix A.

History of the EES Office

North Carolina charged the EES Office in 2007 to provide supports to non-public, early childhood educators. These are educators who do not have access to traditional lead education agencies or administrators to conduct evaluations of their teaching but still need supports, mentoring, evaluations, and other professional development opportunities for growth. Prior to using the NCTEP rubric, the evaluators of these teachers utilized a different evaluation tool known as Pre-Kindergarten/Kindergarten Teacher Performance Appraisal Instrument (PKKTPAI). In 2010, the EES Office began to serve additional educators that were part of the "More at Four" program and began utilizing the NCTEP rubric. The NCTEP rubric is linked to the North Carolina Professional Teaching Standards and is the same evaluation tool used in the public-school setting for Pre-K through 12th grade. Since this time, the EES Office has continued to grow, change, and be transformed. The EES Office began developing a system of evaluation by implementing mentor training and writing a detailed manual to support evaluator practice (de Kort-Young, 2016). Then, a conceptual framework was designed (Taylor et al., 2019) which provided six principles to ensure early childhood educators are receiving appropriate supports from the EES Office. This manual and conceptual framework is used to inform the evaluators to strengthen their own practice. These evaluators are provided with unique supports to provide mentorship. This mentoring component is a unique attribute to the evaluation effort because in other evaluation processes, the principal or other external evaluator, may or may not provide specialized support and content knowledge to the teachers that are being evaluated. However, providing external evaluation and mentor support is an area that the EES Office has extensive experience, and has become the main goal within their work.

The NCTEP refers to the systematic procedure of informal, formative, and ongoing mentoring, coaching, and evaluating of teachers. This process includes classroom observations, summative assessments, and formal mentoring. In 2014, North Carolina's Department of Public Instruction (NCDPI) created two hubs to oversee the early childhood educators that are served as part of the initiative; one at the University of North Carolina at Charlotte and one at East Carolina University. These hubs, named the EES Offices, were given the initiative to oversee the NCTEP for all nonpublic early childhood educators with a B-K license (for additional information about the process, see Bottoms et al., 2021). The EES Office employs experts of early childhood education practice and policy to serve as evaluators. The evaluators utilize the NCTEP rubric, which is composed of 30 total items rated on a 0-4 Likert scale, for all formative and summative classroom observations. There are 5 overarching standards, and each standard has between 3-7 elements, for a combined total of 25 elements. The Likert scale categories are 0- Not Demonstrated, 1- Developing, 2- Proficient, 3- Accomplished, and 4- Distinguished. Further detailed information regarding the rubric is found in Chapter III.

The NCTEP rubric utilizes classroom observations which are one of the most popular strategies for teacher evaluation, based on the authentic and direct feedback from the evaluator to the teacher. However, this process requires a large time commitment by teacher peers or school administrators who serve as evaluators (Ho & Kane, 2013). Thus, it is imperative that these observations are aligned with evidence and theory to support the interpretation of feedback given to the educators.

Rater-Mediated Assessment Theory

This study will utilize Rater-Mediated Assessment Theory as the conceptual framework. The term “rater mediated assessment” is defined as any system of assessments that requires a human rating or score (Wang & Engelhard, 2019). When conceptualizing rater-mediated assessment for this study, it is important to recall that the purpose is to explore rater behaviors, in an effort to ensure that valid, reliable, and fair formative and summative ratings are provided for educators that reflect their true overall teaching quality. While the sample is a group of evaluators, they are referred to as “raters” when utilizing Rater Mediated Assessment Theory. Thus, the term rater is used in the methodology and results to conceptualize through this lens.

One of the tenets of Rater-Mediated Assessment Theory is that professional raters are highly trained, which is defined by Engelhard & Wind (2018) as “raters who have deep understanding of the assessment system” (p. 80). In the context of the EES Office, raters undergo a selection process to make rater judgments for each of the 30 items on the NCTEP rubric. Raters must use multiple types of information from the classroom observation, such as, observing student interactions, teacher conversations with students, and observing instruction to assign a rating. Furthermore, these raters are licensed early educators in the state of North Carolina and hold B-K licenses themselves, so they have previous experience within the classroom setting and, in many cases, operated model classrooms for other early educators. Whereas school principals have a multitude of daily responsibilities in maintaining a school, these expert evaluators focus almost exclusively on applying expert knowledge to the evaluation process using their professional judgment to make appropriate ratings based on the observations or information received from observing the classroom setting. There is a need for an exploration of this rater judgment through the lens of Rater-Mediated Assessment Theory to ensure accurate ratings of classroom teachers.

Method of Analysis

Within Rater Mediated Assessment Theory, the measurement model is called a Rasch model and aims to provide an understanding of the relationship between the measurement of items observed and comparisons of persons (Engelhard, 2013). Rasch models can be appropriate for any assessment that measures a latent construct and meets the assumptions of the model. This method was first explored in the 1960s with statistician, Georg Rasch. It has since been used to understand rater response patterns in the work of Andrich (1978), Wright and Masters (1982) and more recently, Linacre (1994). The Rasch approach aims to determine if the data fit the model closely enough to provide useful information. If the data adequately fits the model, then it is possible to obtain estimates for each facet on the latent variable (Engelhard, 2013).

Strictness, Leniency and Bias in Rater-Mediated Assessments

The Many Facets Rasch Model (MFRM) is appropriate for analyzing rater response data because it allows for a measure of rater performance corrected for strictness and leniency (Bond et al., 2021). In the practice of high stakes teacher evaluations, strictness or leniency can lead to inconsistent scores, which can lead to inconsistent feedback, which can ultimately affect licensure, professional development and even lead to dismissal from the profession. Therefore, evaluation systems should explore their observation ratings for strictness, leniency, or bias. The practice of exploring bias in teacher evaluation is becoming more common and some of the recent literature illustrates that Black teachers are scored lower than their White counterparts. The literature describes how Black teachers are underscored in two scenarios, when they are in schools with predominately White staff members, and when assigned a rater of a different race (Grissom & Bartanen, 2021; Chi, 2021).

Purpose of the Study and Research Questions

The purpose of this dissertation study is to understand the rater response process of early childhood evaluators through gathering and analyzing interrater reliability information to inform evaluator selection, certification, training, and support used by EES Office in North Carolina. This approach utilizes MFRM to explore the strictness and leniency of raters as well as the interaction between rater and teacher race when utilizing both fictitious teacher profiles and actual caseloads. The research questions are:

1. What are the model-estimated ability levels for the fictitious teacher profiles?
2. How variable are the raters in terms of strictness?
3. Is there evidence of differential rater functioning by race of fictitious teacher profiles?
4. How do the ratings of the fictitious teacher profiles compare with the summative ratings of the caseloads in the field?

Significance & Relevance

Teacher evaluation practice varies across the United States, but often includes a performance or observation assessment, requiring a human rater to make a judgment on some scale of teacher quality. Thus, the concept of measurement invariance is essential because the teachers must receive a fair score, that tells an accurate picture of their classroom and teaching practice. A certain teacher’s rating on a scale of quality should be invariant to the rater making the judgments. It is an important step to ensure invariance is present within an evaluation system.

This study explores the measurement properties of the NCTEP rubric with a group of raters utilizing ten, diverse, fictitious teacher profiles, to act as both a training exercise and as an assessment of rater strictness, leniency, and/or bias. The EES Office evaluators make professional judgments using their own content knowledge and teaching expertise. In addition to evaluating teachers, all of the evaluators hold a B-K license themselves. This group of evaluators has extensive knowledge of the NCTEP rubric. Each evaluator has undergone extensive training and continues to participate in ongoing professional developments to further understand the NCTEP rubric and conceptual framework. This rater-mediated assessment is utilized across the state of North Carolina; thus, the reliability and validity evidence are important pieces for ensuring a valid process for early childhood educators.

Delimitations

This study aims to contribute to the body of knowledge that informs the mentor and evaluation services provided to early childhood educators in the state of North Carolina by investigating strictness, leniency, or bias to strengthen the evaluation assessment system. All mentors and evaluators who had served the EES Office for more than one academic year took part in this study. The initial data set was collected in December 2019 when 45 evaluators rated 10 fictitious teacher profiles utilizing the NCTEP rubric. Then, following another academic year, an additional 12 evaluators participated in the same certification activity.

While the scope of this study first aims to support mentors and evaluators across the state of North Carolina, it also seeks to illustrate a model of continuous improvement for the evaluation assessment system utilized by the EES Office. This strengthens early childhood educator practice by ensuring they receive high quality and fair evaluation supports. While this data may be limited to the EES Office evaluators, it has generalizations into teacher evaluation practice. This is due to the generalization of the Rasch model when determining rater response processes, as well as providing an outline to develop an evaluation assessment system of high reliability.

Limitations

One limitation to this study involves the disruption to the EES Office services because of the COVID-19 pandemic. In March of 2020, most of the early childhood educators served still went into their childcare settings daily and provided services for students and families. The group of mentors and evaluators were asked to stop making in-person classroom visits and instead, attempted to provide services in a virtual format. This included virtual conferences, as well as virtual observations. The state of North Carolina made an exception for educators going through licensure renewal during the pandemic year and allowed for a more flexible process. The EES Office provided flexibility to mentors and evaluators due to the unprecedented nature of the pandemic. The goal of providing formal evaluation and licensure renewal was put on pause, and the goal became to support early childhood educators in any way possible.

As of the 2021-2022 academic year, the state of North Carolina has now resumed full evaluation efforts and supports, but the effects of these decisions are continuing to ripple through the classrooms that are served by EES Office. Contextual variables, like the COVID-19 pandemic, can have key effects on personnel evaluation efforts. Within the Personnel Evaluation Standards by Gullickson & Howard (2009), Accuracy Standard 3 calls for an analysis of context, which requires that any and all contextual variables that could have affected an evaluation be identified and discussed. Therefore, the COVID-19 pandemic had a rippling effect on all EES Office staff, including mentors, evaluators, early childhood educators, site administrators, and research staff that is continuing to be seen.

Conclusion

This chapter illustrates that evaluation systems should provide evidence of interrater reliability through the use of sophisticated measurement models because it will strengthen evaluator practice. Research has not been conducted to further understand how highly qualified, licensed external evaluators use the NCTEP rubric. Furthermore, the primary purpose of a strong evaluation system is to provide guidance, support, and information to early childhood educators that will lead to improved learning outcomes. The EES Office has a unique system of support that is rarely found across the United States because it does include external evaluators, whereas Grissom & Bartanen (2021) found that more than 90% of observations are conducted by principals or administrators.

The remainder of this study is organized into four chapters and appendices. Chapter II presents a review of the literature regarding Rater Mediated Assessment Theory, a review of standard practice in the field of evaluation, and a review of literature on rater response methods. Then, Chapter III provides a detailed description of the methodological aspects of this work as well as statistical model descriptions. Next, Chapter IV will include results from four statistical models, and Chapter V will present conclusions and recommendations for next steps to understanding how this work impacts the field.

Chapter II: Literature Review

This chapter provides an overview of relevant literature on foundational principles of reliability, validity, and fairness. Additionally, it will present how interrater reliability is being measured in teacher evaluation contexts through various coefficients, such as exact percentage agreement and Cohen's Kappa. Then, the literature will be presented on Rater Mediated Assessment Theory, a brief history on invariant measurement, the Many Facets Rasch Model (MFRM), and its applications to further understand the validity, reliability, and fairness of rater-mediated assessments.

There are many advantages to utilizing an advanced measurement model to explore these data. First, the model can directly test for invariant measurement which illustrates to the extent that the raters are agreeing or disagreeing when rating the same target respondent behaviors. The model also calibrates the data to a logit scale. This calibration allows for estimated placements that can be compared to each other, providing the opportunity for patterns of strictness, leniency, and potential bias to be identified. Finally, these advanced models can be useful in understanding how raters are making rater judgments which can further inform the evaluation practices used by the EES Office.

Introduction

Presently, there is no standard for percent agreement or interrater reliability within the evaluation components for most state teacher evaluation processes. However, federal guidelines have encouraged the use of interrater agreement measures to ensure that teachers are receiving fair and valid results (Graham et al., 2012). Yet, it is common for teacher evaluation systems to not report validity or reliability statistics (Herlihy et al., 2014). Additionally, rater-mediated performance assessments, such as those used for teacher evaluation in the state of NC, require special considerations regarding the validity, reliability, and fairness of rater judgments. Due to the widget effect described by Weisberg et al. (2009) it is important that teachers are given appropriate feedback based on their teacher performance. The Rasch model, with adequate rater fit, provides evidence for rater-invariant measurement, and provides additional information on the rater judgment process (Wesolowski et al., 2015). There have not been studies conducted that look at these rater judgments within early childhood evaluator practice. One dissertation study conducted in 2012 aimed to determine the extent of interrater agreement among NC elementary principals using Randolph's kappa coefficient but did not conduct a measurement model or explore for leniency, strictness, or bias (Mazurek, 2012). Furthermore, a recent study aimed to determine the demographic congruence between teachers and raters with the NCTEP data from the academic years of 2013-2014 to 2017-2018. However, these data were removed of the 3.87% of evaluations that were conducted by external evaluators (Chi, 2021).

As illustrated, no research has been conducted to date that examines the level of consistency and agreement in a sophisticated, advanced measurement model among early childhood, external evaluators in NC. However, there are national standards that describe optimal personnel evaluation systems that should act as a guide for the state teacher evaluation processes across the country. These standards exist to impact evaluation and assessment systems seen in various educational settings across the United States. Therefore, a review of these standards in practice is presented to give background on the importance of such systems.

National Standards for Interrater Reliability

There are two sets of widely used standards of practice that can work together to promote a fair, valid and reliable evaluation, testing, or assessment process. The first set of standards are the *Personnel Evaluation Standards* (PE standards) (Gullickson & Howard, 2009), which provide guidelines for how evaluators should treat the personnel that they are evaluating. The second set of standards are the *Standards for Educational and Psychological Testing* (EPT standards), which provides guidelines for how assessment systems should be designed (American Educational Research Association [AERA], American Psychological Association [APA], and National Council on Measurement in Education [NCME], 2014). Together, both sets of standards will illustrate the importance of having an accurate, fair, reliable, and valid teacher evaluation process.

Personnel Evaluation Standards

Gullickson & Howard (2009) describes how the Joint Committee on Standards for Educational Evaluation created a total of 27 standards to abide by in personnel evaluation. The PE standards ensure an ethical evaluation process, as well as fair and useful feedback to be provided to those who are evaluated. They ensure feasible and accurate information is gathered from the evaluation. The definition of personnel evaluation is: “the systematic assessment of a person’s performance and/or qualifications in relation to a professional role, and some specified and defensible institutional purpose” (p. 3). This definition indicates that the process of ensuring proper evaluation is not a singular event, but instead a combination of various means to ensure an accurate process for those involved (Gullickson & Howard, 2009).

There are four attributes of evaluation practice laid out by the PE standards: 1) propriety, 2), utility, 3) feasibility, and 4) accuracy. Each attribute contributes something unique to the overall process and will be described in the context of the current study. First, the propriety standards aim to ensure that those who are being evaluated are treated ethically, and that evaluators abide by legal principles. These include standards regarding interactions with teachers, potential conflicts of interest, and other legalities. Second, the utility standards ensure that evaluations are providing constructive feedback that improves overall teacher practice and teaching quality. This includes the qualification measures taken to ensure evaluators are experts in the field, that they provide relevant feedback and valid service to early childhood educators.

Third, the feasibility standards ensure evaluation procedures can take place in an efficient and timely manner. Evaluations occur in a real-world setting, with classroom observations that are both planned and unplanned, so there are a variety of external factors that may influence an evaluation including sickness, field trips, holidays, or other special events in the early education environment. Evaluations should thus be efficient and in alignment with social, political, and/or government forces (Gullickson & Howard, 2009). For the current study, this includes alignment with the NCDPI procedures.

The fourth attribute illustrated by the PE standards is accuracy, and of key importance for this study because these standards ensure the evaluation procedure provided valid and sound information. This includes valid rater judgments, defined expectations for raters, as well as bias management which are all further explored within this study. This also includes statistical measures of reliability and valid, justified conclusions.

Accuracy Standards. Within the PE standards, there are 11 accuracy personnel evaluation standards, as outlined by Gullickson and Howard (2009), each of which ensures technically adequate evaluations that have produced sound information, therefore leading evaluators to make sound judgments. For the purposes of this study, several that are of key importance are noted. Standard A1 describes the importance of having valid rater judgments to minimize misinterpretation (Gullickson & Howard, 2009). In this study, the rater judgments are analyzed as well as statistical measures of reliability via MFRM. The EES Office has been charged with evaluation services for 15 years, thus designing a system of support to provide valid judgments to early childhood educators is at the heart of this work. This system can ensure that rater judgments are correctly interpreted and contribute to educator quality.

Standard A2 refers to the expectations of evaluators and how they are supported in their work (Gullickson & Howard, 2009). These supports are clearly defined by the NCDPI legislature and are outside of the scope of the current study. The requirements for evaluators and early childhood educators include qualifications, responsibilities, and performance outcomes. Standard A3 is discussed in the limitations section of this chapter due to the implications for contextual variables. Standard A4 refers to the evaluation purpose and overall procedures, which are described in the Resource Manual (de Kort-Young et al., 2016). Standard A5 requires that information collected via evaluation be in alignment with evaluation criteria in the field.

Standard A6 requires that evaluation procedures provide reliable information (Gullickson & Howard, 2009). This is one of the cornerstone standards for the present study because it encompasses evaluator strictness and leniency, as well as other reliability measures to be tested across the fictitious teacher profiles. When implementing the interrater reliability activity, each rater was required to rate the same 10 fictitious teacher profiles, on the same 30 NCTEP rubric standards and elements, therefore providing a controlled assessment, to meet the requirements of this PE standard.

Gullickson and Howard (2009) refers to Standard A8 as bias identification and management, which is addressed in the present study via the utilization of the MFRM in the Bias Test model. This model can be specified to include fictitious teacher profile race as well as evaluator race. These bias indices can then illustrate whether an evaluator had strictness or leniency towards White teachers or Teachers of Color, or both.

Furthermore, Standard A9 indicates that information gathered from evaluations should be analyzed correctly. Errors in rater judgments often occurs when the rater is misinterpreting the scale (Gullickson & Howard, 2009). In this context, the NCTEP rubric attempts to measure the latent construct of overall teacher quality. Through the ratings provided, it is possible to measure the extent of unidimensionality across this construct. One way this study will explore the accuracy of the interpretation of the NCTEP Rubric is through the MFRM approach because it assumes that the underlying attribute being measured is a single, unidimensional latent trait (Bond et al., 2021). It is recommended that procedures be implemented to analyze the results of rater-mediated assessments by comparing judgments across raters. Thus, indicating the need for complex measurement models to best ensure correct placements on the NCTEP rubric are made.

The final two PE standards refer to the final phases of ensuring a proper evaluation through justifying conclusions and planning for meta-evaluation. This process describes how to ensure proper evaluations with appropriate supports so that all evaluators and teachers receive a fair, useful, and true measure of their teaching quality (Gullickson & Howard, 2009). Additionally, these final PE standards illustrate how researchers can use the evaluation data to best impact the assessment system over time.

Standards for Educational and Psychological Testing

The EPT standards state that evaluators providing consistent scores ensures validity for the interpretation of the test score and ensures the construct is appropriately measured (AERA et al., 2014). In the context of the NCTEP, these scores indicate teacher quality, therefore high-stakes decisions are being made based on their professional, rater judgment. It is important that these two terms, test and test score, are defined for the purposes of this work. The term “test” refers to any device or procedure that is evaluated and scored. The term “test score” refers to any number, rating, score, or category resulting from an assessment, including ratings on a rubric (AERA et al., 2014). Both terms defined above will be used throughout the study to describe the application of these definitions within the current context.

When understanding the response process of rater judgment, relevant reliability evidence includes the consistency of interpretation of test scores. This consistency can be modeled using an advanced measurement technique, such as the MFRM. The EPT standards (AERA et al., 2014) describe reliability as the consistency of scores that are produced from an interpretation of a test. Reliability can also be described as the degrees of precision when assigning scores (Andrich & Marais, 2019). It is even suggested that validation of such measures should include how judgments are made, as well as analysis of the construct. Reliability for this study refers to the "consistency of scores across replications of a testing procedure" (AERA, APA, & NCME, 2014, p. 33). When analyzing these data for rater response patterns, it is important that the scores be invariant. The term invariant measurement refers to the degree in which the intended constructs are being measured consistently, and if other factors, such as gender or ethnic group, are adding construct-irrelevant variance (Engelhard, 2013).

While exploring reliability, researchers must abide by the EPT standards. Standard 2.7 describes the need for interrater consistency in scoring and within examinee consistency over repeated measures (AERA et al., 2014, p. 44). This standard emphasizes the importance of understanding the process that raters go through when conducting observations and making ratings onto a scale. It also emphasizes that high interrater agreement may not directly indicate high reliability. Engelhard and Wind (2018) connect the EPT standards to rater-mediated assessment by illustrating that investigating measures of reliability, precision, and errors can be accomplished through exploring the invariant measurement properties of items produced from an instrument. According to AERA et al. (2014), “appropriate evidence of reliability/precision should be provided for the interpretation for each intended score use” (p. 42). Thus, if researchers can utilize an advanced statistical model to find evidence of reliability procedures, with a full explanation of interpretations, this can provide a framework for reliability within rater-mediated assessment practice.

Conclusion on the use of Standards

As indicated by the PE standards and the EPT standards, there are systems and structures in place to assist researchers in designing evaluation and assessment systems. Test scores cannot have reliability without first establishing validity; therefore, reliability is a pre-condition for the validity of the test, which refers to the interpretation of scores and explores evidence that exists for the proposed use of the test (AERA et al., 2014). Validity can also be described as the extent to which an assessment measures the intended underlying trait (Andrich & Marais, 2019). Another important aspect to ensuring reliability and validity is fairness. Fairness requires that all populations of test-takers receive a valid test score interpretation for the relevant subgroups within the population. Additionally, fairness refers to any measurement bias or differential item or rater functioning that may exist within the items. By including a bias interaction term within a measurement model, it is possible to measure the extent to which the property of fairness may or may not be violated through the interaction between rater severity and race of teacher in profile.

Both sets of standards require that high quality evaluation systems and structures are put in place as part of a teacher’s evaluation process. These standards exist in the field to give guidelines for the proper implementation and interpretation of test scores across the assessment systems. When determining how to measure the interrater reliability of the evaluators, different measurement models were considered based on the current literature. The next section of the literature review will explore interrater reliability statistics based on Classical Test Theory (CTT), an explanation of why CTT is inappropriate for this application, and then presents the literature on Rater Mediated Assessment Theory.

Measures of Reliability in Teacher Evaluation Contexts

There are three main approaches to exploring reliability theory. CTT depends on the principle that each test-taker has an inherent true score that is an unknown, hypothetical average score over an infinite number of testing replications. Generalizability Theory (G Theory) attempts to measure how different sources of error (e.g., strictness and bias) contribute to the overall error component. Item Response Theory (IRT) relies on the use of information functions, which can indicate the reliability of individual items through reliability coefficients (AERA et al., 2014). Within IRT, the 1 parameter logistic (1PL) model is mathematically equivalent to the Rasch model.

Interrater reliability is one of the main types of reliability that is measured when conducting teacher evaluations because it attempts to measure the level of consistency when two or more raters make a rating on a rating scale (AERA et al., 2014). In the state of NC, all teacher evaluators use the NCTEP rubric as the classroom observation instrument. This means that all teachers, regardless of content area or specialty, are evaluated on the same 30 constructs (North Carolina State Board of Education, 2018). These instruments require raters to make judgments on a rating scale that attempt to measure the unidimensional, underlying construct of teaching quality based on certain indicators. This type of instrument is known as a rater-mediated assessment (Engelhard & Wind, 2019).

Rater-mediated assessments require human judgment to be used, which may open the possibility for construct irrelevant variance. Specifically, this means rater effects may be introduced into the sets of scores and allow for unwanted variance. Thus, measures of interrater reliability are an important piece of the overall message of teacher evaluation processes. There is more than one way to calculate traditional interrater reliability statistics, including exact percent agreement, weighted Kappa, Cohen’s Kappa, Gwet AC1, Gwet’s AC2, and interclass correlations (ICCs) (Holcomb et al., 2022; Jimenez & Zepeda, 2020). The most simplistic of these interrater reliability statistics is exact percent agreement.

Exact Percent Agreement

Zepeda and Jimenez (2019) identified three main areas of concern regarding the lack of validity in teaching evaluation instruments. First, evaluations are completed infrequently and for too short of an observation period to produce valid results. It is difficult to measure all of the many aspects of teacher quality across one evaluation tool. Second, observation instruments are limited in their ability to determine teacher performance across all levels of practice. Third, they point out the limitations for administrators alone conducting observations.

In a study aimed at understanding the most appropriate use for interrater reliability statistics, Jimenez and Zepeda (2020) explored the typically reported measures of interrater reliability to understand the impact of Kappa with these data. In the study, 42 principals and assistant principals participated in training sessions on how to use the observation tool and were tasked with observing four videos of real lessons delivered by teachers in the district. The results found a Fleiss’s Kappa of .335, Gwet’s AC1 of .595 and, an overall percent agreement level of .69, which is a moderate agreement level. As evident from the results, the Fleiss’s Kappa is much smaller than the Gwet’s AC1, which does indicate that these data reflect the Kappa paradox. The Kappa paradox is described as the effect of high agreement statistics resulting in a low Kappa statistic (Gwet, 2012). The study suggests that more stable interrater reliability coefficients are needed in educational studies, more advanced measures of reliability are needed, and Gwet's AC1 is a more accurate statistic of rater-mediated, observational assessments (Jimenez & Zepeda, 2020).

In a different study with the same population, the participants evaluated the teachers using a well-known evaluation instrument, on which they had undergone professional development and training. The data showed that for high-performing teachers, there was a total percent agreement of 84% and for lower-performing teachers, a total percent agreement of 52%. Also, the researchers analyzed the results by standard on the evaluation instrument used. This evaluation instrument has six or seven overarching standards which are unique to the observation instrument. The results indicate variability across these standards, indicating that some may be easier for evaluators to rate than others. The results showed evaluators were able to identify high-performing teachers with consistency but were unable to determine qualities that defined lower-performing teachers with consistency. In an effort to find valid measures of teacher effectiveness, principals and assistant principals completed evaluations of four videos of teachers teaching lessons to students, across a variety of grade levels and subjects.

In conclusion, evaluators need more professional development in identifying lower-performing teachers, and exact percent agreement alone may not be enough to determine if ratings of teacher performance are reliable (Zepeda & Jimenez, 2019). In fact, it is argued that it is *only* through more sophisticated measurement models that researchers can attempt to determine reliability. This is partly due to the differences between consistency and agreement. It is possible that a rater is consistent in their scoring practice, even if not accurate in their ratings. Eckes (2015) referred to this as the "agreement-accuracy paradox," which describes the phenomenon that raters who agree or have high reliability may or may not necessarily indicate those same ratings have high accuracy (p. 29). Beyond individual ratings, raters could also show rater behavior that agrees within the group, but still inaccurately measures a given teacher's quality. Therefore, it is important that there is a distinction between the agreement within each other or across the group, and within a standard cut score for agreement. Both types of agreement statistics provide meaningful information about rater response, but they are limited without implication on a larger scope. Accordingly, there could be group wide patterns that would only be found if using more advanced statistical models.

Interrater Reliability Statistics

There are other appropriate interrater reliability statistics, beyond exact percent agreement, but each has advantages and disadvantages. For example, both Fleiss's Kappa and Gwet's AC1 assume chance agreement is possible, but each statistic calculates the chance of agreement differently. Kappa relies on the fact that all ratings are independent, and Gwet's AC1 relies on an assumption that only a portion of the ratings are correct by chance and seems to be a more robust statistic. In the following studies, Gwet's AC1 or weighted Kappa is seen as the preferred interrater reliability statistics. However, researchers conclude that there is a clear need for more advanced statistical models to provide a more detailed picture of evaluator judgment.

Holcomb et al., (2022) conducted a study aimed to understand the interrater reliability statistics of the EES Office evaluator ratings by calculating interrater reliability coefficients of weighted Kappa, Gwet's AC1, and Gwet's AC2. As described above, the Kappa paradox has proven to be an issue when calculating agreement statistics. This research aimed to explore how robust the Kappa statistics were with these data, and how well other chance-corrected agreement coefficients performed relative to benchmark levels of interrater reliability (Holcomb et al., 2022, Table 2). Finally, the reliability levels among evaluators were also compared through ICC methods.

Results for Kappa showed a range of .121 to .790, with an overall sample average Kappa at .486 (*SD*=.160). The study indicates 29 evaluators within a range of "good" to "moderate" agreement. Results for weighted Kappa produced a higher coefficient of .667 (*SD*=.143), and 31 evaluators were about .61 agreement. Next, Gwet's AC1 was calculated at .563 (*SD*=.143) with a range of .167 to .832. Gwet's AC2 requires an assumption that each rating category is used more than twice, which was not the case for these data because the "Distinguished" category is not used for any of the fictitious profiles. So, this is not the best statistic to use for these data.

The last analysis from this study calculated ICC values for exact and adjacent agreement by standard. It is important to note that the researchers have a unique definition of the term, *within one agreement.* This is defined as an exact agreement for the first and last categories, with the two middle (Proficient and Accomplished) categories allowed to have partial credit. This is because a teacher who is given either the Proficient or Accomplished scores, received similar professional supports, additionally, the partial credit model allows for evaluator professional judgment. The ICC values for exact agreement were calculated between fictitious teacher profiles within the five standards on the NCTEP Rubric ranged from .521 to .692. This represents a moderate to substantial levels of agreement (Landis & Koch, 1977), which is greatly different than the Kappa value calculated. For adjacent agreement, the ICC values ranged from .521 to .729, indicating even more agreement. The within one agreement shows better statistics that more closely align with the weighted Kappa statistic.

The results of this investigation have implications for the current study because this is an exploration of the interrater reliability agreement statistics with these data. These agreement statistics can be used in conjunction with the measurement model of this dissertation to illustrate the full picture of the interrater reliability process for the EES Office. Together, these two studies can work together to validate the evaluation systems that have been put in place. Additionally, this study confirms what is seen in the literature about the inadequacies of Kappa as well as the robustness of Gwet's AC1. It is also evident that Gwet's AC2 has shortcomings in real-world applications when not all categories are used. This investigation also lends itself to the development of new interrater reliability statistics, such as Lambda, which has been tested for its application across rater-mediated assessment data (Lambert et al., 2021). In conclusion, it is suggested by this research that a MFRM be conducted with these data to determine rater strictness, leniency, and racial bias.

Limitations of Reliability Measures

It is difficult to illustrate a full picture of rater behavior based on exact percent agreements, or statistical measures of reliability alone. Instead, measures of reliability within teacher evaluation should be viewed as a system of support that provides professional development, training, and re-certification. Utilizing a multi-faceted approach to measure rater response is one of the best methodologies when investigating response patterns because the model can include rater severity, bias, model fit statistics, and category use. Within a teacher evaluation system, like one utilized by the EES Office, Wind and Jones (2019) encouraged independent "researchers and practitioners who design, implement, and evaluate teacher observation systems to use MFRM to gather information that promotes psychometrically sound evaluations of teaching quality" (p. 532). Furthermore, the researchers explained the benefits of adjusting estimates of teaching effectiveness to explore differences in rater severity even when the raters have not rated all the teachers.

In this dissertation study, the actual caseloads of the raters will be matched with data from evaluations of 10 fictitious teacher profiles. This attempts to determine if the behavior that exhibits itself in the fictious profile activity generalizes into the field. This study uses several Rasch measurement models, which are seen across the literature as the best way to explore rater response. Additionally, these models provide implications for the high-stakes decisions that are made based on evaluations.

Rater Mediated Assessment Theory

Rasch Measurement Theory is grounded in the scaling tradition of measurement research, meaning there is an underlying latent construct that can be measured on a continuum. This theory was first used by Georg Rasch and has since been further explored to place logits onto a linear continuum that can reflect rater judgments. This study uses the term Rater Mediated Assessment Theory because it incorporates the Rasch measurement principles within the design of a rater mediated assessment. This section will further review Rasch models and Rater Mediated Assessment Theory.

Georg Rasch (1901-1980) was a Danish mathematician, statistician, and psychometrician who developed a class of measurement models that changed the trajectory of how future researchers would handle item analysis. *Essays on Georg Rasch and his Contributions to Statistics* (Olsen, 2003) describe that Rasch began his career by studying mathematics at the University of Copenhagen and worked for various prominent scientists. Then, he began to study statistics with the famed R.A. Fisher. Afterward, he began lecturing at the University of Copenhagen and became well known for his contributions to the field of statistics, including his work on growth models and the development of the Rasch model.

After Georg Rasch's contribution to the field with the dichotomous Rasch model, statisticians were charged with utilizing these models across the field. Linacre (1994) describes how it was first expanded by Andrich (1978) and Masters (1982) to include polytomous, ordered rating scale categories. These early Rasch models included two facets: the test-taker ability and the item difficulty. However, in many cases, observations are used and thus, a third facet was added to include rater judgment. This third facet extended the Rasch model to include "many facets Rasch models" (Linacre, 1994). It was also at this time that other psychometricians began to run variations of the MFRM, like partial credit models.

Principles of Rater Mediated Assessment Theory

The principle that grounds Rasch measurement comes from Rasch (1960) and states that if Person A has an ability that is greater than Person B, then, Person A has a greater probability of solving an item correctly. Additionally, a difficult item will have a higher probability of Person A answering it correctly than Person B. With this understanding, Rasch models illustrate that "the probability of success is dependent upon the difference between the ability of the person and the difficulty of the item" (Bond et al., 2021, p. 11). For rating scales, Rasch measurement aims to estimate properties of persons and items, it aims to determine how much a person knows by taking the observed raw score and comparing it with the ability estimate (Bond et al., 2021).

Bond and colleagues (2021) further explained that one of the grounding features of Rasch measurement is that the latent variable or trait is expressed in all of the items, which then a rater responds to, and is recorded in the performance rating of the participant to have success with the item. If test performance is primarily based upon the rater response and the item difficulty, then Rasch model principles are present (Bond et al., 2021). Additionally, rater-mediated assessments include the presence of rater severity. Thus, leading the way for Linacre (1994) to use Rasch principles to develop an understanding of rater behavior or judgment. The raters are independent, with their own expertise and experience, who then apply that knowledge to the rating scale and rate performances.

Invariant Measurement. Invariant measurement implies that the estimates of a person's ability are not influenced by which rater scores them, and the estimates of the rater severity are not influenced by which person they score (Wesolowski et al., 2015). Additionally, this translates into a process of accountability to ensure that raters are providing reliable and fair ratings to the educators that they serve.

The idea of invariance goes back to Guilford (1936) when investigating rating scales. The observations made at the time were that persons could rate latent traits on a continuum, that ranged from low/easy to high/hard, in reference to cues on a scale. Then, these latent traits could be measured using a measurement model. This became known as Guttman scaling and was the first type of what is now known as a Wright map (Engelhard & Wind, 2018). The Wright map aims to display the data onto a continuum for visual inspection. The property of invariance is essential because researchers must be able to determine if raters are using the scale in rater-mediated assessments in the same, or different, ways.

There are five requirements of invariant measurement, as outlined by Engelhard (2013), which are then expanded to include rated mediated assessments in Engelhard & Wind (2018). Below, they are combined to illustrate how these five principles impact this study. First, person measurements must be independent of items, as well as independent of raters; this is called item-invariant measurement of persons, or non-crossing person response functions. Second, a more able person must have a higher chance of success on an item, than a less able person on the same item. In the rater context, this is called non-crossing cue response functions and refers to a rater who rates a cue with a higher chance of success than a less able rater. Third, the calibration of items must be independent of persons for calibrations. In the rating scale context, this means rating categories are independent of raters, or non-cross category response functions. Fourth, any person must have a better chance of success on an easy item, than on a more difficult item. This is called non-crossing rater response functions and indicates that the location of raters is independent of persons, cues, and rating scale. Finally, persons and items should be simultaneously located on a single underlying latent continuum known as the Wright map (Engelhard, 2013). For rater-mediated assessments, this includes persons, cues, rating scales, and raters.

Each level of the rating scale has inherent properties that make it distinct. There are properties of items and raters that should be invariant across the rating scale, meaning these are constants across the entire task. When conceptualizing this in terms of fictitious teacher profiles, each profile has invariant properties for a Developing teacher, for a Proficient teacher, an Accomplished teacher, and a Distinguished teacher. When raters then make a judgment rating, that should be consistent and in alignment with the invariant properties of the categories.

Measurement is “about the development of meaningful scores that locate persons on a continuum (latent variable or construct) that can be used to make decisions about each person” (Engelhard & Wind, 2018, p. 7). Measurement allows for us to interpret the data to provide specific supports and mentorship based on the estimates as well as providing context to these data. Consider this scenario, a group of raters rate a certain teacher on the NCTEP rubric and all raters rate that teacher at the Developing level. However, the teacher has a true score at the Accomplished level, indicating the raters did not arrive at a valid score. So, the group of raters arrived at the same interpretation of the latent construct, but it was incorrect. Therefore, these raters are reliable, but not valid. Thus, the score of Developing is not a valid score of that teacher's overall teaching quality. Furthermore, this group of raters has been too strict in their ratings, because the true score is Accomplished, indicating higher quality than Developing.

The Many Facets Rasch Model (MFRM)

The MFRM provides information regarding the effects of rater severity or leniency, rater habits of central tendency, and systematic bias (Guo & Wind, 2021). The next section will define the Rasch model, and briefly provide an overview of how to specify the model. Then, a review of work by Wind and Jones (2019) that explores the strictness, leniency, and rater judgment of principals and assistant principals with a MFRM when making ratings on a teacher evaluation observation tool. This section also includes a presentation of research regarding rater judgments in musical performance evaluation, which is closely related to the procedure used in this research. Finally, an overview of identifying rater bias. Furthermore, this section will justify the method of these analyses for rater training, developing, and revising scoring systems and the interpretations of the test (Guo & Wind, 2021).

The MFRM Equation & Understanding the Wright Map

Linacre (1994) describes that no empirical data fit a MFRM exactly, instead, it provides guidelines for what data must have, to provide fair and meaningful measures on a construct. It can identify areas of the rater judgment process or discrepancies in item difficulty. It does this through a simplistic equation where:

$$\ln\left[\frac{p_{nijk}}{p_{nljk-1}}\right] = \theta_n - \delta_l - \alpha_j - \tau_k$$

p_{nijk} = probability of examinee *n* being awarded on item *i* by judge *j* a rating of *k*,

p_{nijk-1} = probability of examinee *n* being awarded on item *i* by judge *j* a rating of *k-1*

θ_n = ability of examinee *n*,

δ_l = difficulty of item *í*,

α_j = severity of judge *j*,

τ_k = difficulty of the step up from category *k-1* to category *k*

k = 1 to M

0 = lowest point on the rating scale

M = highest point on the rating scale

M+1 = points on the rating scale.

This is relevant for the current study due to the implications for rater professional judgment when using the NCTEP rubric. As described by Eckes (2011), it is reasonable to assume that raters who are utilizing the scale understand how to do so. Raters should be familiar with the scale, understand the many uses, and know ways to provide interpretation from the ratings. All the category items on a scale may or may not be used for each evaluation, and there may be a difference in the number of times a category is used over the others (Bond et al., 2021). The Many Facets Rasch Partial Credit (MFR-PC) measurement model allows for facets that represent potential areas of variance, which then allows researchers to calibrate difficulty thresholds because each item operates as a unique rating scale structure (Wesolowski & Wind, 2019).

A Wright map is an important element to Rasch measurement models because it provides a visual framework for the interpretation of the scores (Engelhard & Wind, 2018). Wright maps serve as a visual interpretation of item difficulty and person ability, as well as any other selected facets. In Appendix B, the case study Wright map from Engelhard & Wind (2018) illustrates facets of student ability, rater severity, and item difficulty of four different writing knowledge skills (style, organization, conventions, and sentence formation). The first column indicates the logit scale, ranging from 8 to -8 logits, with 0 indicating center. Then, the next three columns indicate placement on the logit scale for the three facets of interest. Test takers are placed high to low ability, raters are ranked severe to lenient, and writing knowledge skills ranged from most difficult to least difficult. It is important to note that many raters are centered around 0, indicating a lack of variation in their locations compared to the spread seen by student ability, offering evidence for invariant measurement with respect to rater effects.

Utilizing MFRM in Teacher Evaluation Reliability

Wind and Jones (2019) attempted to validate the use of MFRMs when researchers aim to understand rater judgments of a performance observation system. This study included 114 principals who provided their summative classroom observation scores for the previous academic year and included four example teachers they rated at the beginning of the year for training. The researchers focused on three key teaching practices and analyzed rater behavior on these: (1) Cognitive Engagement, (2) Critical Thinking, and (3) Formative Assessment. There were two measurement models; one basic model with three facets of teacher effectiveness, rater severity, and item difficulty, and then a slightly different model that estimated rating scale thresholds to explore how raters utilized the categories on the scale. The purpose of the analysis was to understand rater severity, rater fit, and rater category use.

The results found that principals exhibited a range of severity when assessing classroom teachers. The study analyzed rater fit with mean square error statistics and flagged eight raters outside the expected range. Finally, rating scale category thresholds determine how similar raters are when making ratings, with the goal to understand how ratings are made as a group. This study illustrates that MFRMs should be used consistently within teacher evaluation systems to ensure teachers of high quality and to ensure that the teacher evaluation efforts are valid and reliable. MFRMs can provide researchers or practitioners the ability to identify rater behavior patterns. In conclusion, this research provides further evidence that interrater reliability measures of percent agreement alone are not enough for the validation of teacher evaluation systems. Instead, it emphasizes the importance of utilizing a sophisticated measurement model to understand rater behavior when conducting evaluations (Wind & Jones, 2019).

Utilizing the MFR-PC in Musical Performance Evaluation

While the content is different, the procedure of using the MFR-PC model in understanding rater behaviors of musical performance judges provides similar methodologies to those used in this study. Raters of musical performances hold great importance to those students and educators involved in the process (Wesolowski et al., 2015; Edwards et al., 2019). The research of Wesolowski and colleagues (2015) investigated data-model fit as well as differential rater functioning to determine if items were invariant to the school level of the performance. The four school levels included middle school, high school, collegiate and professional. A sample of 23 expert jazz instructors were given a total of four performances to rate, with two of each recording overlapping between two of the raters. For example, rater 1 rated performance numbers 1, 2, 3, and 4, while rater 2 rated performance numbers 3, 4, 5, and 6. Additionally, the data included an expert rater who rated each performance. This totals to 48 jazz performances being rated (Wesolowski et al., 2015).

The results of the differential rater functioning (DRF) in this study showed that raters may use the rating scale inconsistently across school level groups. The model indicated a wide range in rater severity ($\chi^2(23) = \mathit{10}$, p<.05; *Rel*=.83), with 10 individual raters identified as demonstrating differential rater severity by school level. Specifically, this model allows for individual rater patterns to be explored. Rater 8 in this study shows patterns of consistent overestimation of collegiate level performances and an underestimation of professional level performances. The researchers further explored rater patterns to conclude that rater 8 had patterns outside the probabilistic model, specifically rating professional ensembles on measures of balance and time-feel higher than would be expected by the model.

The results of these data indicate that patterns of interactions for some raters were inconsistent, which could not have been identified through other reliability statistics alone. While the model shows evidence of bias, and can better inform musical performance rater practice, it is always cautioned that this is an interpretation of statistical information and is not a means of causality. While able to detect DRF, this is only a potential source of bias without further investigation. In conclusion, the researchers make three recommendations for rater practice in musical performance evaluations: 1) provide rater training, 2) provide exemplars and anchors to the scale, and 3) a benchmarking/cut score system to be developed (Wesolowski et al., 2015).

Conclusions of Rater Mediated Assessment Theory

In conclusion, MFRMs will provide evidence of rater response patterns and provides evidence of strictness, leniency, and potential bias to be used by the EES Office to support evaluators. This theory gives grounding principles to the current work because it allows for rater response patterns to be explored in more depth. Additionally, MFRMs can provide more insights about overall patterns of rater response that traditional reliability coefficients. The use of MFRMs in this study will allow for an interaction term to be placed on the facets of rater severity and race of teacher profile to indicate areas of strictness or leniency towards a certain subgroup of the population.

Rater Bias

It is suggested that bias should be independently investigated to determine if it is present and how it is or is not connected to the test and/or testing procedure (AERA et al., 2014). A

MFRM gives researchers the ability to test for bias within the context of the model. For this study, evaluators are seen as experts in the field, and any sort of bias should be investigated to ensure a valid process is in place for early educators. The concepts of rater bias and fairness intertwine because interpretations of scores varying across subgroups can be an indicator of bias or unfairness towards a group. This must be explored independently to determine sources of construct-irrelevant variance. Additionally, it is important to ensure that bias is always viewed as an area of "potential" bias, and not a clear indication of bias, when examining rater judgments (Wesolowski et al., 2015).

Eckes (2011) describes how raters can be viewed as independent experts. The implications are twofold. First, this indicates that each rater acts independently of the other raters, thus having their own set of behaviors and patterns to their responses. Second, this indicates that each rater is a content expert in the field and has expertise with utilizing the NCTEP rubric. Both factors, having raters that are independent and who are content experts, are properties that ensure invariant measurement when understanding the rater response process of subgroups of the evaluator population (Engelhard & Wind, 2018). Rater independence can be measured through the MFRM. It is necessary to then compare the observed proportion of exact agreements to the expected proportion, and if the observed proportion is approximately equal to the expected proportion, then there is evidence for rater independence. However, if the observed proportion is much larger than the expected proportion, the raters are acting too similarly (Eckes, 2011).

The idea of rater independence is a key assumption of the MFRM and provides support for why it is important to analyze for potential rater bias. If the raters are making judgments that are awarding unfair scores to a particular teacher, it is important that it is addressed to ensure best practice. Additionally, we expect raters to have a range of professional judgment or a range of professional ratings, but this range should not also indicate any sort of potential bias.

Eckes (2011) described the term rater variability as any variability awarded to examinees that can be associated with characteristics of the rater and not with the performance of the examinee. Similar terms in the literature include rater bias, rater effect, and rater error. Hoyt (2000) defines rater bias as a different interpretation of the rating scale or a unique perception of the target response. However, if raters are trained effectively, this rater bias may be minimal (Praetorius et al., 2012). It is important to understand rater response because it leads to a more informed practice of support for educators. It is recommended that to have a valid, high-quality approach, it should include rater training, repeated ratings on the same educator, and establish interrater reliability through various statistical means (Eckes, 2011).

As indicated earlier, the Rasch model can directly test for invariant measurement, which is defined as the "equivalence of a construct across groups or measurement occasions" (Putnick & Bornstein, 2016). This can be conceptualized as a lack of rater effects. Research indicates that different groups may show bias in their ratings. It was recently found that white teachers receive an average of .15 standard deviations higher of an observation rating than their Black colleagues (Grissom & Bartanen, 2021). This study used observational ratings generated by the classroom observation, as well as value-added-measures like student achievement, to determine teacher quality score. The researchers took the results and determined how many Black teachers would have received a higher rating if this gap did not exist. The results indicated 1300 teachers would have received a higher score of teacher quality. Additionally, it was found that Black teachers are scored substantially lower when racially isolated within a school. The main conclusion from this work indicates that policymakers must understand where the bias is located so that it can be addressed. Therefore, there is a clear advantage to using a Rasch model for investigating racial differences because of the vast amount of information it provides regarding rater response, rater behavior and item characteristics.

Context for the Current Study

The NCTEP rubric is used across the state of NC for all educators regardless of grade or content area (NC SBE, 2021). The rubric has five overarching standards with a collection of elements under each for a total of 25 elements (Appendix A). The entire rubric thus consists of 30 items on a four-point scale: (1) Developing; (2) Proficient, (3) Accomplished, and (4) Distinguished. There is also a “Not Demonstrated” option for formative evaluations, this is not used during final summative evaluations. The five overarching standards are Standard 1: Teachers Demonstrate Leadership; Standard 2: Teachers Establish a Respectful Environment for a Diverse Population of Students; Standard 3: Teachers Know the Content They Teach; Standard 4: Teachers Facilitate Learning for Their Students; Standard 5: Teachers Reflect on Their Practice. For more information regarding the history of the EES Office or NCTEP rubric, see Bottoms et al., 2021.

Each year, evaluators in the EES use the NCTEP rubric for documenting formative growth and progress, as well as summative evaluations and performance ratings to early childhood educators regarding how the NC Professional Teaching Standards are exhibited. This includes classroom conditions, teacher, and child behaviors, naturally occurring artifacts, and documented evidence. Mentors and evaluators are dually trained and provided continual professional development opportunities leading to an in-depth familiarity with the NC Professional Teaching Standards, the NCTEP rubric, and the evaluation process. These efforts are implemented to best understand and support the unique contextual challenges of each teacher.

The EES Office aims to accurately interpret observations in the classrooms and place all teachers on the rubric with equal validity. This requires evaluators to fully engage in training, preparation, and practice to develop professionally as skilled observers and evaluators who function well within the complex nature of teacher performance evaluation and the intricacies of working effectively with adult learners. While the task of evaluation is independent in nature, the role itself is carried out by way of a system of support for each early childhood educator, whose team includes the evaluator, the mentor, and the site administrator. Each mentor and evaluator are also part of a regional team that provides its own internal system of support through training, leadership, collegial support, and job shadowing as they continue to learn and grow in their dual roles as mentor and evaluator.

Since the 2017-18 program year, the EES Office has engaged in coordinated phases to support the development of a comprehensive process of reaching reliability among our collective teams. Products of the implementation phases included the development of a conceptual framework, guiding principles, co-observation process, and an exploration of interrater reliability measures. As well as providing individual and group professional development that was informed by the data collected throughout each phase. Ultimately, the components of all phases led to a culminating event and research study in which all mentors and evaluators participated in an interrater reliability exercise to become certified in their role as an evaluator for the EES Office.

Conclusion

The MFRM is situated within Rater Mediated Assessment Theory because it allows for multiple facets of evaluator behavior to be calibrated. The principles of validity and reliability within a rater-mediated assessment, like the NCTEP rubric, are essential to better understanding its use. There are many statistical measures of reliability, but some of them are very limited in their scope and lack the ability to tell the full story of the data. Therefore, a more advanced model is necessary. For example, the drawback to using Kappa coefficients with observational ratings due to the agreement-accuracy paradox.

This literature review summarized the principles of reliability, validity, and fairness of the EPT Standards (AERA et al., 2014) and PE Standards (Gullickson & Howard, 2009). There are limitations to the validity, reliability, and fairness of the CTT interrater reliability statistics, like exact percent agreement, Kappa, Gwet's AC 1 and AC 2, or ICCs. These measures cannot explain rater behavior in the same way that Rasch measurement theory can by allowing for an exploration of items by placing them on a logit scale score. Rasch measurement theory improves objectivity in the assessment process (Wesolowski et al., 2015) and the steps suggested by these authors indicate rater practice that is already utilized for the EES Office.

The MFRM studies presented by Wesolowski and colleagues (2015), as well as in interrater reliability studies like Zepeda and Jimenez (2019) both illustrate the importance of understanding the reliability, validity, and fairness of evaluator data. The MFRM is a more appropriate way to tell the full story because it places the items onto logits to make them comparable. The use of MFRM within the teacher evaluation context is rare, but present, even if the studies are mainly with principals and not external, expert, evaluators. In conclusion, this literature review has emphasized the need for an advanced statistical model to be applied to these data. The next chapter will explore the methodology utilized in this study.

Chapter III: Methodology

Introduction

It has been noted that there are national foundational standards, like the PE Standards and EPT Standards which can guide the development of assessment systems (AERA et al., 2014; Gullickson & Howard, 2009). The literature in the previous chapter explored the limits to the depth of information that traditional interrater reliability statistics, like Kappa, provide for researchers. This was seen in the work by Jimenez and Zepeda (2020) where it was suggested to use more advanced models of reliability, like Rasch models. It is also suggested that a combination of both interrater agreement statistics and an advanced statistical model be used to determine the full extent of reliability for a teacher evaluation system. An exploration of the interrater agreement statistics for these data has already been conducted (Holcomb et al., 2022). Thus, indicating the need for the application of an advanced measurement model to these data.

When researchers aim to investigate the rater judgment process, the full story of the presented data is best illustrated through empirical models, such as the MFRM. The literature review presented in the previous chapter sets the stage for using Rater Mediated Assessment Theory to understand rater response and judgment patterns of potential strictness, leniency, and/or bias through the utilization of MFRMs. These types of models can provide detailed accounts of the reliability, validity, and fairness within the scores from an observational instrument. This chapter provides details on the methodology of this work, beginning with an overview of the purpose and research questions of this study. Then, a description of the research design which includes the population, sample and instrumentation is provided. Next, an overview of the analysis and descriptions of the three models with evaluation criteria are presented. Finally, an overview of limitations and conclusions.

Purpose and Research Questions

The purpose of this dissertation study is to understand the rater response process through gathering and analyzing interrater reliability information to inform evaluator selection, certification, training, and support used by EES Office. This approach utilizes three MFRMs to explore the patterns of strictness and leniency among evaluators as well as the interaction between evaluator strictness and race of fictitious teacher profile.

The first model aims to answer research question 1 and 2 and is named the Strictness Calibration model. The second model aims to answer research question 3 and is named the Bias Test model. The third model aims to answer research question 4 and adds the full set of summative ratings from each evaluator's actual caseloads to their ratings of the 10 fictitious teacher profiles to analyze rater response patterns. This model is named the Caseloads Added model. Additionally, this model investigates the generalizability of the rater behavior seen in the fictious teacher profile activity to their rater behavior in the field.

This study aims to investigate the following research questions through the use of the MFRMs, which provide logit-scale location estimates, infit and outfit mean square errors, and separation statistics. The study will also investigate graphical displays of statistics via Wright maps.

1. What are the model-estimated ability levels for the fictitious teacher profiles?
2. How variable are the raters in terms of strictness?
3. Is there evidence of differential rater functioning by race of fictitious teacher profiles?
4. How do the ratings of the fictitious teacher profiles compare with the summative ratings of their caseloads in the field?

Population & Sample

The group of participants in this study were 57 evaluators, employed by the EES Office, with at least one year of experience as an evaluator prior to participation. These 57 evaluators indicate the full population of evaluators at the end of this study. These are highly qualified, B-K licensed evaluators, who take part in annual professional development to maintain expertise in the field. To ensure evaluators are providing consistent ratings, 10 fictitious teacher profiles were crafted using real data and artifacts from real classrooms that are served by the EES Office.

These profiles include classroom scenarios, complete with photographs, student work samples, written accounts of situations seen in the field and videos. All 10 profiles were rated on all 30 items on the NCTEP rubric. These data combined one group of evaluators (*n*=45) that participated in the fictitious profile activity in December 2019, and another group of evaluators (*n*=12) that participated in July 2021. Both groups of evaluators rated the same 10 fictitious teacher profiles. The Strictness Calibration model and Bias Test model includes these fictitious profile ratings only.

For the Caseloads Added model, the summative caseloads are added to the fictious profile data, which represents the entire population of B-K licensed teachers that were evaluated in the state for the 2020-2021 academic year. This combined data set includes a total of 552 teachers and 10 fictitious profiles for a total of 562 teachers. The number of evaluators (*n*=57) and number of items on the NCTEP rubric (*n*=30) remain the same.

Instrumentation

In the state of North Carolina, teachers who teach in non-public, preschool settings while holding a B-K License will require evaluations for license renewal. These teachers require evaluation supports offered through the EES Office. The NCTEP rubric is used to evaluate all levels of educators in the entire state. This means pre-Kindergarten teachers are assessed on the same rubric as high school English teachers. The rubric helps identify performance levels corresponding to instructional practices, as they relate to the NC Professional Teaching Standards.

The NCTEP rubric consists of five overall standards with a varying number of elements per standard, for a total of 25 elements. The breakdown of all 30 items on the rubric by standards and elements can be seen in Appendix A. Each of the items is rated on a 1-to-4 item Likert scale: 1) Developing, 2) Proficient, 3) Accomplished, and 4) Distinguished. There is an option to select 5) Not Demonstrated, for formative observation purposes, but this was not an option for the reliability activity because evaluators were expected to rate each fictitious profile on all standards and elements. This provides matched observational ratings for all evaluators on all items.

As stated, the data set includes all 57 evaluator ratings across 10 fictitious teacher profiles on 30 elements. Therefore, each evaluator provided 300 rater judgments. While these ratings are provided by external evaluators, to align with Rater Mediated Assessment Theory, the term "rater" will be used for the remainder of the study to describe their rater behaviors, including strictness, leniency, and potential bias.

For the Strictness Calibration model and the Bias Test model, the facets include fictitious teacher profile ability, item difficulty, rater severity, and race of fictitious teacher profile. In the Bias Test model, an interaction term is added to the facet of rater severity and race of fictitious teacher profile to explore for potential bias. Then the Summative Caseloads model adds the fictitious teacher profile ratings with the real, summative caseloads from classroom observation ratings given to B-K licensed educators in May 2021. By combining the fictitious teacher profile data with the actual early childhood educator caseloads, it is possible to determine how the models compare in goodness of fit, infit and outfit *MSE* statistics, and visual inspection.

Data Collection

The participants of this study did provide their consent to participate in the fictitious profile rating activity. Participants were told how the data from the activity could help inform and improve their evaluation practice. They were asked to complete the profiles over a 10-day period and given time away from their classroom observation duties. All participants were given the profiles in the same order. Prior to the use of these profiles in this activity, these profiles underwent panel discussions to ensure high quality profiles that accurately represents the classrooms that are served by the EES Office (Lambert, Moore et al., 2021).

Data Analysis

The Rasch model analysis is conducted using FACETS software (Linacre, 2020). This software uses a maximum likelihood estimation to determine locations of both persons and items, which is called a joint maximum likelihood estimation (JMLE). This estimation procedure iterates between item calibration and person measurement to jointly obtain parameter estimates. The original Rasch model includes two facets: test-taker ability and item difficulty. However, the NCTEP rubric utilizes ordered, polytomous responses on a Likert scale of one to four. The MFRM allows for multiple facets to be associated with ordered, polytomous responses (Eckes, 2011). The first application of MFRM within rater-mediated assessments was utilized by Linacre (1994) and has since been used across multiple disciplines including educational and psychological measurement (Engelhard, 2013). This procedure is appropriate for analyzing rater response and judgment data because it allows for a measure of evaluator performance corrected for strictness or leniency (Bond et al., 2021).

These MFRMs in the study are calibrated as Partial Credit models, which allows for each facet to be calibrated independently as a unique rating scale structure. The Many Facets Rasch-Partial Credit (MFR-PC) Model allows for facets that represent potential areas of variance, which then allows researchers to calibrate difficulty thresholds because each item operates as a unique rating scale structure (Wesolowski & Wind, 2019). The partial credit model has an advantage with these data because each category on the rating scale (Developing, Proficient, Accomplished, Distinguished) is not used equally. It is more common for educators in the field to have summative ratings of Proficient or Accomplished than it is for them to be Developing or Distinguished. The educators are expected to receive at least a rating of Proficient to renew their licensure. Additionally, a summative rating of Developing would indicate that educator is still lacking key skills to be considered Proficient, which could indicate low teaching quality, and even dismissal from the field. A summative rating of Distinguished would indicate that educator is going above and beyond the expected range of teaching quality.

With this in mind, the MFR-PC model allows for the facets to individually vary with the estimates. This allows for researchers to determine fit by individual facet and it is possible to determine if raters are using the rating scale categories as expected. The MFR-PC model will yield useful information with good model-data fit. This evidence will either indicate evidence of invariant measurement or evidence that shows a lack of invariant measurement (Engelhard & Wind, 2018).

The work of Wesolowski and colleagues (2015) provides a clear procedure for analyzing these data because they use MFR-PC to explore model-data fit, rater errors and differential rater functioning with FACETS software. In the context of the present study, invariance can be conceptualized as the extent to which ability estimates of fictitious teacher profiles are invariant to which rater rates them and estimates of rater strictness and leniency are invariant to which fictitious teacher profile the rater is rating.

One of the advantages to the family of Rasch models is that it transforms the data onto a common, log-odds scale. The model "uses probabilistic distribution of responses as a logistic function of person and item parameters in order to define a unidimensional latent trait" (Wesolowski et al., 2015, p. 151). The unidimensional latent trait for these data is teaching quality (Lambert, Bottoms, & Holcomb, 2021). This transformation onto the same scale allows for each facet to be displayed together, which can then be conceptualized more traditionally as independent and dependent variables, with each facet acting as independent variables and the logit scale score as the dependent variable. This provides a way to conceptualize traits that are difficult to measure, such as rater strictness, item difficulty, fictitious teacher profile ability, and potential rater bias. Furthermore, if any two teacher profiles are found to have the same teaching quality, but group differences are present, this could be due to some amount of rater error. Therefore, empirical evidence should be investigated to understand if any individual biases are present (Wesolowski et al., 2015).

The purpose of this study is to use three different MFRMs to investigate rater strictness, leniency, and potential bias. The first model is the Strictness Calibration model and includes fictitious teacher profile ability estimates, severity of rater, difficulty of item, and race of teacher profile. Strictness Calibration model answers research question one and two because it will place fictitious teacher profiles onto a logit scale to determine overall quality and models the strictness of raters on these fictitious profiles. The Bias Test model builds upon the Strictness Calibration model to include an interaction term on severity of rater and race of teacher profile. This Bias Test model helps explore rater behavior in the context of the race of fictitious teacher profile and aims to answer the third research question. Finally, the Summative Caseloads model is used to illustrate the extent to which the fictitious profile ratings align with ratings of actual early childhood educators in the field. This model answers the fourth research question because it examines the extent to which the behaviors displayed by the raters when completing the fictitious profile training activity are also displayed when evaluating teachers within the field.

General Evaluation Criteria for Rasch Models

Each MFRM will produce logit-scale placements for each facet, infit and outfit mean square errors (*MSE*) and separation statistics. Researchers can use *MSE* infit and outfit statistics to analyze the amount of misfit present. Engelhard and Wind (2018) describe an acceptable range of infit and outfit *MSE*s of 0.5 to 2.0. However, due to the high stakes' nature of this work and to ensure that early childhood educators are receiving reliable, valid, and fair evaluations, more precise measurement statistics are used. In this study, infit and outfit *MSE* statistics have an acceptable range of 0.7 to 1.3, as suggested by Stemler and Tsai (2008). Infit *MSE* statistics that are less than 0.7 indicate less variation in the responses than is expected. Infit *MSE* statistics that are larger than 1.3 indicate too much and unpredictable variation (Stemler & Tsai, 2008). While MFRMs produce both infit and outfit statistics, infit statistics are able to provide more information because more weight is given to the performances of persons closest to the item's difficulty value (Bond et al., 2021). Outfit statistics are based on the conventional sum of squared standardized residuals, infit statistics are a weighted sum (Bond et al, 2021).

Engelhard & Wind (2018) address the importance of determining model-data fit for MFRMs. First, data-model fit is highly dependent upon the context of the study, the nature of the rating scale and how the assessment system was designed. It is vital that interpretations of misfit are analyzed in light of the latent variable that is being represented. Secondly, there are communities of practice within Rater Mediated Assessment Theory that have agreed upon guidance regarding the best use of data-model fit. The model produces chi-square statistics for model fit; however, experts in the field indicate the importance of using a variety of visual and statistical information to determine model data fit (Engelhard & Wind, 2018). Thus, it is the responsibility of the researcher to explore techniques to determine areas of misfit within standard practice. Finally, determining numerical estimates of fit is only one piece of the model and visual graphs, like the Wright map, should be utilized.

Strictness Calibration Model

The first research question aims to understand the model-estimated ability levels for the fictitious teacher profiles and places the 10 fictious teacher profiles onto a logit scale by ability. This allows for researchers to explore if the fictitious profiles reflect the amount of variability in the quality across NC B-K teachers. If the profiles exhibit a range of variability that matches the distribution of teacher quality in the field, this will provide evidence that the fictitious profiles are well-constructed. Well-constructed indicates that the fictitious profiles match the variety of classrooms seen in the field and indicates that the profiles have consistent ratings. The second research question aims to calibrate rater strictness and leniency. The model will standardize the individual rater statistics to a logit score where a 0 indicates no evidence of strictness or leniency, and anything outside the range of -0.5 logits to 0.5 logits is considered strict or lenient.

The Strictness Calibration model provides empirical evidence of reliability, validity and fairness and answers the first two research questions. This model is the baseline to understand to the extent invariance exists across a group of raters in terms of strictness and fictitious teacher profile ability. The following equation follows the general form from Wesolowski et al., (2015), and has been adapted to the context of the current study:

$$\ln\left[\frac{p_{nijmk}}{p_{nijmk-1}}\right] = \theta_n - \lambda_i - \delta_j - \gamma_m - \tau_k$$

Where:

p_{nijmk} = the probability of fictitious teacher profile *n* rated by rater *i* on item *j* with race *m* receives a rating in category *k,*

$p_{nijm(k-1)}$ = the probability of fictitious teacher profile *n* rated by rater *i* on item *j* with race *m* receives a rating in category *k-1*

θ_n = the logit-scale location (e.g., ability) of fictitious teacher in profile *n,*

λ_i = the logit-scale location (e.g., severity) of rater *i,*

δ_j = the logit-scale location (e.g., difficulty) of item *j,*

γ_m= the logit-scale location (e.g., ability) of race of fictitious teacher in profile *m,*

τ_k = the logit-scale location where rating scale categories *k* and *k-1* are equally probable for rater *i.*

Within this analysis, it is expected that a spread of fictitious teacher profile ability levels is seen across the logit scale. This spread should reflect what is seen in the field, which would mean all four levels of teacher quality (Developing, Proficient, Accomplished, Distinguished) are displayed. The spread of rater severity should be minimal, with an ideal placement on logit placement 0, indicating no strictness or leniency. However, due to human raters and ranges of professional quality, the researcher is utilizing a small range (-0.5-0.5) of tolerable variability within the logit-scale location. This intends to take into account the professional judgment that is present for these raters. It is expected that these raters are not always in exact agreement because they are human raters, but small variations are acceptable. If raters fall outside of this range, then they have indications for either strictness or leniency.

Finally, the model uses both statistical means and visual inspection to determine model fit. Experts in the field indicate the importance of using a variety of visual and statistical information to determine model data fit (Engelhard & Wind, 2018). So, a visual inspection of the Wright map will be used as well as an exploration of the chi-square fit statistics.

Bias Test Model

The third research question will analyze fairness using an interaction term to explore for evidence of differential rater functioning (DRF). DRF can determine if raters are exhibiting patterns of strictness or leniency across subgroups of the population (Wesolowski et al., 2015). For this study, that subgroup is by race of fictitious teacher profile. If there is no DRF by race, then this group of raters will successfully be able to rate these 10 fictitious teacher profiles without displaying evidence of potential racial bias. If the results yield statistically significant interaction terms, then evidence of potential racial bias is present. This interaction term is between the facets of rater strictness and race of fictitious teacher profile. For example, if a rater shows evidence of strictness towards both teachers of color and white teacher profiles, this may not be evidence of racial bias and may be evidence of overall strictness. However, if a rater is lenient towards White teachers and strict towards teachers of color, this could be evidence of bias because different subgroups of the population are receiving statistically different results. This bias can lead to misinformed teacher practice and systemic bias.

The Bias Test model includes an interaction term to address the third research question. This model allows the researcher to understand the extent to which rater severity is invariant to race of teacher profile. The following equation also follows the general form of Wesolowski et al., (2015), has been adapted to the context of the current study, and includes some of the same variables as the first model:

$$\ln\left[\frac{p_{nijmk}}{p_{nijmk-1}}\right] = (\theta_n - \lambda_i - \delta_j - \gamma_m - \tau_k) - \gamma_m\delta_j$$

Where:

$\gamma_m\delta_j$= interaction between rater severity and race of fictitious teacher in profile

The third research question can be answered through the overall bias statistics produced by FACETS (Linacre, 2020). This bias statistic is suggested to be below 1.0 (Zwick, 1999). However, to indicate the high stakes' nature of this work, the bias statistics will have an acceptable range of -0.5-0.5. Thus, any rater with a bias statistic outside that range will be flagged for potential bias.

Summative Caseloads model

The fourth research question aims to compare the Strictness Calibration model with this Summative Caseloads model, which includes all the real, summative, caseloads of evaluation ratings for these raters. It is possible to compare these two models by comparing the infit and outfit *MSE*, the logit scale locations for item difficulty, rater severity, and teacher ability, as well as separation statistics. It is possible to investigate across all facets to see the differences in the model estimates when the summative ratings are added. The ideal results would calibrate each facet in a similar manner to the first model, indicating that raters are consistent in their rater behavior because their behavior in the fictitious profile activity matches their behavior out in the field. It will be able to determine which raters have different strictness estimates than the first model because it will place the data onto the logit scale.

The third model removes the interaction term between rater severity and race of fictitious teacher profile, and the facet of race. These indices were removed from the original model because the racial identity of the early childhood educators in the summative caseloads is not known. This model can evaluate if the teacher ability estimates, rater severity or item difficulty remain consistent when the ratings are added to the data. This may identify rater behaviors of strictness and leniency within their practice, and the extent to which these behaviors are generalized in the field. The following model was used:

$$\ln\left[\frac{p_{nijk}}{p_{nijk-1}}\right] = \theta_n - \lambda_i - \delta_j - \tau_k$$

Where:

θ_n = the logit-scale location (e.g., ability) of teacher in profile and summative ratings n

The fourth research question will be explored by comparing fit statistics among teacher profile estimates, item difficulty estimates, and rater severity estimates. Each of these estimates that are displayed in the Summative Caseloads model will be compared to the Strictness Calibration model. If any of the caseload model statistics are larger than the absolute value of 0.5, they will be flagged as an area of misfit.

Limitations

One limitation to this study involves the disruption to the EES Office services because of the COVID-19 pandemic. Most of our educators were still teaching in person at childcare settings and providing services for students and families. However, the group of mentors and evaluators were asked to stop making in-person classroom visits and instead, attempted to provide services in a virtual format. An exception was made for educators going through licensure renewal and they were given a more flexible process. There was a shift within the EES Office. The goal became supporting early childhood educators in any way possible, instead of providing formal evaluation and licensure renewal.

As of the 2021-2022 academic year, the EES Office has resumed business as usual evaluation efforts and supports, but the effects of COVID-19 are continuing to ripple through the classrooms that are served by EES Office. It is important that any contextual variables, like global pandemics, are mentioned within research. This is even more important when dealing with evaluation of personnel (Gullickson et al., 2008).

Conclusion

The purpose of the study is to understand rater behaviors in the context of fictitious teacher profiles as well as an exploration of the extent to which these behaviors may be seen in the field. This analysis uses three MFRMs to explore areas of strictness, leniency, or bias as well as provide a common scale so that these different facets can be compared to each other. These results will contribute to the field by informing rater practice so that specific, targeted supports can be provided. Additionally, this analysis allows for areas of potential rater bias to be identified to ensure all early childhood educators are being provided with fair, valid, and accurate evaluations. The next chapter will present results.

Chapter IV: Results

Introduction

As stated in Chapter I, teacher evaluation systems have been shown to treat teachers as if they are replaceable parts, coined the “widget effect” (Weisberg et al., 2009). This practice can lead to teachers taking part in systems and structures that do not provide valid, fair, and reliable evaluations that support their professional growth. Furthermore, there are statistical analyses that can be conducted to examine the extent of agreement among the evaluators of evaluation systems using rater mediated assessments. This type of assessment requires an expert rater, whether internal or external, to make ratings onto a rubric regarding the performance of a practitioner’s ability. Thus, leading researchers to understand the rater response process through the analysis of interrater reliability information within the teacher evaluation practice. Specifically, this study aims to understand the selection, certification, training, and support provided to evaluators employed by the EES Office. This chapter is organized with each of the four research questions being answered with relevant tables. The full data tables of each facet with corresponding statistics can be seen in the Appendices. The two Wright maps produced from the data analysis can be found in Appendix M & N. This chapter concludes with a brief summary of findings, which are explored further in Chapter V.

Research Question 1: What are the model estimated ability levels for the fictitious teacher profiles?

The first research question aimed to determine the model-estimated ability levels for the fictitious teacher profiles. This places the fictitious teacher ability estimates onto a logit scale, which allows researchers to know if the profiles display the appropriate range of variability that is reflected in the field. The fictitious teacher profile activity asked raters to make summative ratings on all 30 elements of the NCTEP rubric regarding the diverse teacher profiles. The EES Office resource manual describes that in the field, most early educators (70-80%) are Proficient, while a small percentage are Developing (0-10%), Accomplished (5-15%) or Distinguished (0-5%) (de-Kort Young et al., 2016). Therefore, a similar spread is expected in the fictitious teacher profiles.

First, based on visual inspection of the Wright map, there is a reasonable spread across the group of fictitious teacher profiles (Appendix M). The statistics show appropriate model-data fit. The results indicated that ability estimates can be assigned with a high degree of reliability. Each of the ability estimates for the fictitious teacher profiles had small standard errors (.04-.06). The person separation index (31.10) was high, and the person reliability measure was high (.99). There are two chi-square significance tests provided by the Facets output, which provide information about the teacher profile ability model (Linacre, 2020). The fixed effect chi-square was statistically significant ($\chi^2(9) = 8504.4$, $p < .05$) which indicates that the teacher profiles have differing ability levels. The random chi-square was not statistically significant ($\chi^2(8) = 9.0$, $p = .34$), which indicates the teacher profile ratings can be regarded as a random sample from a normally distributed population.

Each profile was analyzed to determine the model-estimated ability of the teachers, which allows researchers to understand the extent that the profiles reflect the teacher abilities in the field. This was done utilizing the threshold estimates from the model and is illustrated in the Wright map. Table 1 illustrates the individual teacher profile statistics. The column of teacher ability shows the spread across the logit scale. Teacher profile 5 (1.80 logits) was identified as a high-quality teacher, who should receive an Accomplished rating on the NCTEP rubric. The expert panel rated this fictitious profile as Accomplished across all five Standards. On the other hand, the results indicated that teacher profile 4 (-2.78 logits) is Developing and should receive a Developing rating on the NCTEP rubric. The expert panel rated this fictitious teacher profile as Developing across all five Standards. The results indicate two Developing fictitious teacher profiles (20%), six Proficient fictitious teacher profiles (60%), and two Accomplished fictitious teacher profiles (20%). This is similar to the spread that is expected across the field (de-Kort Young et al., 2016).

Finally, the infit and outfit mean square error (*MSE)* statistics were analyzed. Teacher profiles with infit *MSE* statistics that are lower than 0.7 indicate muted rating patterns, and teacher profiles with outfit *MSE* statistics that are higher than 1.3, indicate noisy rating patterns. None of the teacher profiles fall outside of the expected range of 0.7 to 1.3. Therefore, the answer to research question 1 shows that these teacher profiles are variable, and that the model-estimated ability levels reflect what is seen in the field.

Table 1

Calibration of Fictitious Teacher Profiles in the Strictness Calibration Model

Teacher	Observed	Teacher Ability		Mean Square Errors *(MSE)*	
Profile	avg rating	(logit)	*SE*	Infit	Outfit
1	2.07	0.17	0.04	0.90	0.95
2	1.78	-0.81	0.04	1.01	1.03
3	2.26	0.83	0.04	0.85	0.86
4	1.25	-2.78	0.06	1.17	1.19
5	2.54	1.80	0.05	1.30	1.34
6	2.21	0.81	0.04	0.95	0.95
7	1.38	-2.30	0.05	0.96	0.96
8	1.63	-1.17	0.05	0.97	0.97
9	2.16	0.63	0.04	0.97	0.98
10	2.35	1.28	0.04	0.91	0.90

Research Question 2: How variable are the raters in terms of strictness?

The second research question aimed to understand how variable the raters are in their strictness of ratings. The model calibrates the rater strictness onto the logit scale and appropriate model-data fit is illustrated. First, based on visual inspection of the Wright map, there was some spread in rater strictness across the group, with most raters (*n*=40) falling in the acceptable range of -0.5 logits to 0.5 logits. The MSE statistics of each rater is analyzed. Each of the ability estimates for rater strictness had small standard errors (.11). The person separation index (5.0) was high and indicates the raters could be divided into 5 groups based on the strictness. The person reliability measure was high (.96). The fixed effect chi-square was statistically significant ($\chi^2(56) = 1426.4$, $p < .05$), indicating that the raters do have evidence of variability in their ratings. The random chi-square was not statistically significant ($\chi^2(55) = 53.9$, $p = .52$), which indicates the rater strictness ratings can be regarded as a random sample from a normally distributed population.

The directionality of the rater severity measure on the logit scale indicates that raters 27 and 35 were the strictest of the group, with logit placements of 1.51 and 1.56. On the opposite end of the logit scale, 22 is the most lenient rater with a logit scale placement of -0.99. Nine total raters exhibited muted rating patterns. This includes rater 2 (Infit *MSE*= 0.66), 23 (Infit *MSE*=0.60), 39 (Infit *MSE*=0.40), 41 (Infit *MSE*=0.64), 46 and 47 (Infit *MSE*=0.67). Rater 49 (Infit *MSE*=0.57), 56 (Infit *MSE*=0.64) and 57 (Infit *MSE*=0.58) also exhibit muted rating patterns. These raters exhibit patterns that are more uniform than are predicted by the model.

On the other hand, raters with noisy rating patterns indicate unexpected category use and responses that are more random than the model expected. Noisy rating patterns have infit and outfit MSE statistics above 1.3. These raters are 3 (Infit *MSE*=2.26, Outfit *MSE*=2.83), 11 (Infit

MSE=1.64, Outfit *MSE*=1.87), 19 (Infit *MSE*=1.55, Outfit *MSE*=2.06), 20 (Infit *MSE*=1.94, Outfit *MSE*=1.96), 24 (Infit *MSE*=1.61, Outfit *MSE*=1.84), 36 (Infit *MSE*=2.87, Outfit *MSE*=3.34), 42 (Infit *MSE*=1.44, Outfit *MSE*=1.46), and 45 (Infit *MSE*=1.31, Outfit *MSE*=1.37). Table 2 provides full statistics on the unexpected rating patterns of the raters, specifically the columns of strictness logits and *MSE* statistics. In conclusion, the answer to research question 2 illustrates that this group of raters do vary in their strictness, thus providing evidence of no measurement invariance.

Table 2

Raters with Unexpected Rating Patterns in the Strictness Calibration Model

Raters	Observed	Strictness		Mean Square Errors *(MSE)*	
ID	avg rating	(logit)	*SE*	Infit	Outfit
2	2.12	-0.58	0.11	0.66	0.65
3	1.81	0.55	0.11	2.26	2.83
11	1.87	0.33	0.11	1.64	1.87
19	1.75	0.80	0.11	1.55	2.06
20	2.11	-0.54	0.11	1.94	1.96
23	1.90	0.21	0.11	0.60	0.67
24	2.10	-0.49	0.11	1.61	1.84
36	2.17	-0.76	0.11	2.87	3.34
39	1.79	0.64	0.11	0.40	0.39
41	1.68	1.06	0.12	0.64	0.55
42	1.98	-0.06	0.11	1.44	1.46
45	2.08	-0.43	0.11	1.31	1.37
46	1.96	0.01	0.11	0.67	0.66
47	1.85	0.39	0.11	0.67	0.68
49	1.87	0.33	0.11	0.57	0.56
56	2.14	-0.64	0.11	0.64	0.67
57	1.99	-0.10	0.11	0.58	0.65

Research Question 3: Is there evidence of differential rater functioning (DRF) by race of fictitious teacher profiles?

The third research question aimed to understand if there are any group differences across rater severity based on the race of the teacher in the fictitious teacher profile. The fixed effect chi-square was statistically significant ($\chi^2(1) = 29.0$, $p < .05$), indicating that the raters do have evidence of variability in their ratings based on the race or ethnicity of the teacher in the profile. The bias main effect logit is 0.08 for White teachers and -0.08 for Teachers of Color (TOC), which indicates an overall bias of 0.16 (see Table 3). This bias effect is very small in practice. Recall that bias statistics should be below 1.0 (Zwick, 1999). Therefore, this potential bias is minimal in its overall group effect. The areas of bias that are illustrated in this statistically significant finding can be explained when exploring individual rater patterns. Individual interaction terms are analyzed to determine areas of bias in individual rater strictness by race of profile. The Bias Test model provides individual rater bias statistics and t-statistics, which can be conceptualized as a z-score because they are reported with infinite degrees of freedom (Linacre, 2020). Due to the high stake's nature of this work, a smaller acceptable range of bias (-0.5 logits to 0.5 logits) is used, than the traditional range of 1.0.

Table 3

Bias Main Effect for Rater Strictness by Race of Teacher Profile

Race	Observed	Bias		Mean Square Errors *(MSE)*	
of Teacher	avg rating	(logit)	*SE*	Infit	Outfit
White	1.92	0.08	0.02	0.98	1.00
TOC	2.01	-0.08	0.02	1.00	1.03

Five raters have bias interaction terms outside of the acceptable range of agreement. Rater 3 (-0.52 bias) shows one area of bias, when providing ratings towards White teacher profiles and does not show overly strict ratings toward TOC profiles (0.47 bias). However, this bias term is very close to the accepted boundary. Raters 10, 11, 24, and 36 have evidence of bias towards both White teachers and TOC. For all four of these raters, the bias interaction term indicates a higher rating is given for White profiles and a lower rating is given for TOC profiles. In other words, these raters are rating White teachers with leniency and TOC with strictness. Additionally, all five of these raters have t-score statistics that are above the absolute value of

2.0, indicating unexpected differences in responses. Therefore, the answer to research question 3 illustrates that as a group, these raters do not show evidence of differential rater functioning. However, five raters do show evidence of needing further exploration into the rater behaviors (see Table 4).

Table 4

Raters with Unexpected Bias Terms and t-Scores in the Bias Test model

Rater		White Teacher Profiles						TOC Teacher Profiles		
ID	Obs.	Exp.	Bias	*SE*	t-score	Obs.	Exp.	Bias.	*SE*	t-score
3	246	265	-0.52	0.17	-3.12	297	278	0.47	0.15	3.03
10	291	269	0.54	0.16	3.46	260	282	-0.54	0.16	-3.38
11	301	274	0.66	0.15	4.24	260	287	-0.67	0.16	-4.16
24	351	308	1.07	0.16	6.56	278	321	-1.04	0.16	-6.62
36	381	319	1.68	0.18	9.27	270	332	-1.50	0.16	-9.49

Wright Map for Strictness Calibration model

As stated in Chapter III, it is important to use both statistics and a visual inspection of a Wright map to illustrate the full extent of interrater reliability and rater strictness. This Wright map displays the statistical information for the facets of fictitious teacher profile ability, rater strictness, racial bias main effect, item difficulty, and NCTEP rubric category thresholds (Appendix M). When interpreting the Wright map, the far-left column displays the logit scale, which ranges from 2.5 logits to -3.0 logits. Then, each column displays a different facet onto this logit scale. Upon visual inspection, the fictitious teacher profile ability estimates, and rater strictness estimates have a spread across the logit scale. The racial bias main effect and the item difficulty facet show minimal variability because they are located between the boundary of -0.5 logits to 0.5 logits. Then, the last column displays the NCTEP rubric category thresholds. It is possible to see which overall rating of teacher quality the fictitious teacher profiles have by taking an average of each of the 30 items' individual category threshold statistics. This provides a general guideline for where the teacher ability will fall on the NCTEP rubric.

The fictitious teacher profiles are displayed along their logit score location, these range from 1.80 logits to -2.78 logits. Each of these fictitious teacher profiles indicates which rating of teacher quality is exhibited by that profile. The fictitious teacher profiles that are of the highest ability are 5 and 10, and the lowest ability profiles are 7 and 4. The NCTEP rubric category thresholds displays the Developing threshold below -1.72 logits, the Proficient threshold between -1.72 logits to 1.18 logits, and the Accomplished threshold is above 1.18 logits.

The three middle columns display rater strictness, racial bias main effect, and item difficulty. These columns take the information and place their logit scores across the scale to indicate how facets function across each other, as well as within each other. Rater strictness estimates showed that the most strict raters are 27 and 35, and the most lenient is 22. However, most of the raters (*n*=40) fall within the acceptable range.

The racial bias main effect shows a minimal effect across the group, as discussed in the section on research question 3. Finally, the item difficulty placements show that 28 items load in the range of -0.5 logits to 0.5 logits, with two items slightly outside this range. It is expected that some items would be difficult, average, and easy to endorse in the field. The two items that are more difficult to endorse are Standard 2A and Standard 2E. Standard 2A states "teachers provide an environment in which each child has a positive, nurturing relationship with caring adults" and Standard 2E states "Teachers work collaboratively with the families and significant adults in the lives of their students". This demonstrates that these standards may be the most difficult to identify and endorse in the field.

Research Question 4: **How do the ratings of fictitious teacher profiles compare with the summative ratings of their caseloads in the field?**

The fourth research question aimed to understand how the rater behavior exhibited on the fictitious teacher profile activity is, or is not, displayed in the field. The answer to this question can be explored using the two measurement models and comparing various indicators from the Strictness Calibration model to the rater patterns of strictness or leniency displayed in the Summative Caseloads model. The measurement model statistics, including chi-square and *MSE*, are described first. Then, a description of the Summative Caseloads model Wright map. Finally, the two models are compared using the boundary of the absolute value of 0.5 among individual statistics.

In the Summative Caseloads model, appropriate model-data fit is found. Each of the ability estimates for the fictitious teacher profiles had small standard errors (.05-.06). The person separation index (4.0) was high, and the person reliability measure was high (.94). To determine if the fictitious teacher profiles have the same ability level, chi-square statistics were analyzed. The fixed effect chi-square was statistically significant ($\chi^2(561) = 23688.3$, $p < .05$) which indicates that the teacher profiles have differing ability levels. The random chi-square was not statistically significant ($\chi^2(560) = 496.6$, $p = .97$), which indicates the teacher profile ratings can be regarded as a random sample from a normally distributed population.

Additionally, teacher profiles 4, 5, 7, and 8 had logit scale locations that are more than ±0.5 difference from the Strictness Calibration model (see Table 5). This indicates that the teacher quality exhibited in the fictitious teacher profile is even more Developing than was exhibited in the Strictness Calibration model. Finally, the infit and outfit *MSE* are analyzed and none of the teacher profiles fall outside of the expected range of 0.7 to 1.3 (see Table 6). Teacher profiles 2, 5, and 9 had infit and outfit *MSE* slightly over the acceptable range of 1.3. However, none of these teacher profiles have indicators outside the acceptable range, meaning they are well-constructed.

Table 5

Comparison of Logit Locations for Teacher Profiles
Across the Strictness Calibration Model and Summative Caseloads Model

Teacher Profiles ID	Strictness Calibration (logit)	Summative Caseloads (logit)	Difference (logit)	Is the difference >|0.5|?	Did the interpretation of ability change?
1	0.17	0.38	0.21	No	No
2	-0.81	-1.08	-0.27	No	No
3	0.83	1.31	0.48	No	No
4	-2.78	-3.59	-0.81	**Yes**	No
5	1.80	2.50	0.70	**Yes**	No
6	0.81	1.07	0.26	No	No
7	-2.30	-2.90	-0.60	**Yes**	No
8	-1.17	-1.76	-0.59	**Yes**	No
9	0.63	0.81	0.18	No	No
10	1.28	1.68	0.40	No	No

Table 6

Comparison of Infit and Outfit MSE Statistics for Teacher Profiles
Across Strictness Calibration Model and Summative Caseloads Model

Teacher Profile ID	Strictness Calibration infit	Summative Caseloads infit	Difference	Strictness Calibration outfit	Summative Caseloads outfit	Difference
1	0.90	1.32	0.42	0.95	1.35	0.40
2	1.01	1.43	0.42	1.03	1.49	0.46
3	0.85	1.14	0.29	0.86	1.16	0.30
4	1.17	1.27	0.10	1.19	1.37	0.18
5	1.30	1.54	0.24	1.34	1.61	0.27
6	0.95	1.31	0.36	0.95	1.28	0.33
7	0.96	1.09	0.13	0.96	1.09	0.13
8	0.97	1.22	0.25	0.97	1.26	0.29
9	0.97	1.42	0.45	0.98	1.44	0.46
10	0.91	1.15	0.24	0.90	1.15	0.25

Next, the facet of rater strictness is analyzed, and appropriate model-data fit is found. The fixed effect chi-square was statistically significant ($\chi^2(56) = 3919.7$, $p < .05$), indicating that the raters do have evidence of variability in their ratings. The random chi-square was not statistically significant ($\chi^2(55) = 55.1$, $p = .47$), which indicates the rater strictness ratings can be regarded as a random sample from a normally distributed population. Each of the ability estimates for rater strictness had small standard errors (.07-.11). The person separation index (7.77) indicated the raters could be divided into 7 groups based on the strictness. The person reliability measure was high (.98).

The directionality of the rater severity measure on the logit scale indicated that raters 27 and 35 were still the most strict of the group, with logit values of 2.05 and 2.10. On the opposite end of the logit scale, 22 was the most lenient rater with a logit scale placement of -1.25. Raters with infit *MSE* statistics that are lower than 0.7 indicate muted rating patterns, and raters with outfit *MSE* statistics that are higher than 1.3, indicate noisy rating patterns. These raters are 3 (Infit *MSE*=1.72, Outfit *MSE*=1.95), 11 (Infit *MSE*=1.44, Outfit *MSE*=1.54), 19 (Infit *MSE*=1.78, Outfit *MSE*=2.31), 20 (Infit *MSE*=1.59, Outfit *MSE*=1.62), 24 (Infit *MSE*=1.54, Outfit *MSE*=1.62), 36 (Infit *MSE*=3.50, Outfit *MSE*=3.95), 42 (Infit *MSE*=1.68, Outfit *MSE*=1.68), and 45 (Infit *MSE*=1.29, Outfit *MSE*=1.33). Table 7 provides full statistics on the unexpected rating patterns of the raters. When making comparisons across the group of raters' infit and outfit *MSE* statistics, it is evident that two raters have a range outside of the acceptable difference. These are rater 3 and rater 36. The full tables illustrating all rater statistics can be found in the appendix. The patterns exhibited by Rater 3 and Rater 36 illustrate that the patterns of strictness and leniency exhibited when rating the fictitious teacher profiles in the interrater reliability activity are even more so exhibited in the field.

Table 7

Raters with Unexpected Rating Patterns in Summative Caseloads Model

Raters	Observed	Strictness		Mean Square Errors *(MSE)*	
ID	avg rating	(logit)	*SE*	Infit	Outfit
3	1.83	0.72	0.08	1.72	1.95
11	2.07	0.41	0.09	1.44	1.54
19	1.89	1.05	0.12	1.78	2.31
20	2.19	-0.70	0.08	1.59	1.62
24	2.08	-0.64	0.10	1.54	1.62
36	2.17	-0.96	0.12	3.50	3.95
42	2.00	-0.09	0.11	1.68	1.68
45	2.24	-0.56	0.09	1.29	1.33

There is an additional facet of item difficulty that can be analyzed for model comparisons and appropriate model-data fit is exhibited. The fixed effect chi-square was statistically significant ($\chi^2(29) = 696.9$, $p < .05$), indicating that the raters do have evidence of variability in their ratings. The random chi-square was not statistically significant ($\chi^2(28) = 27.8$, $p = .47$), which indicates the item difficulty ratings can be regarded as a random sample from a normally distributed population. Each of the item difficulty estimates had small standard errors (.07-.10). The person separation index (4.74) indicated the items could be divided into 4 or 5 groups based on the difficulty. The person reliability measure was high (.96). The logit locations of 25 items were within appropriate range. This illustrates the consistency within the NCTEP rubric.

Wright Map for Summative Caseloads Model

The next step in the results is to analyze the Wright map for the Summative Caseloads model (Appendix N). This Wright map can be interpreted similarly to the Strictness Calibration model. The racial bias main effect is not present within this model, because the researcher did not have the racial identities of all early childhood educators served in the field. Additionally, due to the large sample of teachers in this model (*n*=562), only the fictitious teacher profiles are displayed. When interpreting the Wright map, the far-left column displays the logit scale, which ranges from 2.5 logits to -3.5 logits. Then, each column displays a different facet onto this logit scale. Upon visual inspection, the teacher ability estimates, rater strictness estimates and item difficulty information display a spread across the logit scale, with some individual raters exhibiting estimates beyond the boundary of -0.5 logits to 0.5 logits.

The fictitious teacher profiles are displayed along their logit score location, these range from 2.50 logits to -3.59 logits. The teacher profiles that are of the highest teacher quality are 5 and 10, with the lowest quality being 7 and 4. It is possible to see which overall teacher quality rating the fictitious teacher profiles have by taking an average of each of the 30 items' individual item category threshold statistics. This provides a general guideline for where the teacher ability will fall on the NCTEP rubric. The final column of the Wright map illustrates these category thresholds. The Developing threshold is below -2.20 logits, the Proficient threshold is between -2.20 logits to 2.10 logits, and the Accomplished is above 2.10 logits.

The middle two columns display rater strictness and item difficulty. These columns take the information and place their logit scores across the scale to indicate how facets are functioning across each other, as well as within each other. Rater strictness shows the most strict raters are 27 and 35, and the most lenient is 22. However, fewer raters (*n*=32) fall within the acceptable range than in the Strictness Calibration model. Finally, the item difficulty placements show that 25 items load in the acceptable range of -0.5 logits to 0.5 logits, with five items outside the acceptable range. Standard 4H states "Teachers use a variety of methods to assess what each student has learned" above 0.5 logits. Standard 2C states "Teachers treat students as individuals", and Standard 2 is one of the overarching standards which states, "Teachers establish a respectful environment for a diverse population of students". Both of these items fell below -0.5 logits. Standard 2E and Standard 2A maintained their location of below -0.5 logits.

Comparing Rater Strictness Across the Two Models

The goal of research question 4 is to understand the extent to which the rater strictness exhibited by the raters in the fictitious profiles is exhibited in the real, summative caseloads in the field. The process of drawing comparisons across the facet of rater strictness between the Strictness Calibration model and Summative Caseloads model requires a visual inspection of both Wright maps. Additionally, an exploration for any individual raters with infit and outfit *MSE* statistics that changed by more than the absolute value of 0.5 logits from one model to the other. Across these statistics, the researcher can determine if any of the overall strictness thresholds (strict, acceptable, lenient) have changed.

When comparing the logit placements of both models, rater strictness logit estimates showed two raters (27 and 35) with a difference of more than 0.5 logit difference between the individual *MSE* statistics. These two raters both have acceptable *MSE* statistic ranges (0.6-1.3), and both show logit ratings higher than the Strictness Calibration model. Therefore, these raters may be more strict in their evaluation practice in the field than they exhibited in the activity alone. This finding is critically important for EES Office leadership because this strictness can negatively affect early educators. However, the overall strictness category did not change because these same raters exhibited rater patterns of strictness in the Strictness Calibration model. Five total raters had logit placements that changed their overall strictness category. Therefore, these raters may exhibit rater patterns of strictness or leniency in the field that was not demonstrated in the fictitious teacher profile interrater reliability activity alone. These raters are seen in table 8.

Table 8

Overall Changes in Logit Locations for Rater Strictness Across Strictness Calibration Model and Summative Caseloads Model

Rater ID	Strictness Calibration (logit)	Summative Caseloads (logit)	Difference (logit)	Is the difference >|0.5|?	Did the interpretation of strictness change?
3	0.55	0.72	0.17	No	**Yes**
6	-0.44	-0.58	-0.14	No	**Yes**
10	0.45	0.58	0.13	No	**Yes**
14	0.47	0.60	0.13	No	**Yes**
20	-0.54	-0.70	-0.16	No	**Yes**
24	-0.49	-0.64	-0.15	No	**Yes**
27	1.51	2.05	0.54	**Yes**	No
35	1.56	2.10	0.54	**Yes**	No
50	-0.47	-0.61	-0.14	No	**Yes**
55	-0.55	-0.71	-0.16	No	**Yes**

Raters can be compared across the various models and those that continually exhibited unexpected rating patterns can be further investigated. For example, when comparing the *MSE* statistics of both models, it is evident that Rater 3 shows statistics that are beyond the acceptable difference range (Appendix J). However, the statistics also show that this rater is more consistent with their ratings when the entire caseload is added to the model because of the lower *MSE* statistics. On the other hand, Rater 36 exhibits not only *MSE* statistics that are beyond the acceptable difference range, but also higher *MSE* statistics. This indicates that this rater is even more inconsistent with their ratings in the field than they are in the interrater reliability activity. However, none of the raters changed their overall strictness category.

Conclusion

In this chapter, the results were displayed for the four research questions. The results indicate that the raters employed by the EES Office exhibit some spread in rater strictness across the group and do exhibit similar patterns of strictness or leniency in the field. This indicates a lack of measurement invariance among these data. When comparing the Strictness Calibration and Summative Caseloads models, the Wright maps display information regarding individual rater behavior. The teacher profile ability estimates displayed in all three models show that the interrater reliability activity does provide useful information. This information allows for the leadership team to monitor raters employed by EES Office. The data shows areas in which individual raters differ from the overall patterns exhibited by the group. These individual raters who display statistics outside the expected boundary will be provided individual plans to support their growth. Finally, the comparison across the two models illustrate that the interrater reliability activity process used by the EES Office has high reliability. Therefore, the assessment system that has been developed and described in this study could be replicated to become a national or statewide model for teacher evaluation efforts.

Chapter V: Conclusions

Introduction

Teacher licensure evaluation systems are present in every public school district across the country. However, these vary from state to state, and even with state department guidance, these evaluation efforts can vary from district to district or even, school to school. Inconsistent evaluations lead to teachers that do not receive valid, fair, and reliable information regarding their own teaching practice. Weisberg et al. (2009) suggest that evaluation systems share four characteristics. First, these evaluation systems are able to differentiate between teachers by their effectiveness. Second, the evaluators conducting teacher evaluations are highly trained and held accountable for the reliability of their work. Third, evaluation systems have clear consequences for teachers who are not able to reach the minimum qualification of teacher quality. Fourth, evaluation systems provide a conversation regarding professional skills and instructional practices to improve the teacher's practice. Each of these elements is present in the evaluation system utilized by the EES Office in North Carolina. The purpose of this research was to understand the rater response process of the early childhood teacher performance evaluators employed by the EES Office.

The previous four chapters have laid the foundation for the results and conclusions presented in this chapter. This work began with a description of the problem within teacher evaluation practices, including the practice of treating teachers' as "interchangeable widgets" (Weisberg et al., 2009). Then, a review of the literature about PE standards (Gullickson & Howard, 2009) and EPT standards (AERA et al., 2014) in the field, as well as a description of Rater Mediated Assessment Theory. Chapter II concluded with a review of studies that explored interrater reliability in teacher evaluation efforts. Next, the methodology and analysis chapters described the utilization of MFRMs to understand the spread of fictitious teacher profile quality, rater strictness or leniency and potential racial bias. Finally, the Strictness Calibration model and the Summative Caseloads model are compared to illustrate how the behavior exhibited by the raters in the field compares to the rater behavior exhibited in authentic, summative evaluations from their real caseloads. This chapter will provide the implications for the findings of each research question, as well as recommendations for future research.

Research Question 1 Findings

In the EES Office evaluation results, it is expected that a spread of teacher quality is displayed. It is reported that most educators should fall in the Proficient range (70-80%), with Developing, Accomplished and Distinguished educators being less than 20% of the entire group (de-Kort Young et al., 2006). Therefore, it is expected that seven or eight of the profiles would be rated at Proficient. The results of this analysis showed that six profiles received a Proficient rating, and two profiles each received a Developing or Accomplished rating. This indicates that the teacher profiles do exhibit a similar range of quality as is seen in the field. Additionally, these results align with the phased quality improvement model research because the leadership team wanted to create fictitious teacher profiles that reflected actual practices in the field (Lambert, Moore et al., 2021). Furthermore, these profiles are well-crafted and do model a spread of teacher quality that would be expected in the field.

Research Question 2 Findings

The second research question aims to understand the strictness and leniency exhibited by the raters across all the teacher profiles, as well as the extent to which invariance exists across this group of raters. If the property of invariance is upheld, it would indicate that fictitious teacher profiles received a valid, fair, and reliable rating, regardless of the rater who rated them.

Furthermore, the logit placements showed that several raters (*n*=17) have strictness or leniency logit placements that are outside of the acceptable logit range. This means the property of invariance is not present for these raters with this data. This may indicate some teacher profiles could be unfairly rated due to the strictness or leniency exhibited by some individual raters. This finding indicates that some of these raters could be assigning unfair ratings in their summative caseloads with actual early educators. The lack of invariance within rater strictness illustrates the need for further investigation into the individual rater response patterns. Therefore, the fourth research question, which compares the Strictness Calibration model with the Summative Caseloads model, is explored to determine the extent this behavior is exhibited in the field. Additionally, the raters who consistently exhibit unexpected rater patterns have individualized support plans that include co-observations with the project coordinator and specific professional development goals. The project coordinator serves on the leadership team of EES Office and is the only researcher with the cross walk that can reveal the identity of the individual evaluator. Furthermore, she is responsible for ensuring the raters who exhibit unexpected rater patterns are held accountable for those actions in the field.

Research Question 3 Findings

The third research question aims to understand if any group differences exist among the rater strictness rating and the race of the fictitious teacher profiles. The bias main effect was found to be statistically significant, indicating the potential for bias. However, this bias is so minimal (0.16), that it is reasonable to conclude that across the whole group of EES Office raters, the potential bias is not impacting the overall teacher profile ability estimates. Instead, the individual bias interaction terms indicate that a small number of raters (*n*=5) have patterns indicating potential bias.

Research indicates the importance of bias management within personnel evaluation (Gullickson & Howard, 2009). Therefore, these five raters who are identified as having potential bias will have personalized support. Each evaluator had a meeting with the project coordinator for the EES Office. In these meetings, the evaluator and project coordinator worked together to determine several relevant professional goals for the rest of the academic year. Then, some of the classroom observations that evaluator conducts as part of their normal caseload will include the project coordinator, so they can have discussions regarding teaching practice.

Research Question 4 Findings

The fourth research question aims to compare the model data-fit information and individual facet statistics from the Strictness Calibration model and the Summative Caseloads model. This includes exploring the extent to which invariance exists across the group of raters in the field by investigating the patterns of individual raters. The comparisons of these two models allow for researchers to determine if the same behavior exhibited in the fictitious teacher profiles with 10 teachers, is exhibited in the caseloads of real early educators. For the 2020-2021 academic year, this was 552 educators. This research question is answered using a combination of the Wright maps, logit estimates, and MSE statistics for each of the facet estimates. This includes the teacher profile facet, rater strictness facet, and item difficulty facet. Each of these facets was addressed briefly in Chapter IV. Furthermore, misfit is best identified by researchers who have a deep understanding of the latent construct, which in this case is teacher quality (Engelhard & Wind, 2018).

The teacher profile facet illustrates a minimal change in logit location from one model to the other. However, the change does not affect the overall teacher ability. For example, teacher profile 4 changed in their logit score by -0.81. This was over the absolute value of 0.5, but it did not change the overall interpretation of the teaching quality classification. This profile was low quality in both models (below -0.5 logits), and it did it change the overall NCTEP rubric rating (Developing). The fictitious teacher profiles *MSE* statistics are within the appropriate range, which further indicates how well-crafted these profiles are because they accurately depict the early educator behavior seen in the field. This provides evidence that these fictitious teacher profiles do display the appropriate amount of variability and are a good representation of the early childhood educators served by the EES Office.

The rater strictness facet illustrates that the logit locations change from one model to the other. For eight raters, this change in logit location changed their overall strictness category and illustrates that patterns in their rater behavior are even more present in the field. This indicates that the strictness and leniency seen in the Strictness Calibration may be a conservative estimate of strictness. Instead, raters are even more strict or lenient than was originally displayed in the field. Raters 3, 10 & 24 have consistently shown unexpected rating patterns across all MFRMs. Rater 3 shows a slight bias (-0.52) against White teachers and no evidence of bias towards TOC. This means this rater awards White teachers a higher score on the rubric, by 0.5 of a category rating. In practice, this is minimal effect because of the nature of category thresholds of the NCTEP rubric. However, raters 10 and 24 show the opposite pattern of behavior. Rater 10 exhibits a bias of 0.54 towards White teachers, and a bias of -0.54 against TOC. Rater 24 exhibits a bias of 1.07, indicating White teachers may be given a rating as large as one full category rating above TOC of the same teaching quality. In practice, this provides evidence that an early childhood educator in the field may be awarded a lower teacher quality rating based on their race by this rater. Thus, it is imperative that these raters participate in the individualized plan described in the previous section with the project coordinator to ensure this rater behavior is not exhibited in the field.

The item difficulty facet illustrates a minimal change in logit location from one model to the next. The facet of item difficulty was not the focus of this research, partially because the NCTEP rubric has been used for close to 15 years to evaluate teaching quality and is a high-quality instrument. However, some of the items are more difficult to endorse, or more difficult to make placements than others. An example of this is Standard 4H, which illustrates that teachers use a variety of assessments with students. This may be difficult to endorse on a rubric for a fictious teacher profile, without the full experience of meeting the teacher, being in the classroom while they are taking part in assessment procedures and learning the academic successes of the student population.

Recommendations

In NC, there is a long-standing belief in the importance of early childhood education. In the 2022-2023 academic year, the EES Office will celebrate its fifteenth year of serving early educators. This interrater reliability study is the last stage in a phased quality improvement model. This phased model began with developing initial training, creating a resource manual, and developing a conceptual framework. Then, the EES Office collected feedback from early educators and analyzed those results. The last step in this process included the creation of these fictitious teacher profiles and conducting the detailed investigation into interrater reliability (Lambert, Moore et al., 2021).

Early Educator Support Office

On a fundamental level, the work has meaning for the EES Office leadership staff. The Quality Assurance leadership team can use these findings to implement evaluator supports based on the statistical analysis. This includes having individual meetings with the project coordinator and having supervised classroom observations. These findings also indicate that these profiles are well-crafted and are a good indicator of evaluator strictness or leniency in the field. These profiles should be used in further interrater reliability activities. Additionally, if more profiles are crafted to add to the overall pool of profiles to provide sampling techniques, an investigation of their reliability should be explored.

These results further indicate the professional nature of the evaluation work conducted within the EES Office. It indicates that most of the evaluators on the team do provide valid, fair, and reliable results to the early educators that are served in the field. For the over 900 early educators that are served, these results indicate that they have professional evaluators working with them to improve their teaching practice. In fact, these early educators may be receiving more information about their teaching ability than their public-school counterparts. Other Birth through Kindergarten licensed educators may only receive summative and formative evaluations by an administrator, which are solely for licensing. They may not receive the professional development supports that are provided by EES Office.

The EES Office utilizes a research-based approach when it comes to providing evaluations. It is critical that those serving as evaluators are recertified every couple of years to further ensure reliability. It is important to note that while these rater response patterns are exhibited across more than one model, there could still be hidden patterns of strictness, leniency, or bias within real evaluator practice. Therefore, it is key that interrater reliability is conducted on a routine basis by the EES Office leadership team.

Future Research

While these findings impact the EES Office leadership staff, evaluators, and early educators that are served, these findings also implicate a state or national model. There are standards in place for personnel evaluation, and it is suggested that evaluation efforts be seen as a combination of various elements to ensure accuracy (Gullickson & Howard, 2009). Teacher evaluation assessment systems are encouraged to have reliable practices that treat all teachers with fairness and validity. This study could be used as a model for other teacher evaluation assessment systems, beyond just early childhood educators.

As illustrated by the phased quality improvement model, each of the steps has been strategic to provide fair and valid services to early educators. Teacher evaluations are high stakes because student outcomes are high stakes. Research has shown that teachers improve when taking part in capacity building or organization support structures (Stoll et al., 2006). Additionally, high-quality teachers improve student learning outcomes (Vescio et al., 2008). By implementing the tenets of the interrater reliability process that is utilized by the EES Office, teachers receive the support they need, which may lead to child success (Opper, 2019).

For other organizations or state leadership departments, the first steps to creating such a system of interrater reliability begins with a candid conversation regarding the goals of evaluation training. This includes discussions on how to best meet the needs of the teachers served and what benefits arise from the evaluation training. Additionally, it is important to consider any resources that are given to evaluators to support the reliability of the practice. The EES Office has a resource manual (de Kort Young et al., 2016) which provides detailed information regarding the evaluation process, the NCTEP rubric and applications in the field. In conclusion, there are many important steps to create an evaluation system, like that seen in the EES Office.

Summary

As illustrated in Chapter II, there is no research to date that has been conducted regarding the level of consistency in rater behavior and agreement of evaluators who serve B-K licensed educators in the state of NC. This research provides an exploration of the interrater reliability of a set of external evaluators on an activity of ten fictitious teacher profiles. These profiles displayed a variety of evidence, including photographs, videos, written descriptions, and student artifacts. These profiles were evaluated by the EES Office leadership team, and a consensus was reached on their overall abilities of the fictitious teacher profiles. Then, these fictitious profiles were used to explore the rater response process and provide individualized supports to evaluators. In conclusion, the EES Office evaluators have created an assessment system that provides valid, fair and reliable evaluation services to their early educators each year.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (Eds.). (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Andrich, D. (1978) A rating formulation for order response categories. *Psychometrika 43*, 561-573.

Andrich, D., & Marais, I. (2019). *A course in Rasch measurement theory: Measuring in the educational, social and health sciences*. Springer.

Bond, T. G., Yan, Z., & Heene, M. (2021). *Applying the Rasch model fundamental measurement in the human sciences* (4th ed.). Routledge Taylor & Francis Group.

Bottoms, B. L., Holcomb, T. S., Vestal, A. R., & Lambert, R. G. (2021). The development of a systematic approach to evaluating early childhood teachers using the North Carolina teacher evaluation process. *Center for Educational Measurement and Evaluation at the University of North Carolina at Charlotte.* https://ceme.charlotte.edu/sites/ceme.charlotte.edu/files/media/Tech%20Report-%20Dec%2021.pdf

Chi, O. L. (2021). *A classroom observer like me: The effects of demographic congruence between teachers and raters on observation scores* [Working paper]. Wheelock Educational Policy Center, Boston University. https://wheelockpolicycenter.org/wp-content/uploads/2021/02/ClassroomObserver_WP2021-1_FINAL.pdf

de Kort-Young, A.M, Lambert, R.G., Rowland, B., Vestal, A., & Ward, J. (2016). Resource manual for administrators and principals supervising and evaluating teachers of young children. *Center for Educational Measurement and Evaluation at the University of North*

Carolina at Charlotte. https://earlyeducatorsupport.uncc.edu/mentors-evaluators/evaluation-resource-manual

Eckes, T. (2011). *Introduction to many-facet rasch measurement: Analyzing and evaluating rater-mediated assessments.* Peter Lang GmbH, Internationaler Verlag der Wissenschaften.

Edwards, A. S., Edwards, K. E., & Wesolowski, B. C. (2019). The psychometric evaluation of a wind band performance rubric using the Multifaceted Rasch Partial Credit Measurement Model. *Research Studies in Music Education*. DOI: 10.1177/1321103X18773103

Engelhard, G. (2013). *Invariant measurement: Using Rasch models in the social, behavioral, and health sciences*. New York: Routledge.

Engelhard, G., & Wind, S. A. (2018). *Invariant measurement with raters and rating scales: Rasch models for rater-mediated assessments.* Routledge.

Engelhard, G. & Wind, S. A. (2019) Introduction to the special issue on rater-mediated assessments. *Journal of Educational Measurement.* 56(3). 475-477.

Graham, M., Milanowski, A., & Miller, J. (2012). *Measuring and promoting inter-rater agreement of teacher and principal performance ratings*. Washington D.C.: Center for Educator Compensation Reform, U.S. Department of Education.

Grissom, J. A., & Bartanen, B. (2021). Potential race and gender biases in high-stakes teacher observations. *Journal of Policy Analysis and Management*. 1-31.

Guilford, J. P. (1936). *Psychometric methods*. McGraw-Hill.

Gullickson, A. R., & Howard, B.B. (2009). *The Personnel Evaluation Standards: How to assess systems for evaluating educators (2nd ed.).* Thousand Oaks, CA: Corwin Press.

Guo, W & Wind, S. A. (2021). Examining the impacts of ignoring rater effects in mixed-format tests. *Journal of Educational Measurement. 58*(3). 364-387.

Gwet, K. (2012). *Handbook of interrater reliability: The definitive guide to measuring the extent of agreement among raters* (3rd edition). Advanced Analytics, LLC.

Herlihy, C., Karger, E., Pollard, C., Hill, H. C., Kraft, M. A., Williams, M., & Howard, S. (2014). State and local efforts to investigate the validity and reliability of scores from teacher evaluation systems. *Teachers College Record*, *116*(1), 1-28.

Ho, A.D., & Kane, T.J. (2013). The Reliability of Classroom Observations by School Personnel. Research Paper. MET Project.

Holcomb, T.S, Lambert, R.G., & Bottoms, B.L., (2022) Reliability Evidence for the NC Teacher Evaluation Process Using a Variety of Indicators of Inter-Rater Agreement. *Journal of Educational Supervision, 5*(1), 27-43, https://doi.org/10.31045/jes.5.1.2

Hoyt, W. T. (2000). Rater bias in psychological research: When is it a problem and what can we do about it? *Psychological methods*, *5*(1), 64.

Jimenez, A. M., & Zepeda, S. J. (2020). A comparison of Gwet's AC1 and Kappa when calculating inter-rater reliability coefficients in a teacher evaluation context. *Journal of Education Human Resources*, *38*(3), 290-300. https://doi.org/10.3138/jehr-2019-0001

Lambert, R.G., Holcomb, T. S., & Bottoms, B. L. (2021a). *Examining inter-rater reliability of evaluators judging teacher performance: Proposing an alternative to Cohen's Kappa.* Center for Educational Measurement and Evaluation, University of North Carolina at Charlotte. https://ceme.charlotte.edu/sites/ceme.charlotte.edu/files/media/IRR%20lambda%20paper%2006-2021.pdf

Lambert, R. G., Bottoms, B. L., & Holcomb, T. S. (2021). Inter-rater reliability of evaluators using the North Carolina Teacher Evaluation Process. Symposium presented at the 2021 North Carolina Association for Research in Education Conference (Virtual).

Lambert, R. G., Moore, C. M, Bottoms, B. L, Vestal, A., & Taylor, H. (2021). Use of Rasch modeling and focus group interviewing to inform the training of teacher evaluators. *International Journal of Multiple Research Approaches, 13*(2), 1-15.

Landis, J. R., & Koch, G. G. (1977) The measurement of observer agreement for categorical data. *Biometreics, 33,* 159-174. https://doi.org/10.2307/2529310

Linacre, J. M. (1994) Many Facet Rasch Measurement. MESA Press, Chicago, IL.

Linacre, J. M. (2020) Facets computer program for many-facet Rasch measurement, version 3.83.4. Beaverton, Oregon: Winsteps.com

Masters, G. N., (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149-174.

Mazurek, S. A. (2012). *Interrater Reliability among Elementary Principals Using the North Carolina Teacher Evaluation Process.* Available from Social Science Premium Collection. (1697499260; ED551396). https://www.proquest.com/dissertations-theses/interrater-reliability-among-elementary/docview/1697499260/se-2

North Carolina State Board of Education, North Carolina Teacher Evaluation Process (2021). Retrieved from https://sites.google.com/dpi.nc.gov/ncees-information-and-resource/home

North Carolina Division of Child Development and Early Education (NCDHHS) (n.d.). *Early educator support, licensure and professional development unit.* https://ncchildcare.ncdhhs.gov/Services/EESLPD

Olsen, L. W. (2003). *Essays on George Rasch and his Contributions to Statistics*. Unpublished PhD Thesis, University of Copenhagen.

Praetorius, A.K., Lenske, G., & Helmke, A. (2012). Observer ratings of instructional quality: Do they fulfill what they promise? *Learning and Instruction*, *22*(6), 387-400.

Putnick, D. L., & Bornstein, M. H. (2016). Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental Review, 41*, 71-90.

Rasch, G. (1960). Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests.

Stemler, S. E., & Tsai, J. (2008). Best practices in interrater reliability: Three common approaches. In J. W. Osborne (Ed.), *Best practices in quantitative methods* (pp. 29-49). Sage.

Stoll, L., Bolam, R., McMahon, A., Wallace, M., & Thomas, S. (2006). Professional learning communities: A review of the literature. *Journal of educational change*, *7*(4), 221-258.

Taylor, H., Vestal, A., Saperstein, D., Stafford, C., & Lambert, R. G. (2019). The early educator support, licensure, and professional development (EESLPD) office conceptual framework: A narrative describing supporting early childhood educators as part of the North Carolina teacher evaluation process. *CEME Technical Report*. https://ceme.charlotte.edu/sites/ceme.charlotte.edu/files/media/EESLPD%20Conceptual%20Framework%20Report%202019.pdf

Vesico, V., Ross, D., & Adams, A. (2008). A review of research on the impact of professional learning communities on teaching practice and student learning. *Teaching and Teacher Education 24*(1), 80-91.

Wang, J. & Engelhard, G. (2019). Conceptualizing rater judgments and rating processes for rater mediated assessments. *Journal of Educational Measurement 56*(3). 582-609. DOI: 10.1111/jedm.12226

Weisberg, D. Sexton, S., Mulhern, J. & Keeling, D. (2009). The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness. *Education Digest: Essential Readings Condensed for Quick Review*

Wesolowski, B. C., Wind, S. A. & Engelhard, G. (2015) Rater fairness in music performance assessment: Evaluating model-data fit and differential item functioning. *Musicae Scientiae 19*(2). 147-170. DOI: 10.1177/102986-4915589014

Wind, S. A., & Jones, E. (2019). Not just generalizability: A case for multifaceted latent trait models in teacher observation systems. *Educational Researcher*, *48*(8), 521-533.

Wright, B. & Masters, G. (1982). Rating scale analysis. Chicago, IL: MESA Press.

Zepeda, S. J, & Jimenez, A. M. (2019). Teacher evaluation and reliability: Additional insights gathered from inter-rater reliability analyses. *Journal of Educational Supervision, 2*(2), 11-26.

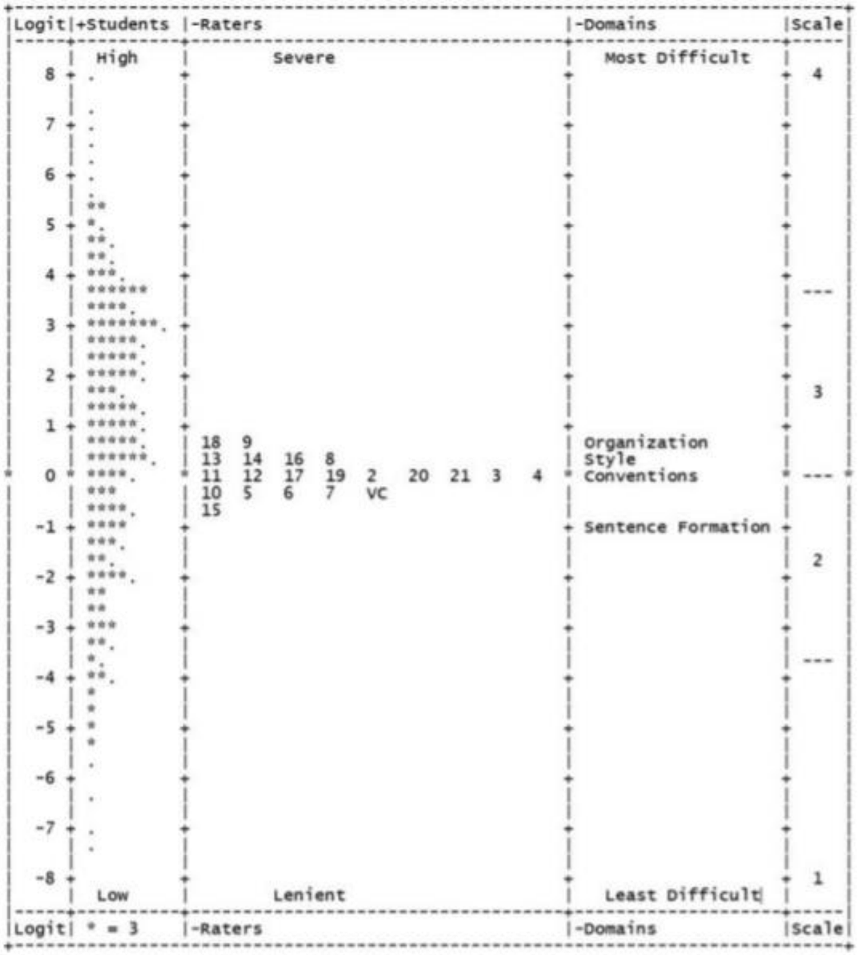
Zwick, R., Thayer, D. T., & Lewis, C. (1999). An Empirical Bayes Approach to Mantel-Haenszel DIF Analysis. *Journal of Educational Measurement*, *36*(1), 1–28. http://www.jstor.org/stable/1435320

Appendix A

North Carolina Professional Teaching Standards and Elements

Standard	Element	Description
I		***Teachers Demonstrate Leadership***
	1a	Teachers lead in their classrooms
	1b	Teachers demonstrate leadership in the school
	1c	Teachers lead the teaching profession
	1d	Teachers advocate for schools and students
	1e	Teachers demonstrate high ethnical standards
II		***Teachers Establish a Respectful Environment for a Diverse Population of Students***
	2a	Teachers provide an environment in which each child has a positive, nurturing relationship with caring adults
	2b	Teachers embrace diversitu in the school community and in the world
	2c	Teachers treat students as individuals
	2d	Teachers adapt their teaching for the benefit of students with special needs
	2e	Teachers work collaboratively with the families and significant adults in the lives of their students
III		***Teachers Know the Content They Teach***
	3a	Teachers align their instruction with the *North Carolina Standard Course of Study*
	3b	Teachers know the content appropriate to their teaching speciality
	3c	Teachers recognize the interconnectedness of content areas/disciplines
	3d	Teachers make instruction relevant to students
IV		***Teachers Facilitate Learning for Their Students***
	4a	Teachers know the ways in which learning takes place, and they know the appropriate levels of intellectual, physical, social, and emotional development of their students
	4b	Teachers plan instruction appropriate for their students
	4c	Teachers use a variety of instructional methods
	4d	Teachers integrate and utilize technology in their instruction
	4e	Teachers help students develop critical-thinking and problem-solving skills
	4f	Teachers help students work in teams and develop leadership qualities
	4g	Teachers communicate effectively
	4h	Teachers use a variety of methods to assess what each student has learned
V		***Teachers Reflect on Their Practice***
	5a	Teachers analyze student learning
	5b	Teachers link professional growth to their professional goals
	5c	Teachers function effectively in a complex, dynamic environment

Appendix B



```
+-----------------------------------------------------------------------------------------+
|Logit|+Students |-Raters                                          |-Domains           |Scale|
|-----+----------+-------------------------------------------------+-------------------+-----|
|     |  High    |          Severe                                 |  Most Difficult   |     |
|   8 + .        +                                                 +                   +  4  |
|     |          |                                                 |                   |     |
|     | .        |                                                 |                   |     |
|   7 + .        +                                                 +                   +     |
|     | .        |                                                 |                   |     |
|     | .        |                                                 |                   |     |
|   6 + .        +                                                 +                   +     |
|     | .        |                                                 |                   |     |
|     | **       |                                                 |                   |     |
|   5 + *.       +                                                 +                   +     |
|     | **.      |                                                 |                   |     |
|     | **.      |                                                 |                   |     |
|   4 + ***.     +                                                 +                   +     |
|     | ******   |                                                 |                   | --- |
|     | ****.    |                                                 |                   |     |
|   3 + *******. +                                                 +                   +     |
|     | *****.   |                                                 |                   |     |
|     | *****.   |                                                 |                   |     |
|   2 + *****.   +                                                 +                   +     |
|     | ***.     |                                                 |                   |  3  |
|     | *****.   |                                                 |                   |     |
|   1 + *****.   +                                                 +                   +     |
|     | *****.   | 18  9                                           | Organization      |     |
|     | ******.  | 13  14  16  8                                   | Style             |     |
*   0 * ****.    * 11  12  17  19  2   20  21  3   4               * Conventions       * --- *
|     | ***      | 10  5   6   7   VC                              |                   |     |
|     | ****.    | 15                                              |                   |     |
|  -1 + ****     +                                                 + Sentence Formation +     |
|     | ***.     |                                                 |                   |     |
|     | **.      |                                                 |                   |  2  |
|  -2 + ****.    +                                                 +                   +     |
|     | **       |                                                 |                   |     |
|     | **       |                                                 |                   |     |
|  -3 + ***      +                                                 +                   +     |
|     | **.      |                                                 |                   |     |
|     | *.       |                                                 |                   | --- |
|  -4 + **.      +                                                 +                   +     |
|     | *        |                                                 |                   |     |
|     | *        |                                                 |                   |     |
|  -5 + *        +                                                 +                   +     |
|     | *        |                                                 |                   |     |
|     | .        |                                                 |                   |     |
|  -6 +          +                                                 +                   +     |
|     | .        |                                                 |                   |     |
|     |          |                                                 |                   |     |
|  -7 + .        +                                                 +                   +     |
|     | .        |                                                 |                   |     |
|     |          |                                                 |                   |     |
|  -8 +          +                                                 +                   +  1  |
|     |  Low     |          Lenient                                |  Least Difficult  |     |
|-----+----------+-------------------------------------------------+-------------------+-----|
|Logit| * = 3    |-Raters                                          |-Domains           |Scale|
+-----------------------------------------------------------------------------------------+
```

Note. From *Invariant Measurement with Raters and Rating Scales* (p. 173), by G. Engelhard, Jr., and S.A. Wind, 2018, Routledge (https://doi.org/10.4324/9781315766829). Copyright 2018 by Taylor & Francis

Appendix C

Calibration of the Rater Strictness in Strictness Calibration Model

Evaluator	Observed	Strictness		Mean Square Errors (*MSE*)	
ID	avg rating	(logit)	*SE*	Infit	Outfit
1	2.00	-0.16	0.11	1.07	1.07
2	2.12	-0.58	0.11	0.66	0.65
3	1.81	0.55	0.11	2.26	2.83
4	2.00	-0.13	0.11	1.05	1.09
5	1.75	0.80	0.11	0.74	0.71
6	2.08	-0.44	0.11	1.48	1.43
7	2.05	-0.34	0.11	1.12	1.10
8	2.22	-0.93	0.11	0.75	0.76
9	1.99	-0.10	0.11	0.91	0.94
10	1.84	0.45	0.11	1.01	1.02
11	1.87	0.33	0.11	1.64	1.87
12	2.15	-0.70	0.11	0.84	0.86
13	2.05	-0.31	0.11	0.83	0.82
14	1.83	0.47	0.11	0.93	0.95
15	1.86	0.37	0.11	0.97	0.92
16	2.14	-0.66	0.11	1.01	1.01
17	1.75	0.78	0.11	0.96	0.91
18	1.89	0.27	0.11	0.77	0.76
19	1.75	0.80	0.11	1.55	2.06
20	2.11	-0.54	0.11	1.94	1.96
21	1.96	0.00	0.11	0.89	0.86
22	2.23	-0.99	0.11	1.00	0.99
23	1.90	0.21	0.11	0.60	0.67
24	2.10	-0.49	0.11	1.61	1.84
25	2.01	-0.17	0.11	0.76	0.80
26	2.06	-0.36	0.11	1.02	1.04
27	1.57	1.51	0.12	0.70	0.61
28	1.85	0.42	0.11	0.81	0.76
29	2.07	-0.41	0.11	0.77	0.77
30	1.95	0.02	0.11	0.85	0.84
31	2.02	-0.22	0.11	0.73	0.74
32	1.92	0.16	0.11	0.97	0.95
33	2.04	-0.29	0.11	0.97	1.00
34	1.96	0.01	0.11	1.17	1.11
35	1.56	1.56	0.12	0.76	0.61

36	2.17	-0.76	0.11	2.87	3.34
37	1.87	0.34	0.11	0.71	0.71
38	1.97	-0.05	0.11	0.70	0.69
39	1.79	0.64	0.11	0.40	0.39
40	1.70	0.99	0.12	0.76	0.76
41	1.68	1.06	0.12	0.64	0.55
42	1.98	-0.06	0.11	1.44	1.46
43	1.98	-0.07	0.11	0.98	0.97
44	1.89	0.24	0.11	0.85	0.83
45	2.08	-0.43	0.11	1.31	1.37
46	1.96	0.01	0.11	0.67	0.66
47	1.85	0.39	0.11	0.67	0.68
48	2.17	-0.74	0.11	0.93	0.88
49	1.87	0.33	0.11	0.57	0.56
50	2.09	-0.47	0.11	0.99	0.98
51	2.19	-0.82	0.11	0.74	0.75
52	1.98	-0.07	0.11	0.93	0.92
53	2.04	-0.29	0.11	0.90	0.89
54	1.93	0.11	0.11	0.78	0.82
55	2.11	-0.55	0.11	1.02	1.00
56	2.14	-0.64	0.11	0.64	0.67
57	1.99	-0.10	0.11	0.58	0.65

Appendix D

Calibration of the Item Difficulty in Strictness Calibration Model

Item	Observed	Item Diff.		Mean Square Errors (*MSE*)	
Number	avg rating	(logit)	*SE*	Infit	Outfit
1	1.96	0.05	0.08	0.86	0.91
2	2.09	-0.44	0.08	1.39	1.38
3	2.01	-0.18	0.08	1.10	1.09
4	1.95	0.06	0.08	0.99	0.98
5	1.94	0.11	0.10	1.15	1.18
6	2.22	-0.94	0.08	0.92	0.94
7	1.86	0.42	0.09	1.25	1.29
8	2.09	-0.44	0.08	1.01	1.03
9	2.09	-0.50	0.08	1.16	1.16
10	2.19	-0.92	0.08	1.23	1.26
11	1.88	0.32	0.08	0.89	0.94
12	1.92	0.18	0.08	0.91	0.95
13	1.89	0.28	0.08	0.99	1.06
14	1.94	0.11	0.08	0.92	0.97
15	1.84	0.46	0.08	0.86	0.91
16	1.92	0.18	0.07	0.87	0.91
17	2.05	-0.30	0.08	0.99	0.99
18	1.98	-0.09	0.09	1.10	1.09
19	1.83	0.49	0.08	1.00	1.02
20	1.84	0.44	0.08	1.15	1.16
21	1.93	0.13	0.08	1.02	1.03
22	1.84	0.43	0.08	0.78	0.81
23	1.85	0.41	0.08	0.91	0.93
24	2.06	-0.38	0.08	1.13	1.12
25	1.96	0.03	0.08	0.94	0.95
26	1.96	-0.02	0.08	0.92	0.90
27	2.09	-0.54	0.08	0.89	0.89
28	1.89	0.25	0.08	0.83	0.88
29	1.86	0.39	0.08	0.81	0.82
30	1.96	0.02	0.08	0.89	0.90

Appendix E

Bias Interaction Terms and t-Scores for All Raters in Bias Test Model

Rater	White Teacher Profiles					TOC Teacher Profiles				
ID	Obs.	Exp.	Bias	*SE*	t-score	Obs.	Exp.	Bias.	*SE*	t-score
1	288	294	-0.14	0.16	-0.91	313	307	0.14	0.15	0.91
2	313	311	0.04	0.15	0.25	323	325	-0.04	0.16	-0.25
3	246	265	-0.52	0.17	-3.12	297	278	0.47	0.15	3.03
4	311	293	0.44	0.15	2.81	288	306	-0.43	0.16	-2.79
5	249	256	-0.20	0.17	-1.18	275	268	0.18	0.16	1.13
6	290	306	-0.38	0.16	-2.45	335	319	0.39	0.16	2.46
7	300	301	-0.03	0.16	-0.21	316	315	0.03	0.15	0.21
8	342	326	0.40	0.16	2.49	323	339	-0.39	0.16	-2.51
9	291	291	-0.01	0.16	-0.06	305	305	0.01	0.15	0.06
10	291	269	0.54	0.16	3.46	260	282	-0.54	0.16	-3.38
11	301	274	0.66	0.15	4.24	260	287	-0.67	0.16	-4.16
12	318	316	0.04	0.16	0.25	328	330	-0.04	0.16	-0.24
13	317	300	0.40	0.15	2.57	297	314	-0.40	0.15	-2.57
14	257	269	-0.31	0.16	-1.89	293	281	0.29	0.15	1.84
15	256	273	-0.43	0.16	-2.66	302	285	0.40	0.15	2.59
16	315	315	0.00	0.15	0.02	328	328	0.00	0.16	-0.01
17	273	257	0.42	0.16	2.66	252	268	-0.42	0.16	-2.59
18	265	277	-0.30	0.16	-1.84	301	289	0.28	0.15	1.80
19	270	256	0.36	0.16	2.26	254	268	-0.35	0.16	-2.19
20	302	310	-0.19	0.15	-1.21	331	323	0.19	0.16	1.23
21	274	287	-0.33	0.16	-2.10	314	301	0.32	0.15	2.08
22	338	329	0.23	0.16	1.48	332	341	-0.23	0.16	-1.49
23	272	279	-0.18	0.16	-1.12	299	292	0.17	0.15	1.10
24	351	308	1.07	0.16	6.56	278	321	-1.04	0.16	-6.62
25	280	294	-0.35	0.16	-2.23	322	308	0.34	0.16	2.22
26	295	302	-0.18	0.16	-1.14	323	316	0.18	0.16	1.14
27	219	231	-0.35	0.18	-2.00	252	240	0.31	0.16	1.92
28	275	271	0.11	0.16	0.68	279	283	-0.10	0.16	-0.66
29	300	304	-0.10	0.16	-0.67	322	318	0.11	0.16	0.68
30	274	286	-0.31	0.16	-1.95	312	300	0.30	0.15	1.93
31	298	296	0.04	0.16	0.25	308	310	-0.04	0.15	-0.25
32	288	281	0.17	0.16	1.10	287	294	-0.17	0.16	-1.08
33	280	299	-0.47	0.16	-2.99	332	313	0.47	0.16	2.99
34	284	287	-0.07	0.16	-0.46	303	300	0.07	0.15	0.46
35	235	229	0.17	0.17	1.00	233	239	-0.17	0.17	-0.98

36	381	319	1.68	0.18	9.27	270	332	-1.50	0.16	-9.49
37	260	274	-0.35	0.16	-2.18	300	286	0.33	0.15	2.13
38	287	289	-0.06	0.16	-0.38	305	303	0.06	0.15	0.38
39	257	262	-0.13	0.16	-0.81	279	274	0.12	0.16	0.78
40	243	249	-0.17	0.17	-0.99	266	260	0.15	0.16	0.94
41	241	247	-0.16	0.17	-0.92	263	258	0.14	0.16	0.88
42	290	290	0.00	0.16	0.01	303	303	0.00	0.15	-0.01
43	293	290	0.06	0.16	0.40	301	304	-0.06	0.15	-0.40
44	283	278	0.13	0.16	0.85	285	290	-0.13	0.16	-0.84
45	286	305	-0.47	0.16	-2.99	338	319	0.47	0.16	3.00
46	277	287	-0.25	0.16	-1.56	310	300	0.24	0.15	1.54
47	263	272	-0.23	0.16	-1.39	293	284	0.21	0.15	1.36
48	324	318	0.14	0.16	0.87	326	332	-0.14	0.16	-0.87
49	269	274	-0.13	0.16	-0.82	292	287	0.12	0.16	0.80
50	303	307	-0.09	0.15	-0.59	324	320	0.09	0.16	0.60
51	329	321	0.18	0.16	1.18	327	335	-0.18	0.16	-1.18
52	298	290	0.18	0.16	1.18	296	304	-0.18	0.15	-1.17
53	297	299	-0.06	0.16	-0.36	315	313	0.06	0.15	0.37
54	287	283	0.10	0.16	0.63	292	296	-0.10	0.16	-0.62
55	294	310	-0.39	0.16	-2.53	340	324	0.40	0.16	2.54
56	308	314	-0.14	0.15	-0.91	333	327	0.14	0.16	0.92
57	288	291	-0.08	0.16	-0.53	308	305	0.08	0.15	0.53

Appendix F

Calibration of Rater Strictness in Summative Caseloads Model

Rater	Observed	Strictness		Mean Square Errors *(MSE)*	
ID	avg rating	(logit)	*SE*	Infit	Outfit
1	1.96	-0.22	0.09	1.02	0.98
2	2.01	-0.74	0.08	0.78	0.73
3	1.83	0.72	0.08	1.72	1.95
4	2.10	-0.19	0.08	0.95	0.96
5	1.83	1.05	0.08	0.80	0.76
6	2.01	-0.58	0.08	1.21	1.15
7	2.15	-0.44	0.11	1.28	1.29
8	2.22	-1.17	0.12	0.93	0.95
9	2.15	-0.14	0.07	0.82	0.80
10	2.24	0.58	0.10	1.14	1.18
11	2.07	0.41	0.09	1.44	1.54
12	2.07	-0.89	0.07	0.74	0.70
13	2.00	-0.41	0.10	0.95	0.94
14	2.04	0.60	0.07	0.84	0.83
15	1.98	0.46	0.08	0.78	0.76
16	2.32	-0.85	0.08	0.97	0.93
17	1.84	1.03	0.08	0.92	0.90
18	1.95	0.33	0.10	0.79	0.73
19	1.89	1.05	0.12	1.78	2.31
20	2.19	-0.70	0.08	1.59	1.62
21	1.86	-0.02	0.10	0.94	0.92
22	2.35	-1.25	0.09	0.91	0.88
23	2.06	0.25	0.08	0.66	0.64
24	2.08	-0.64	0.10	1.54	1.62
25	2.03	-0.23	0.08	0.59	0.56
26	2.09	-0.47	0.10	1.12	1.14
27	1.70	2.05	0.10	0.92	0.87
28	1.98	0.53	0.11	0.93	0.89
29	2.17	-0.53	0.08	0.91	0.89
30	2.14	0.01	0.10	1.00	0.98
31	2.16	-0.29	0.08	0.87	0.86
32	2.03	0.19	0.11	1.15	1.15
33	2.02	-0.38	0.09	1.02	1.04
34	2.21	0.00	0.08	1.14	1.15
35	2.00	2.10	0.07	0.67	0.60

36	2.17	-0.96	0.12	3.50	3.95
37	1.87	0.43	0.13	0.96	0.97
38	2.01	-0.08	0.10	0.83	0.79
39	2.00	0.84	0.09	0.45	0.41
40	1.85	1.33	0.10	1.01	1.03
41	1.93	1.42	0.09	0.84	0.77
42	2.00	-0.09	0.11	1.68	1.68
43	1.96	-0.11	0.08	0.99	0.97
44	1.88	0.30	0.10	0.99	0.97
45	2.24	-0.56	0.09	1.29	1.33
46	2.00	0.00	0.09	0.80	0.77
47	2.01	0.50	0.10	0.80	0.79
48	2.04	-0.95	0.08	0.77	0.71
49	1.99	0.41	0.09	0.55	0.49
50	2.09	-0.61	0.12	1.25	1.34
51	2.18	-1.04	0.11	0.90	0.87
52	2.06	-0.11	0.11	1.06	1.04
53	2.22	-0.38	0.09	0.96	0.93
54	1.86	0.12	0.10	0.90	0.89
55	2.08	-0.71	0.10	0.95	0.95
56	2.14	-0.82	0.12	0.78	0.78
57	1.98	-0.14	0.09	0.57	0.54

Appendix G

Calibration of Fictitious Teacher Profiles in Summative Caseloads Model

T. Profile	Observed	Ability		*MSE*			
ID	avg rating	(logit)	*SE*	Infit	Std. infit	Outfit	Std. outfit
1	2.07	0.38	0.05	1.32	7.00	1.35	7.10
2	1.78	-1.08	0.05	1.43	9.00	1.49	9.00
3	2.26	1.31	0.05	1.14	4.20	1.16	4.30
4	1.25	-3.59	0.06	1.27	7.80	1.37	6.30
5	2.54	2.50	0.05	1.54	9.00	1.61	9.00
6	2.21	1.07	0.05	1.31	8.20	1.28	6.80
7	1.38	-2.90	0.05	1.09	3.30	1.09	2.50
8	1.63	-1.76	0.05	1.22	7.20	1.26	7.30
9	2.16	0.81	0.05	1.42	9.00	1.44	9.00
10	2.35	1.68	0.05	1.15	5.40	1.15	4.80

Appendix H

Calibration of Item Difficulty in Summative Caseloads Model

Item	Observed	Item Diff.			Mean Square Errors *(MSE)*		
Number	avg rating	(logit)	*SE*	Infit	Std. infit	Outfit	Std. outfit
1	2.03	0.13	0.06	0.88	-3.00	0.89	-2.20
2	2.11	-0.25	0.06	1.37	8.30	1.38	7.00
3	2.08	-0.20	0.07	1.11	2.50	1.15	2.90
4	2.02	0.11	0.07	1.03	0.70	1.01	0.10
5	2.11	-0.48	0.07	1.11	2.40	1.10	1.70
6	2.35	-1.39	0.07	0.96	-1.10	0.96	-0.60
7	1.95	0.46	0.07	1.24	4.90	1.28	4.70
8	2.20	-0.68	0.06	0.97	-0.70	0.95	-0.90
9	2.08	-0.21	0.07	1.19	4.20	1.18	3.40
10	2.24	-1.17	0.07	1.17	4.20	1.22	3.80
11	1.96	0.40	0.06	0.89	-2.60	0.89	-2.30
12	2.02	0.15	0.06	0.89	-2.60	0.87	-2.60
13	1.95	0.45	0.07	0.98	-0.50	1.00	0.10
14	2.04	0.08	0.06	0.88	-2.90	0.89	-2.30
15	1.94	0.53	0.07	0.87	-3.00	0.87	-2.70
16	1.99	0.29	0.06	0.86	-3.50	0.87	-2.50
17	2.10	-0.32	0.07	0.96	-0.90	0.91	-1.80
18	2.06	-0.12	0.07	1.18	3.80	1.18	3.10
19	1.94	0.52	0.07	1.02	0.30	1.00	0.00
20	1.93	0.54	0.06	1.17	3.70	1.17	3.10
21	2.03	0.05	0.07	0.98	-0.50	0.95	-0.90
22	1.89	0.74	0.06	0.86	-3.50	0.88	-2.30
23	1.94	0.52	0.07	0.91	-2.00	0.90	-1.90
24	2.10	-0.35	0.07	1.14	3.10	1.12	2.20
25	1.99	0.23	0.07	1.00	0.10	0.99	-0.10
26	2.06	-0.12	0.07	0.88	-3.00	0.81	-3.80
27	2.17	-0.74	0.07	0.86	-3.70	0.79	-4.30
28	1.98	0.30	0.07	0.81	-4.60	0.79	-4.50
29	1.94	0.50	0.07	0.81	-4.60	0.76	-4.80
30	2.03	0.03	0.07	0.90	-2.30	0.83	-3.30

Appendix I

Comparison of Logit Locations for Rater Strictness

Strictness Calibration Model & Summative Caseloads Model

Rater	Model 1	Model 3	Difference	Change is	Change in
ID	(logit)	(logit)	(logit)	>|0.5|?	Strictness?
1	-0.16	-0.22	-0.06	No	No
2	-0.58	-0.74	-0.16	No	No
3	0.55	0.72	0.17	No	No
4	-0.13	-0.19	-0.06	No	No
5	0.80	1.05	0.25	No	No
6	-0.44	-0.58	-0.14	No	**Yes**
7	-0.34	-0.44	-0.10	No	No
8	-0.93	-1.17	-0.24	No	No
9	-0.10	-0.14	-0.04	No	No
10	0.45	0.58	0.13	No	**Yes**
11	0.33	0.41	0.08	No	No
12	-0.70	-0.89	-0.19	No	No
13	-0.31	-0.41	-0.10	No	No
14	0.47	0.60	0.13	No	**Yes**
15	0.37	0.46	0.09	No	No
16	-0.66	-0.85	-0.19	No	No
17	0.78	1.03	0.25	No	No
18	0.27	0.33	0.06	No	No
19	0.80	1.05	0.25	No	No
20	-0.54	-0.70	-0.16	No	No
21	0.00	-0.02	-0.02	No	No
22	-0.99	-1.25	-0.26	No	No
23	0.21	0.25	0.04	No	No
24	-0.49	-0.64	-0.15	No	**Yes**
25	-0.17	-0.23	-0.06	No	No
26	-0.36	-0.47	-0.11	No	No
27	1.51	2.05	0.54	**Yes**	No
28	0.42	0.53	0.11	No	No
29	-0.41	-0.53	-0.12	No	No
30	0.02	0.01	-0.01	No	No
31	-0.22	-0.29	-0.07	No	No
32	0.16	0.19	0.03	No	No
33	-0.29	-0.38	-0.09	No	No
34	0.01	0.00	-0.01	No	No

35	1.56	2.10	0.54	**Yes**	No
36	-0.76	-0.96	-0.20	No	No
37	0.34	0.43	0.09	No	No
38	-0.05	-0.08	-0.03	No	No
39	0.64	0.84	0.20	No	No
40	0.99	1.33	0.34	No	No
41	1.06	1.42	0.36	No	No
42	-0.06	-0.09	-0.03	No	No
43	-0.07	-0.11	-0.04	No	No
44	0.24	0.30	0.06	No	No
45	-0.43	-0.56	-0.13	No	No
46	0.01	0.00	-0.01	No	No
47	0.39	0.50	0.11	No	No
48	-0.74	-0.95	-0.21	No	No
49	0.33	0.41	0.08	No	No
50	-0.47	-0.61	-0.14	No	**Yes**
51	-0.82	-1.04	-0.22	No	No
52	-0.07	-0.11	-0.04	No	No
53	-0.29	-0.38	-0.09	No	No
54	0.11	0.12	0.01	No	No
55	-0.55	-0.71	-0.16	No	No
56	-0.64	-0.82	-0.18	No	No
57	-0.10	-0.14	-0.04	No	No

Appendix J

Comparison of Infit and Outfit *MSE* Statistics for Rater Strictness
Strictness Calibration Model & Summative Caseloads Model

Rater ID	Model 1 infit	Model 3 infit	Difference	Change is >0.5?	Model 1 outfit	Model 3 outfit	Difference	Change is >|0.5|?
1	1.07	1.02	-0.05	No	1.07	0.98	-0.09	No
2	0.66	0.78	0.12	No	0.65	0.73	0.08	No
3	2.26	1.72	-0.54	**Yes**	2.83	1.95	-0.88	**Yes**
4	1.05	0.95	-0.10	No	1.09	0.96	-0.13	No
5	0.74	0.80	0.06	No	0.71	0.76	0.05	No
6	1.48	1.21	-0.27	No	1.43	1.15	-0.28	No
7	1.12	1.28	0.16	No	1.10	1.29	0.19	No
8	0.75	0.93	0.18	No	0.76	0.95	0.19	No
9	0.91	0.82	-0.09	No	0.94	0.80	-0.14	No
10	1.01	1.14	0.13	No	1.02	1.18	0.16	No
11	1.64	1.44	-0.20	No	1.87	1.54	-0.33	No
12	0.84	0.74	-0.10	No	0.86	0.70	-0.16	No
13	0.83	0.95	0.12	No	0.82	0.94	0.12	No
14	0.93	0.84	-0.09	No	0.95	0.83	-0.12	No
15	0.97	0.78	-0.19	No	0.92	0.76	-0.16	No
16	1.01	0.97	-0.04	No	1.01	0.93	-0.08	No
17	0.96	0.92	-0.04	No	0.91	0.90	-0.01	No
18	0.77	0.79	0.02	No	0.76	0.73	-0.03	No
19	1.55	1.78	0.23	No	2.06	2.31	0.25	No
20	1.94	1.59	-0.35	No	1.96	1.62	-0.34	No
21	0.89	0.94	0.05	No	0.86	0.92	0.06	No
22	1.00	0.91	-0.09	No	0.99	0.88	-0.11	No
23	0.60	0.66	0.06	No	0.67	0.64	-0.03	No
24	1.61	1.54	-0.07	No	1.84	1.62	-0.22	No
25	0.76	0.59	-0.17	No	0.80	0.56	-0.24	No
26	1.02	1.12	0.10	No	1.04	1.14	0.10	No
27	0.70	0.92	0.22	No	0.61	0.87	0.26	No
28	0.81	0.93	0.12	No	0.76	0.89	0.13	No
29	0.77	0.91	0.14	No	0.77	0.89	0.12	No
30	0.85	1.00	0.15	No	0.84	0.98	0.14	No
31	0.73	0.87	0.14	No	0.74	0.86	0.12	No
32	0.97	1.15	0.18	No	0.95	1.15	0.20	No
33	0.97	1.02	0.05	No	1.00	1.04	0.04	No
34	1.17	1.14	-0.03	No	1.11	1.15	0.04	No
35	0.76	0.67	-0.09	No	0.61	0.60	-0.01	No
36	2.87	3.50	0.63	**Yes**	3.34	3.95	0.61	**Yes**
37	0.71	0.96	0.25	No	0.71	0.97	0.26	No

38	0.70	0.83	0.13	No	0.69	0.79	0.10	No
39	0.40	0.45	0.05	No	0.39	0.41	0.02	No
40	0.76	1.01	0.25	No	0.76	1.03	0.27	No
41	0.64	0.84	0.20	No	0.55	0.77	0.22	No
42	1.44	1.68	0.24	No	1.46	1.68	0.22	No
43	0.98	0.99	0.01	No	0.97	0.97	0.00	No
44	0.85	0.99	0.14	No	0.83	0.97	0.14	No
45	1.31	1.29	-0.02	No	1.37	1.33	-0.04	No
46	0.67	0.80	0.13	No	0.66	0.77	0.11	No
47	0.67	0.80	0.13	No	0.68	0.79	0.11	No
48	0.93	0.77	-0.16	No	0.88	0.71	-0.17	No
49	0.57	0.55	-0.02	No	0.56	0.49	-0.07	No
50	0.99	1.25	0.26	No	0.98	1.34	0.36	No
51	0.74	0.90	0.16	No	0.75	0.87	0.12	No
52	0.93	1.06	0.13	No	0.92	1.04	0.12	No
53	0.90	0.96	0.06	No	0.89	0.93	0.04	No
54	0.78	0.90	0.12	No	0.82	0.89	0.07	No
55	1.02	0.95	-0.07	No	1.00	0.95	-0.05	No
56	0.64	0.78	0.14	No	0.67	0.78	0.11	No
57	0.58	0.57	-0.01	No	0.65	0.54	-0.11	No

Appendix K

Comparison of Logit Locations for Item Difficulty
Strictness Calibration Model & Summative Caseloads Model

Item Number	Model 1 Logit	Model 3 Logit	Difference	>0.5?
1	0.05	0.13	0.08	No
2	-0.44	-0.25	0.19	No
3	-0.18	-0.20	-0.02	No
4	0.06	0.11	0.05	No
5	0.11	-0.48	-0.59	No
6	-0.94	-1.39	-0.45	No
7	0.42	0.46	0.04	No
8	-0.44	-0.68	-0.24	No
9	-0.50	-0.21	0.29	No
10	-0.92	-1.17	-0.25	No
11	0.32	0.40	0.08	No
12	0.18	0.15	-0.03	No
13	0.28	0.45	0.17	No
14	0.11	0.08	-0.03	No
15	0.46	0.53	0.07	No
16	0.18	0.29	0.11	No
17	-0.30	-0.32	-0.02	No
18	-0.09	-0.12	-0.03	No
19	0.49	0.52	0.03	No
20	0.44	0.54	0.10	No
21	0.13	0.05	-0.08	No
22	0.43	0.74	0.31	No
23	0.41	0.52	0.11	No
24	-0.38	-0.35	0.03	No
25	0.03	0.23	0.20	No
26	-0.02	-0.12	-0.10	No
27	-0.54	-0.74	-0.20	No
28	0.25	0.30	0.05	No
29	0.39	0.50	0.11	No
30	0.02	0.03	0.01	No

Appendix L

Comparison of Infit and Outfit *MSE* Statistics for Item Difficulty
Strictness Calibration Model & Summative Caseloads Model

Item Number	Model 1 infit	Model 3 infit	Difference	Model 1 outfit	Model 3 outfit	Difference
1	0.86	0.88	0.02	0.91	0.89	-0.02
2	1.39	1.37	-0.02	1.38	1.38	0.00
3	1.10	1.11	0.01	1.09	1.15	0.06
4	0.99	1.03	0.04	0.98	1.01	0.03
5	1.15	1.11	-0.04	1.18	1.10	-0.08
6	0.92	0.96	0.04	0.94	0.96	0.02
7	1.25	1.24	-0.01	1.29	1.28	-0.01
8	1.01	0.97	-0.04	1.03	0.95	-0.08
9	1.16	1.19	0.03	1.16	1.18	0.02
10	1.23	1.17	-0.06	1.26	1.22	-0.04
11	0.89	0.89	0.00	0.94	0.89	-0.05
12	0.91	0.89	-0.02	0.95	0.87	-0.08
13	0.99	0.98	-0.01	1.06	1.00	-0.06
14	0.92	0.88	-0.04	0.97	0.89	-0.08
15	0.86	0.87	0.01	0.91	0.87	-0.04
16	0.87	0.86	-0.01	0.91	0.87	-0.04
17	0.99	0.96	-0.03	0.99	0.91	-0.08
18	1.10	1.18	0.08	1.09	1.18	0.09
19	1.00	1.02	0.02	1.02	1.00	-0.02
20	1.15	1.17	0.02	1.16	1.17	0.01
21	1.02	0.98	-0.04	1.03	0.95	-0.08
22	0.78	0.86	0.08	0.81	0.88	0.07
23	0.91	0.91	0.00	0.93	0.90	-0.03
24	1.13	1.14	0.01	1.12	1.12	0.00
25	0.94	1.00	0.06	0.95	0.99	0.04
26	0.92	0.88	-0.04	0.90	0.81	-0.09
27	0.89	0.86	-0.03	0.89	0.79	-0.10
28	0.83	0.81	-0.02	0.88	0.79	-0.09
29	0.81	0.81	0.00	0.82	0.76	-0.06
30	0.89	0.90	0.01	0.90	0.83	-0.07

Appendix M

Measure	Fictitious Teacher Profiles	Raters	Race of Teacher in Profile	Items	Rubric Rating Scale
Logits	High Quality	Strict	High Quality	Difficult	
2.5					
2					
	5 (A)				
1.5		27 35			
	10 (A)				Accomplished
					3
					Proficient
1		40 41			2
	3 6 (P)	5 17 19			
	9 (P)	39			
0.5		3		S4.E S4.A	
		10 14 28		S4.F S4.H S2.B S.5A	
		47 37 49 15 11		S4 S3.A	
	1 (P)	18 23 44		S3.C S3	
		32 54	White	S4.B S3.B S4.G S1.E S3.D	
0		30 34 46 21 38 57		S1.D S1.A S5.C S5 S1 S4.D	
		42 43 52 9 25 4 1	TOC	S1.C	
		31			
		13 33 53 7 26		S4.C S5.B	
		45 29 6 50		S2.C S1.B	
-0.5		24 20 55		S2.D S2	
		2 56 16			
		12 36 48			
	2 (P)	51			
		8		S2.E S2.A	
-1		22			
	8 (P)				
-1.5					Proficient
					2
					Developing
					1
-2					
	7 (D)				
-2.5					
	4 (D)				
-3					
	Low Quality	Lenient	Low Quality	Easy	

Appendix N

Measure	Fictititous Teacher Profiles	Raters	Items	Rubric Rating Scale
Logits	High Quality	Strict	Difficult	
2.5	5 (A)			
		35		Accomplished
2		27		3
				Proficient
				2
	10 (P)			
1.5				
		41		
	3 (P)	40		
1	6 (P)	5 17 19		
	9 (P)	39		
		3	S4.8	
		10 14		
0.5		11 15 28 47	S4.1 S4.5 S4.6 S5.1 S4	
	1 (P)	37 49	S2.2 S3.1 S3.3	
		18 44	S4.2 S3	
		23 32	S5.3	
		54	S1.1. S1.4 S3.2	
0		21 30 34 38 42 46	S3.4. S4.7 S5	
		9 25 43 52 57	S4.4. S1	
		1 4	S1.2 S1.3 S2.4	
		18 31	S4.3 S5.2	
		7 13 26 33 53		
-0.5		29 45	S1.5	
		6 24 50	S2.3	
		2 20 55	S2	
		16 56		
		12		
-1	2 (P)	36 48 51		
		8	S2.5	
		22		
			S2.1	
-1.5				
	8 (P)			
-2				Proficient
				2
				Developing
				1
-2.5				
	7 (D)			
-3				
-3.5	4 (D)			
	Low Quality	Lenient	Easy	