




2021

RATE TO MEASURE MATHEMATICS TEACHING: USING THE MANY-FACET RASCH MODELING TO REEVALUATE THE MATHEMATICS CLASSROOM OBSERVATION PROTOCOL FOR PRACTICES (MCOP2)

Chunling Niu

University of Kentucky, chunling.niu@gmail.com

Author ORCID Identifier:

 <https://orcid.org/0000-0002-9106-0417>

Digital Object Identifier: <https://doi.org/10.13023/etd.2021.158>

[Right click to open a feedback form in a new tab to let us know how this document benefits you.](#)

Recommended Citation

Niu, Chunling, "RATE TO MEASURE MATHEMATICS TEACHING: USING THE MANY-FACET RASCH MODELING TO REEVALUATE THE MATHEMATICS CLASSROOM OBSERVATION PROTOCOL FOR PRACTICES (MCOP2)" (2021). *Theses and Dissertations--Education Sciences*. 90.
https://uknowledge.uky.edu/edsc_etds/90

This Doctoral Dissertation is brought to you for free and open access by the College of Education at UKnowledge. It has been accepted for inclusion in Theses and Dissertations--Education Sciences by an authorized administrator of UKnowledge. For more information, please contact UKnowledge@lsv.uky.edu.

STUDENT AGREEMENT:

I represent that my thesis or dissertation and abstract are my original work. Proper attribution has been given to all outside sources. I understand that I am solely responsible for obtaining any needed copyright permissions. I have obtained needed written permission statement(s) from the owner(s) of each third-party copyrighted matter to be included in my work, allowing electronic distribution (if such use is not permitted by the fair use doctrine) which will be submitted to UKnowledge as Additional File.

I hereby grant to The University of Kentucky and its agents the irrevocable, non-exclusive, and royalty-free license to archive and make accessible my work in whole or in part in all forms of media, now or hereafter known. I agree that the document mentioned above may be made available immediately for worldwide access unless an embargo applies.

I retain all other ownership rights to the copyright of my work. I also retain the right to use in future works (such as articles or books) all or part of my work. I understand that I am free to register the copyright to my work.

REVIEW, APPROVAL AND ACCEPTANCE

The document mentioned above has been reviewed and accepted by the student's advisor, on behalf of the advisory committee, and by the Director of Graduate Studies (DGS), on behalf of the program; we verify that this is the final, approved version of the student's thesis including all changes required by the advisory committee. The undersigned agree to abide by the statements above.

Chunling Niu, Student

Dr. Kelly Bradley, Major Professor

Dr. Jane Jensen, Director of Graduate Studies

RATE TO MEASURE MATHEMATICS TEACHING:
USING THE MANY-FACET RASCH MODELING TO REEVALUATE THE
MATHEMATICS CLASSROOM OBSERVATION PROTOCOL FOR PRACTICES
(MCOP²)

DISSERTATION

A dissertation submitted in partial fulfillment of the
requirements for the degree of Doctor of Philosophy in the
College of Education
at the University of Kentucky

By

Chunling Niu

Lexington, Kentucky

Co- Directors: Dr. Kelly Bradley, Professor of Education

and Dr. Margret Schroeder, Professor of Education

Lexington, Kentucky

2021

Copyright © Chunling Niu 2021
<https://orcid.org/0000-0002-9106-0417>

ABSTRACT OF DISSERTATION

RATE TO MEASURE MATHEMATICS TEACHING: USING THE MANY-FACET RASCH MODELING TO REEVALUATE THE MATHEMATICS CLASSROOM OBSERVATION PROTOCOL FOR PRACTICES (MCOP²)

Rater-mediated classroom observation protocols are increasingly being used for teaching performance assessments, which makes identifying and controlling for various rater effects a central issue to ensure the rating quality. A series of validation studies under the classical test theory framework, including content validity, interrater reliability, and structure analysis, have been completed for the 16-item Mathematics Classroom Observation Protocol for Practices (MCOP²).

However, the MCOP² data have never been investigated under the Rasch framework. Due to the methodological limitations of the CTT approach for rater-mediated assessments, it is imperative to examine the MCOP² validity and reliability using the MFRM modeling technique to implement dimensionality analysis, item-level analysis, rater effects control, and ratee and rater ability level calibration.

To that end, two existing samples of the MCOP² data were obtained and analyzed, where twelve raters were asked to rate 237 math classroom observations, using the MCOP² classroom observation protocol. The data were analyzed under the MFRM framework, using Facets 3.83.3.

Results of the Facets analysis showed that both the MCOP² subscales (i.e., Student Engagement & Teacher Facilitation) were valid, unidimensional, and highly reliable rater-mediated performance measures across raters, ratees, and study samples. However, rater-item bias analyses revealed a type of intra-rater inconsistency, where some raters tended to rate more severely than other raters on certain items while more leniently on some other items.

The overall findings are promising in that they provide systematic preliminary psychometric evidence for the viability of the MCOP² protocol to be used for math teachers' self-assessment and/or peer-assessment along with other designated raters in the future studies.

KEYWORDS: Teaching Performance Assessments, Rater-Mediated Performance, MCOP², MFRM, Rater Effects, Rater Bias Analysis

Chunling Niu

(Name of Student)

05/14/2021

Date

RATE TO MEASURE MATHEMATICS TEACHING:
USING THE MANY-FACET RASCH MODELING TO REEVALUATE THE
MATHEMATICS CLASSROOM OBSERVATION PROTOCOL FOR PRACTICES
(MCOP²)

By
Chunling Niu

Dr. Kelly Bradley

Co-Director of Dissertation

Dr. Margret Schroeder

Co-Director of Dissertation

Dr. Jane Jensen

Director of Graduate Studies

05/14/2021

Date

DEDICATION

This dissertation is dedicated to my parents, the late Yi Niu and Zhuangru He, who taught me about persistence and hope. Also, I dedicate this work to my fellow doctoral students, Summer, Bruce, LJ, and Rainbow, whose company and friendship made me feel so young and special. Finally, I dedicate this dissertation to my darling boy, Han Niu, for your wonderful presence in my life. You have always been the one reason for me to keep striving for the better. I thank God every day for the great privilege of becoming your Mama.

ACKNOWLEDGMENTS

First, I would like to express my deepest appreciation to my committee chair, Professor Kelly Bradley, who so willingly stepped up to take the role of my dissertation committee chair in the time of dire need. She exemplified the attitude and substance of a true scholar and kept pushing me for quality scholarly work by continually and convincingly providing me invaluable guidance and advice. Without her encouragement and mentorship, the completion of this work would not have been possible.

Next, I would like to thank my committee co-chair, Professor Margret Schroeder, who took precious time out of her very busy schedules to ensure my access to the secondary research data needed for my dissertation project. Her genius in breaking down complex research problems with simplicity and clarity, and her directness and sincerity in communicating with research collaborators have always been such an amazing inspiration to me.

I also would like to thank my committee member, Dr. Brent Harrison, for his kind willingness to take on an extra heavy workload (i.e., reading someone else's lengthy dissertation research) under tight time constraints in addition to his already busy working schedules. I especially appreciated him listening to my concerns and plans related to my future research and career, and I enjoyed "brainstorming" with him on various alternatives/possibilities to solve my problems.

Sincere thanks also go out to Dr. Shannon Sampson, for her firm and selfless commitment in supporting student researchers such as myself to succeed both academically and personally. I will always remain grateful for her on-point suggestions and challenging thoughts which contributed to the continuous improvement of my

dissertation work quality.

Finally, a very special thank you goes out to Dr. Morris Grubbs, for his patiently reading and reviewing my dissertation draft as my outside examiner. Your praise and kind words, as well as your very detailed editing advice, meant a lot to me in this journey of completing my second doctorate program.

TABLE OF CONTENTS

ACKNOWLEDGMENTS	iii
LIST OF TABLES	ix
LIST OF FIGURES	xi
CHAPTER I: STATEMENT OF THE PROBLEM	1
Introduction	1
Rater-Mediated Performance Assessment	2
Teaching Performance Assessment	3
Mathematics Classroom Observation Protocol for Practices	4
Many-faceted Rasch Model (MFRM)	6
The Problem Defined	7
Purpose of the Study	9
Research Questions	9
Significance of the Study	11
MFRM-Based Validation Study	11
MFRM-Calibrated Teacher Peer- and Self-Assessments	12
Limitations of the Study	16
CHAPTER II: REVIEW OF THE LITERATURE	19
Introduction	19
Summary of Earlier MCOP ² Validation and Empirical Studies	21
Development of MCOP ²	21
Existing MCOP ² Validation Studies	25

Existing MCOP ² Empirical Studies.....	29
Limitations in the MCOP ² Research	32
A Calibrated Framework for Rater-Mediated Assessment (MFRM).....	34
MFRM-Based Dimensionality Analysis	35
Rater Effects	38
Rater severity.....	39
Rater centrality.	41
Rater misfit.	43
Interrater Reliability	44
Multi-Facet Calibration Controlling for Rater Effects	48
CHAPTER III: METHOD.....	52
Introduction	52
Research Questions	53
Research Design	54
Participants	54
Ratees	55
Raters.....	56
Mathematics Classroom Observation Protocol for Practices (MCOP ²)	57
Data Analysis.....	60
Analyses Plan for Research Question One	60
Analyses Plan for Research Question Two	66
Analyses Plan for Research Question Three	67
Analyses Plan for Research Question Four	69

Analyses Plan for Research Question Five	70
Analyses Plan for Research Question Six	73
Analyses Plan for Research Question Seven.....	75
Ethical Standards.....	75
CHAPTER IV: RESULTS	77
Introduction	77
Research Questions	78
Descriptive Statistics	80
Analyses for Research Question One	84
Local Independence.....	84
Unidimensionality	86
Overall Model Fit	89
Rater Fit and Item Fit	89
Analyses for Research Question Two	92
Analyses for Research Question Three	98
Analyses for Research Question Four	100
Analyses for Research Question Five.....	101
Analyses for Research Question Six	107
Analyses for Research Question Seven.....	115
Summary.....	118
CHAPTER V: DISCUSSION AND CONCLUSIONS.....	121
The Study in Brief	121
Discussion.....	122

Sample Characteristics	122
Research Question 1	125
Research Question 2	134
Research Question 3	136
Research Question 4	139
Research Question 5	141
Research Question 6	143
Research Question 7	145
Implications	147
Limitations.....	151
Future Research.....	152
APPENDIX A: IRB Approval Letter	154
APPENDIX B: Mathematics Classroom Observation Protocol for Practices.....	155
APPENDIX C: Mathematics Classroom Observation Protocol for Practices: Descriptors Manual	159
REFERENCES	178
VITA.....	200

LIST OF TABLES

Table 1. Descriptive Statistics for Demographic Variables in the Combined Sample ($N = 159$).....82

Table 2. Descriptive Statistics for Participants' MCOP² Raw Scores in the Respective AL, UK, & Combined Samples.....83

Table 3. Summary of the Local Independence (Yen's Q_3) Statistics for the MCOP² Protocol, Student Engagement Subscale, and Teacher Facilitation Subscale.85

Table 4. Summary of the PCA Statistics for the MCOP² Protocol, Student Engagement Subscale, and Teacher Facilitation Subscale.....87

Table 5.1. Percentages of Rater Mean-Square Fit Statistics for the Student Engagement Subscale and the Teacher Facilitation Subscale.90

Table 5.2. Percentages of Item Mean-Square Fit Statistics for the Student Engagement Subscale and the Teacher Facilitation Subscale.....91

Table 6.1. Summary of the MFRM Analysis Statistics for the Student Engagement Subscale.....96

Table 6.2. Summary of the MFRM Analysis Statistics for the Teacher Facilitation Subscale.....97

Table 7.1. Summary Statistics of the Interaction Analysis for the Student Engagement Subscale.....104

Table 7.2. Summary Statistics of the Interaction Analysis for the Teacher Facilitation Subscale.....106

Table 8.1. Ratees' Observed and Fair Scores on Student Engagement by Sites, Service Types, or Classroom Grade Levels	116
Table 8.2. Ratees' Observed and Fair Scores on Teacher Facilitation by Sites, Service Types, or Classroom Grade Levels	118

LIST OF FIGURES

Figure 1. The MCOP2 Theoretical Model.	58
Figure 2. The MCOP ² Scoring Roadmap	59
Figure 3.1. The Student Engagement Subscale Wright Map	93
Figure 3.2. The Teacher Facilitation Subscale Wright Map... ..	94
Figure 4.1. The Plot Illustrating the Rater by Item Bias Interactions for the Student Engagement Subscale	105
Figure 4.2. The Plot Illustrating the Rater by Item Bias Interactions for the Teacher Facilitation Subscale	107
Figure 5.1. Summary Statistics of the Rating Scale Functioning for the Student Engagement Subscale	108
Figure 5.2. The Student Engagement Subscale Probability Category Curves (PCCs)..	110
Figure 6.1. Summary Statistics of the Rating Scale Functioning for the Student Engagement Subscale	109
Figure 6.2. The Teacher Facilitation Subscale Probability Category Curves (PCCs)....	110
Figure 7.1. The Student Engagement Subscale Category Information Function	112
Figure 7.2. The Teacher Facilitation Subscale Category Information Function... ..	112
Figure 8.1. The Student Engagement Subscale Item Information Functions	114
Figure 8.2. The Teacher Facilitation Subscale Item Information Functions.....	114

CHAPTER I:
STATEMENT OF THE PROBLEM

Introduction

Teaching performance assessment has been managed unscientifically and measured poorly for too long. The various vague, multivariate definitions of effective teaching performance that exist in pertinent literature have made accurate measurement/assessment of teaching performance nearly impossible, because these definitions tend to mix values, emotions, personality, behaviors, processes, teaching contexts, and even outcomes into one or more unidimensional latent traits. To make matters worse, teaching performances are often assessed by human raters from different backgrounds (e.g., teacher educators, mentors, cooperating classroom teachers, school leaders, peer teachers, students, etc.) Thus, the results of the teaching performance assessment are inevitably subject to serious human bias, and cannot be psychometrically compared across teacher preparation programs, teaching subjects, student populations, schools, and other demographic samples.

The recent research on teaching performance has shifted the focus from the standardization of teaching practice towards the complexity of teacher-student interactions in the co-constructed classroom learning environment (Gomez, Kyza & Mancevice, 2018). Consequently, in practice, classroom observation protocols with rubric-based rating scales are increasingly used and valued for teaching performance assessment to collect rich, real-time data on pedagogical practices. Such protocols are designed to quantify teaching performance on a number of carefully selected, observable

behavioral dimensions. Although this approach may be unable to capture the full complexity of teaching through a single statistical measure, it can provide a potential common reference framework to assess and compare teaching practices across raters, teaching education programs, classrooms, schools, and regions. Thus, it is critical to identify the factors that introduce construct-irrelevant sources of variance and to control for their adverse effects upon the validity and reliability of these classroom observation protocols. To that end, the current study applies a Rasch technique (i.e., many-facet Rasch modeling) to account for rater variability and other construct-irrelevant variances in analyzing data collected via the observational protocols.

Rater-Mediated Performance Assessment

In typical performance assessments, examinees are required to create a response or perform a task for a particular constructed-response item or task, rather than choose the correct answer from the test-given alternatives. Human raters are then trained and employed to analyze, interpret, and evaluate the examinees' responses/task performances to assign scores/ratings that reflect the true proficiency levels for individual examinees as intended by the assessment measures. Naturally, the process of such performance assessments mediated by human raters is complex and indirect, and very vulnerable to a variety of measurement errors, such as rater variability/effects and other construct-irrelevant variances (Eckes, 2009; Han, 2019).

Among others, the validity and reliability of the interpretation and use of ratings from rater-mediated performance assessments are primarily threatened by various rater effects (Eckes, 2015; Wind, 2019). To address the challenge, Standard 6.9 in the *Standards for Educational and Psychological Testing* (American Educational Research

Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME), 2014) recommend documenting and correcting “any systematic source of scoring errors” when using the scores/ratings from rater-mediated performance assessments (p. 118). Specifically, the *Standards* suggest that the rating quality in rater-mediated performance assessments should be evaluated and analyzed in a way that

monitors possible effects on scoring accuracy of variables such as scorer, task, time, or day of scoring, scoring trainer, scorer pairing, and so on, to inform appropriate corrective or preventive actions . . . Systematic scoring errors should be corrected, which may involve rescored responses previously scored, as well as correcting the source of the error. (AERA, APA, & NCME, 2014, p. 118)

Teaching Performance Assessment

Feiman-Nemser (2012) proposes the notion that teacher learning should be viewed as a continuum that extends across the professional lifespan. In line with this conceptualization, related empirical research literature review suggests that appropriate usage of teaching performance assessments for both pre-service and in-service teachers are promising in improving teacher learning and teacher effectiveness (Darling-Hammond, Newton, & Wei, 2013; Wei & Pecheone, 2010). Consequently, a consensus within the teaching profession has gradually formed to increase the implementation of performance-based assessments for both formative and summative evaluation of teaching effectiveness (Knight et al., 2014), because compared to other forms of knowledge-based teacher evaluation, teaching performance assessments focus on the proficiency/effectiveness of teachers in applying their subject-matter and/or pedagogical

content knowledge to the specific classroom contexts during the act of teaching (Santagata & Sandholtz, 2019).

Another unique advantage of teaching performance assessments lie in the fact that they are by design linked to established professional teaching standards that acknowledge the complexity of actual teaching practice and promote evidence-based effective teaching (Darling-Hammond, 2010; NBPTS, 2000; Sato, 2014). Empirical studies also support the positive relationships between scores on a teaching performance assessment (e.g., the Performance Assessment for California Teachers [PACT]) and student achievement gains (Darling-Hammond et al., 2013).

Despite the strong potentials of teaching performance assessments as discussed above, however, concerns remain among researchers and practitioners regarding the validity and reliability of using teaching performance assessments to capture teachers' true teaching quality. For example, disagreements have been reported in some studies between pre-service teachers' scores on a teaching performance assessment and teacher educators' judgments about their teaching qualifications (Sandholtz & Shea, 2012; Tellez, 2016). Furthermore, researchers also find that some teachers can deliberately "tailor" their classroom instruction (often in ways that contradict their everyday teaching practice) to cater to the specific standards of classroom observation protocols (Meuwissen, Choppin, Cloonan, & Shang-Butler, 2016). These issues warrant the urgent need for further research on the development, validation, and use of teaching performance assessments to promote teacher and teaching effectiveness.

Mathematics Classroom Observation Protocol for Practices

Developed by a team of math teacher educators at the University of Alabama

(Gleason, Zelkowski, Livers, Dantzler, & Khalilian, 2014), the Mathematics Classroom Observation Protocol for Practices (MCOP²) is a K-16 mathematics classroom instrument designed to measure the degree of alignment of the mathematics classroom with the Standards for Mathematical Practice from the Common Core State Standards in Mathematics (NGACBP & CCSSO, 2010); “Crossroads” and “Beyond Crossroads” from the American Mathematical Association of Two-Year Colleges (AMATYC 1995; AMATYC 2006); the Committee on the Undergraduate Program in Mathematics Curriculum Guide from the Mathematical Association of America (Barker et al., 2004); and the Process Standards of the National Council of Teachers of Mathematics (NCTM, 2000). The instrument contains 16 items originally intended to measure three primary constructs (student engagement, lesson content, and classroom discourse) as validated by a review of over 150 individuals self-identified as mathematics teacher educators from a mixture of mathematics departments and departments or colleges of education (Gleason et al., 2014). Each of the 16 items also contains a full description of the item with specific requirements for each rating level (Gleason & Cofer, 2014).

As a teaching performance assessment exclusively designed for math teachers, MCOP² focuses on both direct and dialogic instruction encompassing classroom interactions for the development of student math conceptual understanding, specifically examining teacher facilitation and student engagement (Watley, 2017; Zelkowski & Gleason, 2016; Zelkowski, Gleason, & Livers, 2017). Presently MCOP² has gone through a series of systematic validation studies under the classical test theory (CTT) framework (Gleason, Livers & Zelkowski, 2017), including content validity, interrater reliability, and structure analysis. However, the observation protocol has never been

investigated under the Rasch framework for dimensionality, item-level analyses, rater effects control, and ratee and rater ability level calibration.

Many-faceted Rasch Model (MFRM)

The Many-Facet Rasch model (MFRM) is appropriate for analysis of multiple variables or facets at the same time possibly influencing assessment results. MFRMs belong to the family of Rasch models such as rating-scale models (RSM), partial credit models (PCM), linear logistic test models (LLTM) (Kubinger, 2009), the mixed Rasch model (Baghaei & Carstensen, 2013), and others. MFRM approach has been used extensively in the areas such as language testing, educational and psychological measurement, and health sciences. A typical assessment scenario where MFRM can be applied may involve a four-category rating scale and raters to evaluate performance of a test taker.

The scenario described above defines a three-facet situation with test takers, tasks, and raters as the three facets. This three-facet situation can be expressed as follows:

$$\ln\left[\frac{p_{pljk}}{p_{pljk-1}}\right] = \theta_p - \delta_l - \alpha_j - \tau_k, \quad (1)$$

where p_{pljk} is the probability of test taker p receiving a rating of k from rater j on task l ;
 p_{pljk-1} is the probability of test taker p receiving a rating of $k-1$ from rater j on task l ;
 θ_p is the proficiency of test taker p ; δ_l is the difficulty of task l ; α_j is the severity of rater j ,
and τ_k is the difficulty of receiving a rating of k relative to $k-1$ (threshold parameter). In this three-facet rating scale model (Linacre & Wright, 2002), different facets such as test takers, tasks, and raters can be regarded as independent variables (IVs) that affect the log odds as dependent variables (DV).

Unlike other IRT models where item discrimination is estimated freely on an item-by-item basis, the Rasch model scales the discrimination parameters for all items equally to 1. This constraint allows the Rasch models to place both item difficulty and person ability on the same equal-interval log-odds (logit) scale. MFRM extends the typical Rasch model (that involves test-takers and items as its only two facets) by permitting the addition of facets, or sources of variability. MFRM can model all these facets jointly and analyze the pattern of examinee responses, rater scores, item functioning in the form of fit statistics that help detect aberrant behavior on any of the facets (Sims et al., 2020).

MFRM has been readily employed to control for rater effects (Eckes, 2015; Engelhard, 1992), and is widely accepted as a robust statistical mechanism that adjusts for rater effects and identifies outlying raters or examinees, resulting in a modified “fair average score” that represents a more accurate assessment. Moreover, MFRM can also provide scale diagnostic data regarding rater use of the scale in terms of both consistency and consensus (Knoch & Chappelle, 2018).

Another advantage of using MFRM is that they are robust against missing data (often resulting from not-fully crossed research designs where not all raters rate all examinees) as they are only evaluated for observed data points. There is no requirement to impute for unobserved data (Linacre, 1993, 1995, 2001).

The Problem Defined

Rater-mediated classroom observation protocols are increasingly being used for teaching performance assessments, which makes identifying and controlling for various rater effects a central issue to ensure the rating quality. Researchers employ two common

approaches to evaluate the rating quality in rater-mediated assessment: the number-correct score approach under the classical test theory (CTT) framework, and the latent trait modeling approach under the item response theory (IRT)/Rasch framework.

For the CTT number-correct score approach, different indices of interrater reliability (IRR) are computed and compared, for instance via absolute interrater agreement or Cohen's kappa (Cohen, 1960). However, empirical research literature shows ample evidence that the CTT approach may result in unintended interpretations of a scoring rubric (Eckes, 2008), biased ratings due to power dynamics among raters (Hoyt & Kerns, 1999), or the need for costly and time-consuming training programs that often fail to produce a high degree of agreement (Barrett, 2001).

Furthermore, Chen and his colleagues (2020) point out two significant theoretical flaws with this IRR approach. First, this approach tends to ignore the unique impacts on rating quality resulting from the complex interactions between raters' expertise, observational protocol rubrics, and classroom environments. Second, this approach attributes rater drift (rater scores begin to vary over time or across occasions) to rater training (e.g., increased familiarity with the rubrics after practice) or consensus-making efforts (e.g., raters who originally assign different scores discuss with each other to reach an agreement for a certain examinee on a certain item/task), rather than to important differences in the rating data itself (Hoskens & Wilson, 2001).

A series of validation studies under the CTT framework (Gleason, Livers & Zelkowski, 2017), including content validity, interrater reliability, and structure analysis, have been completed for the 16-item Mathematics Classroom Observation Protocol for Practices (MCOP²). However, the MCOP² data have never been investigated under the

Rasch framework. Due to the methodological limitations of the CTT approach for rater-mediated assessments as discussed above, it is imperative to examine the MCOP² validity and reliability using MFRM. Thus, the study aims to evaluate the MCOP² rating quality from another psychometric perspective, which employs the MFRM modeling technique to implement dimensionality analysis, item-level analysis, rater effects control, and rater and rater ability level calibration.

Purpose of the Study

Based on the discussions above, the purpose of this study is to evaluate a math classroom observation protocol (MCOP²) under a Rasch measurement framework for calibrating rater assessment of math teachers' instructional performance, which combines the Rasch sub-dimensional modeling for internal structure validation and the Many-Facet Rasch Model (MFRM) for rater effects control.

Research Questions

There is a total of seven empirical research questions (ERQs) based on the above-mentioned purpose of the study. This section elaborates on the relationships between the research purpose and the related empirical questions.

Research Questions 1-7

The research purpose (i.e., to evaluate the validity and reliability of the MCOP² classroom observation protocol under the MFRM framework) guides the following seven research questions.

1. To what extent do the observed rating data obtained from the MCOP² instrument fit the MFRM modeling? This question is evaluated by testing the MFRM model assumptions, including local independence, unidimensionality,

overall model fit, rater fit, and item fit.

2. To what extent does the MCOP² observation protocol separate observed teachers into distinct levels of proficiency? Such a separation is evaluated by examining the examinee facet in the MFRM analysis.
3. To what extent do raters differ in terms of the relative severity with which they rate observed teachers? This question is evaluated by examining the rater facet in the MFRM analysis.
4. To what extent do raters consistently rate the teaching performance of observed teachers? This question is evaluated by investigating possible interactions between raters and observed teachers using the MFRM analysis.
5. To what extent do raters consistently rate the teaching performance of observed teachers across the MCOP² items? This question is evaluated by examining investigating possible interactions between raters and the MCOP² items using the MFRM analysis.
6. To what extent can the score levels of the MCOP² items be distinguished, without certain score levels being either underused or overused? This question is evaluated by examining both the graphic indicators (i.e., Item Characteristic Curves, and Item Information Functions) and the statistical indicators (i.e., item category ordering for individual raters, and rater fit indices).
7. To what extent are the rater behaviors associated with the professional background characteristics (i.e., in-service vs. pre-service teachers, schools, and teaching grade levels) of the observed teachers? This question is

evaluated by examining possible interactions between raters and the facets indicating observed teachers' professional background in the MFRM analysis.

Significance of the Study

In this section, scholarly significance is discussed in the following two aspects: the first aspect focuses on the systematic MFRM-based validation of the MCOP² classroom observation protocol for math teachers' teaching performance, whereas the second aspect highlight the important implications of using the MFRM-calibrated MCOP² ratings for promoting efficient and effective teacher peer- and self-assessments.

MFRM-Based Validation Study

Various types of psychometric techniques have been applied to identify and control for the sources of rater variability in rater-mediated performance assessments. Among those, Rasch models (especially MFRM) have gained wide recognition for their methodological robustness and easy adaptability for a rich variety of empirical research contexts. For instance, studies on high-stakes assessments (e.g., language assessments) have focused on using MFRM analyses to investigate the reliability of rater judgments, rater biases, and the relationship between rater bias and rater training (McNamara & Knoch, 2012). Applications of MFRM research in other research fields also include outpatient performance assessment (Kramer, Kielhofner, Lee, Ashpole & Castle, 2009), creative writing assessment (Barbot, Tan, Randi, Santa-Donato & Grigorenko, 2012), and behavior analysis (Mannarini, 2009).

However, MFRM-based analyses have rarely found their way to rater-mediated teaching performance assessments, such as teacher observation protocols; and the existing few MFRM-related studies in this field only focus on examining and handling

rater effects for a specific sample of rating data (Chen, Yim, Kogen, Stieff, & Superfine, 2020). Thus, the current study is one of the first research efforts to adopt the MFRM framework to systematically examine the validity and reliability of a rater-mediated classroom observation protocol for math teachers: the Mathematics Classroom Observation Protocol for Practices (MCOP²).

Identifying and controlling for rater effects is essential for improving the reliability of rating quality obtained from observational protocols. Rater effects may take different forms and can be hidden by or confused for other parts of an assessment system. By examining various facets (rater, item, examinee, and others) as well the possible interactions among them, and the functioning of the rating scale, the MFRM measurement approach is used in this study to verify aspects of the MCOP² assessment system that function as intended, as well as to detect the aspects that are potentially problematic.

As an alternative to the Cohen's kappa method under the CTT framework, the MFRM is applied here to establish interrater reliability and to account for rater variability at once. Such a MFRM approach provides a robust psychometric framework to assess and compare teaching based on observational ratings of teacher practices (Johnson, Zheng, Crawford, & Moylan, 2019; Jones & Bergin, 2019), which can be used to compliment research based on other data sources to present a fuller and more accurate characterization of teacher practice (Chen et al., 2020). In sum, the information gained from the findings of this study would hopefully help improve the psychometric properties of the MCOP² observational protocol.

MFRM-Calibrated Teacher Peer- and Self-Assessments

According to Sluijsmans and Prins (2006), teacher peer assessment can be utilized as an effective tool for promoting important teacher learning for four reasons. First, peer assessments can motivate peer learning and peer communication among teachers that help to form a learning community (Johnson, Johnson, Holubec, & Holubec, 1994; Shachar & Sharan, 1994; Verloop & Wubbels, 2000). Second, peer assessment fosters teachers' critical reflection and analysis and the development of reflection skills necessary for making reliable judgments about peers' work (Birenbaum, 1996; Sambell & McDowell, 1998). Third, teachers can readily transfer the skills they have learned from peer assessments to their own classroom settings and improve the ability to design assessments and make critical judgements about their student performances. Lastly, teachers are expected to rely heavily on their peers' judgments to estimate the effectiveness of their performances in the school setting (Brown, Rust, & Gibbs, 1994). Thus, being able to interpret the work of colleagues and peers is a prerequisite for professional development and for improving teachers' functioning in the profession (Verloop & Wubbels, 2000).

Sluijsmans and Prins (2006) proceeds to underline performance assessment as the foundation for peer assessment tasks, where judgments are made about the level of achievement attained by comparing teacher performance to predetermined standards. All peer teachers attain the standards, whereby they are expected to make their best judgments about the performance of their peers and negotiate about appropriate criteria for these performances (Boud, 1995; Orsmond, Merry, & Reiling, 1996; 1997; 2000). Similarly, Stiggins describes the unique role of peer performance assessments in promoting professional learning in teacher education as: "Once students (teachers)

internalise performance criteria and see how those criteria come into play in their own and each other's performance, students (teachers) often become better performers" (1991, p. 38).

However, Cabello and Topping (2020) point out several prominent obstacles in effectively implementing peer assessments among teachers, including costs of time and resources for the organizers and participants, teachers' initial reluctance and anxiety to participate in peer assessments, and most importantly, the fact that validity, reliability and fairness of peer assessments may be threatened by potential effects of teachers' social considerations of friendships, popularity, enmity, and perception of criticism, as well as the tendency for the less socially risky option of assigning average scores on peer assessments.

In addition to peer assessment, performance assessments also take the form of self-assessment (i.e., a formative assessment process in which students evaluate their own studies in accordance with predetermined criteria and goals), enabling learners to take more responsibility for their own learning and actively participate in the process of "assessment for learning" (Ballantyne, Hughes & Mylonas, 2002; Matsuno, 2009). Self-assessment familiarizes learners with well-defined performance criteria against which they evaluate their own learning with clear focus and motivate learning from their mistakes. Just as Puhl (1997) argues, the biggest contribution of self-assessment to any learning and teaching process can be understood as "one of the important skills that should be developed for students to take with them when they leave school and then use them for lifelong learning" (p. 28).

Not unlike peer assessment, however, the validity and reliability of self-

assessment is also subject to the influences of rater characteristics and rating contexts, despite its positive effects on learning and on metacognitive knowledge levels (Topping, 2009; Yurdabakan & Oğlun, 2011). Moreover, empirical research shows that low to medium correlations are found between self, peer, and teacher assessments and these ratings are significantly different from each other, where self-assessments are the most lenient while peer assessments are the most severe (Aryadoust, 2015; Farrokhi, Esfandiari & Dalili, 2011; Farrokhi, Esfandiari & Schaefer 2012; Karakaya, 2015).

To tackle the above-mentioned challenges in utilizing self- and peer assessments as a valid and reliability learning assessment tool, the Many-facet Rasch Model (MFRM) has been recommended to determine the reliability of peer and self-assessment scores and mitigate the limitations of classical CTT approaches (Baird, Hayes, Johnson, Johnson & Lamprianou, 2013; Kim, Park & Kang, 2012; Linacre, 1996). The major methodological benefits of MFRM include (a) calibrating raw ratings for performance assessments affected by rater behavior (Mulqueen, Baker & Dismukes, 2000), (b) identifying the interactions between different sources of error (Haiyang, 2010), (c) accounting for more than one source of error simultaneously and producing higher ability estimates for validity (Ilhan, 2016), and (d) providing diagnostic information at the individual level rather than at the group level for raters and ratees (Barkaoui, 2008).

MFRM has been adopted in limited studies to investigate self- and/or peer assessments from various perspectives (Erman Aslanoglu, Karakaya, & Sata, 2020): in some research, the MFRM approach is compared to other theoretical frameworks (Guler, 2008; Macmillan, 2000; Sudweeks, Reeve & Bradshaw, 2005); some researchers use MFRM to examine the ratees' proficiency on the construct/ability assessed as well as the

severity/leniency of the raters at the individual level (Akin & Basturk, 2012; Basturk, 2008; Engelhard & Stone, 1998; McNamara & Adams, 1991; Weigle, 1998; Weigle, 1999); some studies focus on investigating rater bias and factors affecting it (Aryadaust, 2015; Cetin & Ilhan 2017, Farrokhi & Esfandiari, 2011; Saito, 2008; Schaefer, 2008); and some others aim to examine and compare rater sources (Farrokhi, Esfandiari, & Dalili, 2011). Recently, Erman Aslanoglu, Karakaya, and Sata (2020) conducted a MFRM analysis to understand the role of teacher candidates' participation in the assessment process (self- and peer assessment) in improving their scoring behaviors in self- and peer assessments. They found a significant difference in teacher candidates' rating behavior related to the rater types (self vs. peer), in that the raters appear more lenient in self-assessments rather than peer assessments. In addition, raters tend to be more biased when they rate individual performances rather than group performances.

Along this line of MFRM-based research, this research seeks to uniquely contribute to the literature by providing a MFRM-based, latent construct framework to systematically validate a classroom observation protocol for math teaching performance (i.e., MCOP²), and to anchor the parameters of the essential facets involved (i.e., item, raters, and ratees) as the basis for rating calibration in the future, wider application of MCOP² in self- and peer assessments among K-16 math teachers.

Limitations of the Study

All research has limitations, and this study is no exception. Five major limitations are noted in the current investigation.

First, the study is limited in terms of its generalizability to the total population of K-16 math teachers. Only two samples of the MCOP² data are used for the purposes of

this study. To improve the generalizability of the findings, the data from additional test administrations should be considered for comparison of the results. Furthermore, the data from the two samples used for this study are collected from the K-16 math classrooms in only two states: Alabama and Kentucky. The same analysis should be run for rating data collected from other states or regions for comparison of the results. Additionally, the data from the two samples in the current study heavily focus on the observation and assessment of pre-service math teachers, which warrants the need for future research to increase the inclusion of in-service math teachers in their study samples.

A second potential limitation of the study is related to the replicability of the MFRM-based validation analysis for other rater-mediated teaching performance assessments. Although the MFRM-based approach appears more methodologically robust compared to traditional performance assessment methods (e.g., generalizability theory, interrater reliability indices), its application also presents various logistical challenges. Among others, MFRM analyses require up-front planning/rating design, relatively large sample sizes (e.g., at least 30-50 persons in a sample for pilot/exploratory MFRM analysis according to Linacre, 1994), and specialized knowledge of measurement and psychometric principles. In this sense, the current study may hold reference value for measurement and assessment professionals to determine whether the benefits of MFRM overwhelm the additional challenges associated with the technique.

Third, problems with the data quality may affect the findings of the current study. These problems may include the way the data were collected, recorded, and stored, whether any data were missing when the initial rating was done, the security of the data, and possible errors in rating. All these issues would normally be regarded as limitations

of a study, as the data quality issues could affect parameter estimation, such as item, rater, and examinee parameters in the MFRM analysis.

Fourth, the sample size of 159 observed math teachers is relatively small compared to other studies applying the MFRM model; and the sample is not randomly selected, thus possibly limiting the external validity of the study.

Finally, an additional limitation of the study concerns the secondary nature of the MCOP² rating data used for the MFRM analyses. Typically, researchers of any secondary data sources may experience difficulty in fully understanding all the data subtleties or problems encountered in the original data collection, recording, and storage process. This data knowledge can be instrumental in the accurate interpretation of MFRM analysis findings.

CHAPTER II

REVIEW OF THE LITERATURE

Introduction

Chapter I described the recent major trends in the research literature related to rater-mediated teaching performance assessments. Further, the existing gaps in the related research literature were identified regarding the understanding and handling of construct-irrelevant variances that threaten the validity and reliability of the rater-mediated assessments. In line with these identified needs warranting the current study, a brief introduction was also provided about the theoretical foundation and empirical application of the MFRM based analysis. To address the research needs, this proposed study was designed to further examine the MCOP² rating quality under the MFRM framework and its applicability in self- and peer assessments.

Guided by the research purpose, seven empirical research questions were specified to determine to what extent the construct-irrelevant variance (especially rater effects) can be detected and controlled for the MCOP² performance assessment based on

the MFRM analysis.

The study was expected to make unique contributions to the knowledge base of rater-mediated teaching performance assessment by illustrating the application of the psychometric technique (i.e., MFRM) to improve the psychometric properties of typical classroom observation protocols designed for assessing teaching practices. However, five major limitations of the study were also acknowledged as the issue of generalizability, data quality, sample size, modeling restrictions, and limitations associated with the researcher's choice of using secondary data sources.

In this chapter, a review of the literature was conducted using EBSCOhost, ProQuest, and Web of Science accessed through the library at University of Kentucky (UK). Literature was reviewed and reported below on the existing MCOP² validation and empirical studies, and the MFRM analysis as a calibrated framework for rater-mediated teaching performance assessments. The key words *Rater-Mediated Performance Assessment* and *MFRM* were input for the literature search when using the above-mentioned databases. The resources listed in the search findings were then further filtered according to their degree of relevance to the current study.

The first part of Chapter II, **Summary of Earlier MCOP² Validation and Empirical Studies**, covers the considerations, standards, and procedures in the process of developing the MCOP² classroom observation protocol, existing MCOP² validation and empirical studies, and limitations in the MCOP² research literature. The second section in the chapter, **A Calibrated Framework for Rater-Mediated Assessment (MFRM)**, discusses theoretical framework and empirical application of the MFRM analysis for rater-mediated performance assessment including dimensionality analysis, rater effects

and drift, interrater variability, and multi-facet calibration controlling for rater effects.

Summary of Earlier MCOP² Validation and Empirical Studies

This section covers detailed discussion on **Development of MCOP², Existing MCOP² Validation Studies, Existing MCOP² Empirical Studies, and Limitations in the Existing MCOP² Studies.** The end of the section further highlights the need for conducting the proposed dissertation study.

Development of MCOP²

The Mathematics Classroom Observation Protocol for Practices (MCOP²) was initially developed in 2013 by a team of mathematics teacher educators and researchers as a classroom observation instrument. It was primarily designed to measure the degree to which a K-16 mathematics classroom aligns with the standards put forth by national mathematics organizations. These standards include the Standards for Mathematical Practice from the Common Core State Standards in Mathematics (NGACBP & CCSSO, 2010), “Crossroads” and “Beyond Crossroads” from the American Mathematical Association of Two-Year Colleges (AMATYC 1995; AMATYC 2006), the Committee on the Undergraduate Program in Mathematics Curriculum Guide from the Mathematical Association of America (Barker et al., 2004), and the Process Standards of the National Council of Teachers of Mathematics (NCTM, 2000) (Gleason & Cofer, 2014).

The research team initially created 18 items focused on the interactions of the mathematics classroom understood to promote conceptual understanding. In this process, some items were adapted from other instruments and others were developed to incorporate the framework and language of the Standards for Mathematical Practices.

The rubrics for the items were developed through an iterative process involving watching classroom videos as a group and determining specific criteria for each level in the rubric, along with referencing related literature for specific interactions. This process resulted in the development of a user guide with detailed descriptors and rubrics (Gleason, Livers, & Zelkowski, 2015), along with an abridged user guide containing only the rubrics.

The original 18 items in the MCOP² instrument were intended to measure three primary constructs (student engagement, lesson content, and classroom discourse) based on the theoretical framework of the Common Core State Standards in Mathematics (CCSSM). The researchers claimed that these three constructs were validated by a review of over 150 individuals self-identified as mathematics teacher educators from mathematics departments or departments or colleges of education (Gleason, Zelkowski, Livers, Dantzler, & Khalilian, 2014). However, in their initial validation study (Gleason & Cofer, 2014), the results of the exploratory factor analysis pointed to a 2-factor model, instead of a 3-construct framework as originally planned. Thus, the two constructs (i.e., student engagement and classroom discourse) were merged as one sub-dimension in parallel with the “lesson content” sub-dimension.

The finalization of the MCOP² development involved a multistage iterative process over three years based on the standards for scale development (AERA, APA, & NCME, 2014; Bell et al., 2012; DeVellis, 2011). To determine whether it was essential to retain or remove certain items in the instrument, feedback on content validity of the items were gleaned from three rounds of expert panels. The expert panels were comprised of a convenience sample of the members of the Association of Mathematics Teacher Educators (AMTE), who were invited by the MCOP² developers to participate in

an initial online survey asking for feedback on the initial pool of 18 items and their perceived usefulness in measuring various aspects of the mathematics classroom. The initial survey asked the participants to rank the usefulness of the item to measure mathematics instruction on a three-point scale (essential, not essential but useful, and not necessary) and provide comments about the items.

The 164 professionals in the initial expert panel completed the online survey. Based on their responses and the comments, 16 out of the original 18 items were retained with minor edits in the wording of the item, with the largest such change involving changing “Students engaged in flexible alternative modes of investigation/problem solving” to “Students engaged in exploration/investigation/problem solving.” One of the items removed was “Students explored prior to formal presentation.” for possible biases of the item toward specific teaching methods and ambiguity about the meaning of the terms in the item. The second item removed was “The lesson promoted connections across the discipline of mathematics,” for its ambiguity as to what constituted another area of mathematics.

The second and third round expert panels were asked to provide further content validity feedback regarding how the 16 items should be related to the four theoretical factors (i.e., Lesson Design, Lesson Implementation, Student Engagement with the Content, and Student Engagement with Peers), the details of which are presented in the next section, **Existing MCOP² Validation Studies**.

Theoretically, each of the final 16 MCOP² items was created to correlate with one of the *Standards for Mathematical Practice*. For example, Item 8 on the protocol is “The lesson provided opportunities to examine elements of abstraction (symbolic notation,

patterns, generalizations, conjectures, etc.)” It matched the second Standard for Mathematical Practice that instructors should teach their students:

“CCSS.Math.Practice.MP2: Reason abstractly and quantitatively” (NGACBP & CCSSO, 2010). Furthermore, as cited in Gleason & Cofer (2014), Item 9 was also conceptually connected to Part 1 of the CUPM Curriculum Guide which recommends (Barker, et al., 2004):

“For instance, one reason students encounter difficulty in applying mathematics to problems in other disciplines is that they have trouble identify appropriate mathematical procedures when problems are expressed with different symbols than those used in the mathematics classroom....instructors can go beyond conventional x, y notation to use a larger collection of symbols for both constants and variables.” (p. 20)

Whereas operationally, each of the 16 MCOP² items contains a full description of the item with specific requirements for each rating level on a four-category rating scale ranging from 0 to 3. Again, take Item 8 as an example: to give the highest rating of 3, raters must observe “The students have a sufficient amount of time and opportunity to look for and make use of mathematical structure or patterns.,” while the lowest rating of 0 would be justified if raters believe “Students are given no opportunities to explore or understand the mathematical structure of a situation.” in the lesson (Gleason, Livers, & Zelkowski, 2017).

The MCOP² developers and researchers noted four major benefits of the MCOP² instrument compared to other preexisting classroom observation protocols related to teaching mathematics (Gleason & Cofer, 2014). First, unlike the MCOP², many of the

other protocols were not created specifically for mathematics classrooms, but instead are intended for dual use in both mathematics and science classrooms (Wainwright, Flick, & Morrell, 2003; Walkington et al., 2012). Second, some of the other preexisting protocols were not designed in line with the most recent national standards for mathematics classrooms. Third, the rating rubrics of the MCOP² instrument were written in a clear, concise, and accessible way for peer-to-peer reviews and assessments, and thus it was not necessary for the MCOP² raters to receive any special training. Finally, compared to the generic, lengthy, and subjective preexisting protocols that often contained around 50 items, the finalized MCOP² instrument only had 16 items and had gone through a three-year, multi-stage process of robust validation of its psychometric properties.

Existing MCOP² Validation Studies

A pilot study (Gleason & Cofer, 2014) was conducted at a large southern university to determine if the data collected aligned with the theoretical constructs verified by the expert survey. Thirty-six math classrooms taught by 28 different instructors were observed throughout a semester. The backgrounds of instructors (e.g., graduate teaching assistants, or tenured full professors) and math classrooms (e.g., college algebra, or upper division mathematics) both varied across a wide range.

Based on the expert panel feedback, the 17-item MCOP² that was used for the pilot study was initially designed to measure three constructs (i.e., Student Engagement, Lesson Design and Implementation, and Class Culture and Discourse). Items 1-5 were supposed to measure Student Engagement, Items 6-11 were meant for Lesson Content, and Items 12-17 were classified under Classroom Culture and Discourse (Gleason & Cofer, 2014).

However, based on the results of the exploratory factor analysis (EFA), the original 3-factor assumption was re-examined after a low eigenvalue loading on the third factor. Consequently, the two factors, Student Engagement and Classroom Culture and Discourse, were combined to create a new construct: Student Engagement and Classroom Discourse. The resulting 2-Factor model explained over 50% of the total variance in the pilot study data.

Cronbach's alpha was also calculated for the 17-item protocol as a whole, and for the two factors separately. The overall Cronbach's alpha value for the protocol was .898; whereas the Cronbach's alpha values for the sub-scales of Lesson Content and Student Engagement and Classroom Discourse were calculated as .779 and .907, respectively. Thus, Gleason and Cofer (2014) concluded, “the internal reliabilities are high enough for both sub-scales and the entire instrument to be used to measure at the group level, either multiple observations of a single classroom or single observations of multiple classrooms” (p. 99).

This pilot study marked the initial phase of a multi-stage, reiterative validation process for the MCOP² instrument over a span of three years. First, content validity of the MCOP² items were verified with 164 experts in mathematics teaching education. These experts were invited to participate in three rounds of online surveys. The first survey provided feedback on the initial 18 MCOP² items and their usefulness in measuring various aspects of the teaching practices in a mathematics classroom (Gleason, Livers, & Zelkowski, 2017). Over 94% of the experts rated the items as either “essential” or “not essential, but useful,” rather than “not useful” for measuring the mathematics teaching practices. Based on the first-round expert feedback, two of the original 18 items

were removed from the MCOP² instrument due to ambiguity in wording or definition of special terms. This was followed by a second survey with 26 of the initial 164 experts that agreed to provide additional information. This survey provided the experts with detailed description of each item, the associated theoretical constructs, and the intended purpose of the MCOP². With the information gained from the experts, the structure of the MCOP² instrument was revised.

Gleason, Livers, and Zelkowski (2017) also calculated the inter-rater reliability for using the MCOP² instrument for math teaching performance assessment. Five raters were chosen from various educational and professional backgrounds. Among them, two had doctorates in mathematics education; one rater had a doctorate in mathematics and had been heavily involved in mathematics education research; one rater was a mathematics specialist that worked with secondary teachers and had taught at both the secondary and introductory college level. The fifth rater was a graduate student in mathematics with minimal background in education other than teaching some introductory college math classes.

Five different classroom videos were rated by all five raters. Each rater independently observed and rated the five video-recorded math classrooms without receiving any formal rater training. The sample of the five videotaped math teaching practice was chosen from each of K-2, 3-5, 6-8, 9-12, and undergraduate level math classrooms. Gleason, Livers, and Zelkowski (2017) used the sub-scale score to calculate the intra-class correlation (ICC) among the five raters and reported acceptable inter-rater reliability.

As a result of this rigorous validation process, the finalized MCOP² protocol

contains 16 items measuring two primary constructs (i.e., Teacher Facilitation and Student Engagement) in an interaction-based, co-constructed math classroom environment. However, before the protocol can be used for undergraduate-level math classroom with confidence, the validity and reliability of using the MCOP² protocol needed to be further evaluated in other mathematics classrooms at multiple higher education institutions. Both liberal arts schools and other types of research universities should also be included in the validation study samples to increase their representativeness to reflect the characteristics of the overall population.

To that end, Watley (2017) tested the validity and reliability of the MCOP² protocol with a different study sample that included 110 college mathematics classrooms at the undergraduate level, representing a wide variety of college and university classrooms. In her study, many of the sample classrooms were selected from three large southern doctorate-granting universities with enrollments of approximately 18,000 to 35,000. Other sample classrooms came from eight southern master's and baccalaureate colleges and universities with enrollments between 1,100 to 15,000 students. All these high-education institutions had student populations representing a diversified demographic, ethnic, and cultural backgrounds.

With these selections of institutions, the researcher was able to obtain 46 observations at doctorate-granting universities, 21 observations at master's universities, and 43 observations at baccalaureate college and universities (See Table 5) in this study to overcome any potential bias due to the convenience sampling (Watley, 2017). In this study sample, lower-level undergraduate mathematics lessons were taught in 89 classrooms compared to the 21 upper level classrooms. Seventy-two mathematics faculty

members agreed to participate in this study. Since some instructors teach two or more completely different courses, a total of 110 observations were conducted in the Spring 2016, Fall 2016, and Spring 2017 semester.

A confirmatory factor analysis (CFA) was conducted on the rating data drawn from the new sample of undergraduate math classrooms, and the findings showed that the 16-item MCOP² data fit a two-factor model: Student Engagement and Teacher Facilitation. Items 1-5 and items 12-15 loaded on Student Engagement, while items 4, 6-11, 13, and 16 loaded on Teacher Facilitation. The goodness of fit indices for the MCOP² revealed an acceptable fit for three indices ($\chi^2/df=1.19$, SRMR=.08, and CFI=.90), and a poor fit for the other indices (RMSEA=.09 and GFI=.81).

In terms of internal consistency, the Cronbach's alpha values for the two subscales of the MCOP² were .888 for Student Engagement and .812 for Teacher Facilitation, respectively. Both subscales therefore fell within the satisfactory range for basic research and were near the acceptable levels for individual measurement (Nunnally, 1978, p. 245-246).

Additionally, simple linear regression analyses were also conducted to estimate the relationships between the constructs measured by the Mathematics Classroom Observation Protocol for Practices (MCOP²) and the abbreviated Reformed Teaching Observation Protocol (aRTOP). The findings highlighted that Inquiry Orientation positively predicts higher ratings in both Teacher Facilitation and Student Engagement measured by the MCOP², while better Teacher Facilitation also positively predicts more desirable Student Engagement.

Existing MCOP² Empirical Studies

Unlike the relatively solid literature on the MCOP² validation studies, empirical research has been scarce concerning direct applications of the protocol in high-stake teaching performance assessments, probably due to the administrative and programmatic cost/complexities involved in pushing for the change in the evaluation of teacher and teaching quality evaluation. Thus, the existing body of the MCOP² empirical research seems limited to the fields of mathematical instruction reforms (e.g., active learning) and teacher education/preparation program evaluation studies.

Zelkowski and Gleason (2016) conducted a two-year, mixed method study to investigate the value of using the MCOP² in secondary mathematics teacher preparation programs (SEMA-TPP) to (a) compliment the generalist observation forms currently adopted by teacher educators and local cooperating teachers to assess student teachers' instructional quality, and (b) facilitate preservice teachers' growth and self-learning, especially in their planning of formal observation lessons. Over the two-year study, the researchers examined 59 SEMA-TPP candidates in middle and upper grades mathematics classrooms using both observation forms.

Their findings showed a very strong correlation between scores on the two forms, indicating an accurately scored MCOP² rubric aligned very well to A, B, C, D, F letter grades on the generalist observation forms, even when used by raters from very different backgrounds (e.g., teacher, supervisor, or university faculty). Furthermore, the researchers also found that eighty-three percent (n=49 of 59) of preservice teachers preferred the MCOP² scoring because they knew what to improve on for the next observation; whereas the generalist form was not specific enough without written feedback or discussion. Finally, regarding the MCOP² impact on student teachers' lesson

planning quality, it was found that 26 of 31 (84%) method student lesson plans from two years of MCOP² were scored higher than in the two years prior to the use of the MCOP².

To measure students' perceptions of active learning opportunities (such as forming hypotheses, creating mathematical models and discussing their ideas with others), Bowers and Smith (2016) transformed the MCOP² from a teacher observation tool to a student survey to understand what students thought about the active learning labs and measure the extent to which these labs engaged students in active learning practices. The researchers cut the 16 MCOP² items in half, to ask the eight questions most relevant to student experiences, but then asked each question twice: once about students' experiences in lectures, and once for their experiences in the labs (Bowers & Smith, 2016).

The results of their confirmatory factor analysis showed that the MCOP² student survey had the same 2-factor structure as the original observation tool (Teacher Facilitation and Student Engagement). Moreover, analysis of the MCOP² survey data also suggested that the transformed MCOP² student survey was instrumental in identifying student-perceived specific value-added aspects of active learning that the labs could offer to augment lecture (Bowers & Smith, 2016).

Another MCOP²-related empirical research was conducted by Garrett and her colleagues as a case study to understand the learning and adaptation that could occur when faculty incorporated active learning into existing course structures (Garrett, Guest, Tameru, & Karatas, 2016). In this study, the MCOP² protocol was only used as a general framework to document the key events/activities around the subject of the case study related to the activity learning elements in math classroom instruction.

Similarly, in a mixed method study, Livers et al. (2020) attempted to use the MCOP² protocol to measure the pre-post changes in Teacher Facilitation and Student Engagement after implementing a coaching cycle approach within a larger professional development design that focused on infusing high quality mathematics tasks and differentiation within inclusive elementary mathematics classrooms. Their findings indicated that classroom observations shifted to a more student-centered practice with an increase in co-teaching collaborations and behaviors, in support of the benefit of a coaching component to facilitate and sustain teacher growth and professional development.

Limitations in the MCOP² Research

Three major limitations are noted in the current MCOP²-related validation and empirical research.

The first major limitation lies in the fact that the existing MCOP²-related validation studies exclusively adopt the test score tradition or number-correct approach (Engelhard, Wang, & Wind, 2018) under the classical test theory (CTT) framework to adjust for the rater bias in the rating data, such as rater agreement indices, intraclass correlations, kappa coefficients, and generalizability coefficients (Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Johnson, Penny & Gordon, 2008; von Eye, & Von Eye, 2005). Employing this test score tradition approach, researchers only need to report the percentage of exact and adjacent category usage for operational raters as shown in the raw rating/score distributions. However, this method is based upon a fundamental assumption that the observed ratings can be treated as having categories with equal width to be modeled as equal intervals by using sum scores, which is *never* the case for most of

the rating scales created and used for rubric-based, rater-mediated performance assessments. For instance, on a 3-category rating scale used in a rater-mediated teaching performance assessment, student engagement activities in a classroom that receive a rating of 3 may clearly display better student on-task behaviors, participation, and problem-solving interests than those classrooms rated with a 1 or 2 in terms of student engagement. However, from the psychometric measurement perspective, there is no evidence in support of equal distances in their abilities to engage students between the classrooms that are given the ratings of 1, 2 or 3. Therefore, acceptable interrater reliability indices such as the ICC or kappa coefficients calculated under the CTT framework do not provide sufficient grounds to validate the usage of a rater-mediated performance assessment (e.g., classroom observation protocols) free of any rater bias/effects.

As an alternative method to account for and manage the rater variability, measurement models based on the scaling tradition (Engelhard, 2013) parameterize the structure of rating categories with category coefficients (i.e., thresholds). Thresholds that define rating categories do not need to have equal width (Engelhard & Wind, 2013). As an item response theory (IRT) model specifically designed for rater-mediated assessments (Eckes, 2015), the Many-Facet Rasch model (MFRM, Wright & Linacre, 1989) is a generalized form of the Rasch model that not only adds a rater parameter particularly for rater effects control, but also is capable of accommodating other construct-irrelevant variances as additional facets in the model for parameter estimation and calibration. The MFRM model has been widely used in the detection and management of rater effects. Additionally, several other rater models have also been

proposed, such as the hierarchical rater model (Casabianca, Junker, & Patz, 2016; Engelhard, Wang, & Wind, 2018).

A second limitation in the current MCOP² research is related to the insufficient theoretical, statistical, and empirical support to justify the specific choices that the MCOP² developers made in terms of the rating scale structure (e.g., 4-category scales), definition of ratings (e.g., level of agreement vs frequency etc.), observer training (e.g., 2-day training, online training, no formal training etc.), number of observers needed to code a lesson (e.g., 1 or 2 observers), and number of observations needed in order to determine quality of mathematics instruction (e.g., 1 to 6 observation per rater) (Cerezci, 2020).

Finally, the current MCOP² research is also lacking in extending the usage of the MCOP² protocol from a scientific, grade-bearing instrument for evaluating teaching quality in mathematics classrooms towards a rich teacher professional development resource that can promote math teachers' self-reflection, self-evaluation, self-learning, and continuous professional growth. In other words, the potentials of the MCOP² protocol have not been explored and validated in offering detailed diagnostic information for individual teachers to understand their own weaknesses and strengths in becoming effective math teachers. Further research can also be conducted to investigate how to build an associated coaching model/framework for teacher learning and training by using the MCOP² protocol for the dual purpose of ranking and diagnostic assessments.

A Calibrated Framework for Rater-Mediated Assessment (MFRM)

At least two key advantages are present in using the MFRM to evaluate the rating quality of any rater-mediated performance assessment. First, all facets (e.g., examinee, rater, rating scale, items/tasks, etc.) are placed on the same logit measurement scale,

allowing for comparisons to be made across facets (Bond & Fox, 2015). Second, the MFRM model produces model expected estimates of the true scores/ratings examinees should have received, after accounting for measurement errors related to all included facets. Thus, these model-estimated scores and raw scores can be compared to make inferences about the extent to which rater-assigned raw scores represent examinees' true scores after correcting for measurement errors (Eckes, 2009; Wu & Tan, 2016).

Thus, the following review includes **MFRM-Based Dimensionality Analysis, Rater Effects, Interrater Reliability, and Multi-Facet Calibration Controlling for Rater Effects**. This section details both the theoretical foundation and empirical application of the MFRM approach.

MFRM-Based Dimensionality Analysis

MFRM (e.g., Linacre, 1995, 2007) is an extension of the partial credit modeling (PCM) within the Rasch family to rater-mediated assessment settings. Thus, MFRM can be applied to identify and measure all factors/facets (other than examinee ability and item difficulty) that can systematically influence examinees' rating scores (Bond & Fox, 2007). However, several key assumptions need to be tested prior to applying MFRM to model the data obtained from rater-mediated assessments. One of these assumptions is unidimensionality. Bond and Fox (2007) define the Rasch unidimensionality assumption as "useful measurement involves examination of only one human attribute at a time (unidimensionality) on a hierarchical 'more than/less than' line of inquiry." (p. 41) Specifically with respect to MFRM, Eckes (2005) further clarifies that the main question in rating-based scores is

whether ratings on one criterion followed a pattern that was markedly different

from ratings on the others, indicating that [test-taker] scores related to different dimensions, or whether the ratings on one criterion corresponded well to ratings on the other criteria, indicating unidimensionality of the data (p. 211).

In practice, however, researchers have raised concerns about the Rasch unidimensionality assumption and its appropriateness in rater-mediated performance assessment, since examinees' performance on a task (e.g., teaching, writing, piano performance, etc.) may involve utilizing a variety of abilities/skills, and its complexity/multidimensionality cannot be accurately captured by the models assuming measurement unidimensionality. Thus, Rasch models have been criticized for being simplistic (or reductionistic) and lacking in validity as they reduce multidimensional performance to a single score (Barkaoui, 2013; McNamara, 1996). To address these concerns, Bejar (1983) posits

unidimensionality does not imply that performance on items is due to a single psychological process. In fact, a variety of psychological processes are involved in responding to a set of test items. However, as long as they are involved in unison - that is, performance on each item is affected by the same process and in the same form - unidimensionality will hold (p. 31).

Using a simulation study, Henning (1992) further illustrated that psychological unidimensionality may be present in the context of psychometric multidimensionality, while psychometric unidimensionality may also be present in the context of psychological multidimensionality. Henning therefore concludes that dimensionality is dependent on the samples and supports the application of IRT approaches even in measuring a complex, multidimensional latent ability/trait.

Empirically, several procedures have been proposed to test the MFRM-based unidimensionality assumption, including (a) examining fit statistics, (b) conducting Rasch Factor Analysis (RFA) of residuals, and (c) examining Point biserial correlations (Barkaoui, 2013).

First, all facets must have infit and outfit statistics within the acceptable range (between 0.5 and 1.5) to uphold the unidimensionality assumption (Eckes, 2005; Linacre, 1998; Smith, 2002).

Next, it is recommended for researchers to perform a factor analysis on the residuals that remain after conducting a regular Rasch analysis (Bond & Fox, 2007; Linacre, 1998). This approach is referred to as Rasch Factor Analysis (RFA) or Principal Component Analysis (PCAR) of the standardized residuals. Unlike traditional CTT-based factor analysis based on raw scores, RFA is conducted with interval data in the form of logit measures (Bond & Fox, 2007; Linacre, 1998; Smith, 2002). The purpose of RFA is to determine if one or more other factors (than the measured latent ability/trait) explain the residual variance. If RFA identifies one or more factors suggesting a strong correlation between the item residuals left over from the variances explained by the latent trait, the presence of secondary structures or sub-dimensions within the data will be supported and the unidimensionality assumption cannot be upheld. In this case, researcher should consider modeling separate measures for the dimensions. Regarding the operational criteria in interpreting the RFA results, Smith and Miao (1994) propose that eigenvalues smaller than 1.4 are at the random level and can be ignored, while Linacre (2004) suggests that if the first residual factors explain less than 3.0 units of residual variance, the unidimensionality assumption should be considered met. Other

criteria include that (a) the variance explained by the latent trait be at least 40%, and (b) the variance explained by the first principal component of the residuals be no more than 15% (Linacre, 2006).

Last, Smith (2000) suggests that median point-biserial correlations should be positive and below .30 to support the assumption of unidimensionality. The presence of several median point-biserial correlations greater than .30 would be another indicator of multidimensionality, such as raters defining and using the rating scale in different ways.

Rater Effects

The rating quality obtained from any rater-mediated performance assessment is under the inevitable influence of rater judgment (Eckes, 2009; Myford & Wolfe, 2003), since the ratings directly represent raters' perceptions of examinees' work/performance, interpretations of the rubric, and analysis to determine to what extent the examinees' work/performance aligns with the rubric (Engelhard, 2002). Raters' personal understanding of the rubric and its application in judging examinees' work/performance may very likely disagree with the intended interpretations and uses of the rubric.

Systematic errors in raters' scores that reflect raters' personal characteristics and/or personal interpretations of the rubric is known as rater effects (Bond & Fox, 2015; Eckes, 2009; Myford & Wolfe, 2003; Scullen, Mount, & Goff, 2000). Various forms of targeted rating training programs are often created and implemented to mitigate the adverse impact of these rater effects on the rating quality. Unfortunately, research on rater training reveal that the effects of rater training are very limited in changing the behavior of raters who exhibit rater effects (Knoch, Read, & von Randow, 2007; Raczynski, Cohen, Engelhard, & Lu, 2015; Weigle, 1998).

To systematically examine the impacts of rater effects on student achievement estimates and on classification decisions, Wind (2019) conducts a simulation study and finds that when as few as 10% of the raters exhibit any type of rater effects, substantial changes will be identified in students' classifications within rating scale categories compared with their classifications when no raters exhibited the effects; and as the proportions of "problematic" raters increase, changes in the values of the student achievement estimates and the rank orderings of students will become more pronounced.

Therefore, this section of literature review focuses on three most common types of rater effects: rater severity, rater centrality, and rater misfit (Myford & Wolfe, 2003).

Rater severity. Rater severity/leniency (also called rater harshness or the hawk effect) refers to a rater's tendency to systematically assign lower or higher ratings to student performances, respectively, than one would expect if the rater applied the scoring rubric appropriately (Eckes, 2009, 2015; Engelhard, 1992; Saal et al., 1980). Raters are considered severe if they consistently assign low scores across all examinees, and lenient if they consistently assign high scores across all examinees (Bond & Fox, 2015; Eckes, 2015; Wolfe, 2004).

Severe raters are problematic because when they score examinee performances, examinees tend to receive ratings that underestimate their latent ability/proficiency. Similarly, lenient raters are problematic because they tend to assign ratings to examinees that overestimate their ability/proficiency (Myford & Wolfe, 2003). Regrettably, raters are often found to vary drastically from one another in their severity (Eckes, 2005; Han, 2015; Lunz & Stahl, 1990), contrary to the common assumption held by many researchers of rater-mediated assessments that raters are of similar rating severity after

training and practice (Lunz, Wright, & Linacre, 1990).

Raters' leniency or severity may also change across rating occasions over a period. Related literature indicates that raters tend to become more severe over time, especially across rating periods of several days or more (Leckie & Baird, 2011; de Moira, Massey, Baird, & Morrissy, 2002). However, in a study evaluating rater effects in AP English Literature and Composition essays, Wolfe and his colleagues (2007) find that only 5% of the raters become more severe over time, whereas 16% of the raters actually become more lenient. Hence, it seems that the direction of raters' changes in their leniency and severity may not always be the same and predictable across rating occasions.

Raters' leniency and severity may also vary across rubric dimensions. Specifically, raters may rate more severely on some rubric dimensions compared to other dimensions. For instance, Eckes (2005) find that more than one-third of raters exhibit differential severity across rubric elements the writing assessment in the Test of German as a Foreign Language. Such interactions between rater leniency/severity and rubric elements/dimensions are also referred to as differential rater functioning or bias (Eckes, 2015). Such rater bias can be particularly problematic in compensatory models where examinees are given differential credits by rubric elements/dimensions.

Furthermore, raters' leniency and severity may not always be constant across scoring levels. In a study related to an Oral English Proficiency Test, Yan (2014) finds that raters differ in their severity or leniency depending on scoring levels: raters agree more with one another for tests that score on the passing side of the score levels than for tests that score on the failing side of the score levels. Consequently, raters are unable to

rate consistently in line with the intended scoring criteria across score levels. This effect may seriously threaten the validity and reliability of the ratings in pass-or-fail performance assessment settings.

In sum, rater severity and severity drift effects can vary dramatically across rating occasions, raters, samples, rubric elements/dimensions, and assessment settings, and need to be investigated and handled very carefully in practice.

Rater centrality. Range restriction effects are evident if raters systematically limit their ratings to a subset of the available rating scale categories which fails to capture the true variability of examinee performances across all the rating categories. Range restriction effects can take various forms of raters' overuse of the lowest, middle, or highest categories of a rating scale. Among others, the most frequently discussed type of range restriction effects in related literature is centrality (also referred to as central tendency), or raters' tendency to limit their ratings to the middle category or categories of a rating scale (Wind, 2019; Wolfe & Song, 2015). These central tendency raters tend to assign ratings which underestimate the examinees' latent ability/skill/proficiency when their performances warrant ratings in the highest category. Similarly, those examinees whose performances warrant ratings in the lowest category instead receive ratings from these effects raters that overestimate their ability/skill/proficiency (Wind, 2019).

As cited in Leckie and Baird (2011, p. 400), central tendency is a well-documented phenomenon across various contexts including in the assessment of Advanced Placement English Literature and Composition essays (Myford & Wolfe, 2009), school writing examinations in Georgia (Engelhard, 1994), English as a second language (Knoch, Read, & von Randow, 2007), and writing and speaking in German as a

foreign language (Eckes, 2005). For instance, Knoch et al. (2007) find that rater training and practice increase the central tendency effect of the raters in their scoring of an English writing examination in a New Zealand university. A possible explanation they provide for this phenomenon is that raters are more likely to exhibit central tendency when they are aware that they are being monitored closely. Wolfe et al. (2007) refer to this rater psychological state as a "play-it-safe" effect because raters know that their ratings will be less likely to be questioned if they avoid using the extreme categories of the scale.

Raters may show centrality effects if they are unable to differentiate between the scoring criteria across score levels, especially when scoring criteria are ambiguously worded (Myford & Wolfe, 2003). If raters cannot fully appreciate and apply the scoring criteria differences across score levels with confidence, they may tend to assign ratings around the mid-range of the scoring levels. Consequently, like the halo effect, centrality may also threaten the validity of the rating data by limiting the variability of examinees' ratings. However, unlike the halo effect which often affects the ratings assigned to certain individual performances and results in limited score variability across rubric dimensions for those individual examinees, centrality effect can impact the ratings for all examinees, which results in limited score variability either within or across any of the examinees.

However, it is important to note that the presence of high amount of middle-category scores/ratings does not necessarily indicate rater centrality, since it could reflect the actual distribution of examinees' moderate abilities. To determine if rater centrality is indeed a problem, the variability of ratings across all examinees on each rubric

element/dimension can be evaluated for any individual rater showing a potential central tendency effect. First, the mean rating can be calculated by averaging all examinee scores on each rubric element. Next, a standard deviation around the mean rating can be computed for each rubric element (Saal et al., 1980). A mean rating close to the mid-range of the rating categories with a small standard deviation would indicate the presence of a central tendency effect for that particular rater on the particular rubric element in question.

Rater misfit. Rater misfit occurs when a rater interprets the scoring rubric very differently from the way it is intended to be used, giving rise to a large discrepancy between this rater's ratings and the expected ratings if that rater had applied the rubric appropriately. Such rater misfit effects are also referred to as rater inaccuracy (Wolfe & McVay, 2012), noisy ratings (Wind & Engelhard, 2013), or within-rater rating category disordering (Wind & Engelhard, 2017).

In rater-mediated performance assessments, empirical verification of the intended ordering of the rating scale categories for each rater is as important as the detection of various rater effects in ensuring measure stability and accuracy (fit) (Linacre, 2002, 2010; Wind, 2014). Especially when raters are trained to assign scores according to a set of ambiguously worded rubrics, cross-rater differences in understanding and interpreting each of the defined rating categories can be concerning and cause individual rating category disordering.

According to Barkaoui (2013), rater misfit may pose a more serious threat to general test validity than overfit or test-taker misfit because it indicates divergent behavior from the norm on the part of the raters, and its effect on all other facet measure

estimates can be strong (Bonk & Ockey, 2003). To make matters worse, Rasch models do not adjust scores for rater misfit as they do for rater severity (Bonk & Ockey, 2003, p. 101; Myford & Wolfe, 2000, 2003, 2009).

To address this issue, the divide-by-total IRT models based on adjacent-categories probabilities are appropriate for investigating category disordering. In particular, the Rasch-MFRM partial credit models can be used to yield threshold location estimates for each individual rater, an important indicator of possible rating scale category disordering under the polytomous Rasch framework (Andrich, 2004, 2013, 2015; Andrich, de Jong, & Sheridan, 1997). However, this long-standing diagnostic practice has been strongly questioned, since it is recently found that disordered categories can have both ordered or disordered threshold estimates, and threshold disordering often only reflects the irregular distribution of observations in certain response categories (Adams, Wu, & Wilson, 2012; García-Pérez, 2017; Linacre, 2002, 2012).

Linacre suggests examining the ordering of average category measures (ACMs) as well as their associated outfit indices instead for detecting disordered rating categories for individual raters, when polytomous Rasch models are estimated using the joint maximum likelihood method (e.g., in Winsteps and Facets). Alternative indicators are also proposed for other divide-by-total IRT models, including overall model-data-fit and graphical indices of the option response functions (ORFs)/ item step response functions (ISRFS) among others (García-Pérez, 2017; Muraki, 1993; Wind & Peterson, 2018).

Interrater Reliability

In rater-mediated performance assessments, it is vital to understand how each of the interrater reliability (IRR) indices works to support the validity and reliability of the

rating data. Special caution must be exerted in distinguishing the interrater reliability indices under the CTT framework from those based on the MFRM method, as well as the associated methodological benefits and limitations.

Under the CTT framework, interrater reliability (IRR) can be conceptualized in many different ways (Bramley, 2007; Hayes & Krippendorff, 2007). Among those, two most commonly used classes of IRR indices include a consensus index of interrater agreement and a consistency index of interrater reliability. A consensus index of interrater reliability refers to the degree to which independent raters assign number-identical ratings to a particular examinee on a particular item/task (absolute correspondence of ratings). While a consistency index of interrater reliability refers to the degree to which independent raters assign ratings so that the performance of all examinees is ordered or ranked in an identical way (relative correspondence of ratings) (Eckes, 2011).

Following the CTT approach, two IRR indices can be computed in practice. For consensus indices, exact interrater agreement index (i.e., the number of examinees awarded identical ratings on a particular item/task divided by the total number of examinees commonly rated by the raters) and Cohen's weighted kappa (i.e., this index corrects the interrater agreement for agreement based on chance alone). The weighted kappa should be selected for computing the IRR for ordered categories on the rating scale (Cohen, 1968), where the higher disagreement (two ratings further apart across the rating categories, such as 0 and 4 on a 5-point scale) leads to the higher weight assigned. The values of weighted kappa normally range between 1 (agreement is perfect) and 0 (agreement is no better than by chance alone). Negative values of weighted kappa

suggest worse-than-chance agreements among raters (Mun, 2005).

For consistency indices, the product-moment correlation (Pearson's r) or Kendall's τ - b coefficients can be calculated. Pearson's r represents the linear relationship between two raters' ratings. Kendall's τ - b represents the degree of relative correspondence between rank orderings of examinee performances assigned by two raters. The values of both Pearson's r and Kendall's τ - b fall within the $[-1,1]$ range, with higher values indicating stronger correlation between two raters' ratings.

However, contrary to the common assumption pervasive in the CTT-based validation research on rater-mediated performance assessments, high IRR values (as mentioned above) alone do not provide sufficient evidence in support of the psychometric quality of the rating data. In other words, a conclusion cannot be drawn solely based on high CTT-based IRR values that raters are expected to assign ratings accurate enough to reflect examinees' "true" latent ability/skill/proficiency when applying a rater-mediated performance assessment instrument.

To further clarify the limitations of using the CTT-based IRR indices in rater-mediated performance assessment validation studies, Eckes (2011) proposes a term "agreement accuracy paradox", referring to the fact that high agreement and/or high consistency only reflect the homogeneity of raters and ratings to a certain extent, and should not be considered as equal to high rating accuracy (Henning, 1996). On the other hand, low consensus and/or low consistency may only reflect heterogeneity of raters and ratings unrelated to the misuse of the measure/rubric and does not necessarily indicate inferior rating data quality. Such false assumption can even lead to severe consequences such as dismissing or replacing the raters deemed "unreliable/problematic", which may

negatively affect the overall rating data quality.

In contrast, the rater-related reliability indices based on the MFRM analysis are conceptualized very differently from the CTT-based IRRs. First, the reliability index associated with the rater facet does not refer to the traditional index of inter-rater agreement; instead, it indicates the ability of the MFRM analysis to reliably/consistently separate raters into different levels of severity. Therefore, a low reliability index close or equal to zero is quite desirable, as it suggests that raters rate examinee performances at about the same level of severity and they are interchangeable in the rating process (McNamara, 1996; Weigle, 1999). Similar to the rater-facet reliability, a low separation index also indicates that the assumption of equivalence among raters is upheld after calibrating the estimated measures of all facets in the MFRM model (Lunz et al., 1996; Weigle, 1998). Fixed χ^2 values associated with the rater facet can also be used to test the assumption that all raters are equal in their level of severity, where a low χ^2 value shows that raters are closely aligned in terms of severity (Weigle, 1998).

Further, the MFRM analysis also yields statistical reports on the observed and expected percentages of exact rater agreements. Empirically, both these interrater agreement indices are calculated just like the CTT-based interrater agreement index, referring to the proportion of examinees who receive number-identical observed or expected ratings from their common raters. If the observed agreement rate is too low compared to the expected agreement rate, the MFRM model should not be used for predicting interrater agreement with confidence. Whereas if the observed agreement rate is much higher than the expected rate, it could be possible that raters are under the influence of external circumstances (e.g., being trained to avoid too much disagreement

with other raters) to force agreement with each other, which may lead to compromised quality of independent ratings for those raters (Linacre, 1989).

Finally, the point-biserial correlation for each rater (also referred to as the “single rater—rest of the raters” (SR/ROR) correlation) measures the correspondence between that rater’s ratings and the total ratings of all other raters that rated the same examinees’ performances. The mathematical formula used to compute this correlation coefficient is a many-facet version of the Pearson product-moment correlation (Linacre, 2001).

Myford and Wolfe (2003) clearly explain the empirical standards in interpreting the point-biserial (SR/ROR) correlation coefficients:

SR/ROR correlations less than .30 are considered to be somewhat low, while correlations greater than .70 are considered to be high for a rating scale composed of several categories. However, as the number of rating scale categories decreases, these rule-of-thumb values should be relaxed. For example, it is not uncommon to see SR/ROR correlations no higher than .20 in dichotomous ratings. If a SR/ROR correlation is near zero or negative for a given rater, then that rater rank orders ratees in a manner different from the other raters’ rank ordering (p. 410).

Multi-Facet Calibration Controlling for Rater Effects

The MFRM model is an extension of the single-facet rating scale Rasch model (Andrich, 1978) and single-facet partial-credit Rasch model (Masters, 1982), which allows for multiple facets to be included in the evaluation of polytomous-scored assessment items. Specifically in rater-mediated performance assessments, the rater facet, item/task facet, and other facets that contribute to the construct-irrelevant variances

of the measurement can all be added to the original examinee ability/skill/proficiency facet in the MFRM model to systematically evaluate the rating scores.

In a MFRM analysis, the log-odds of each transition between adjacent rating scale categories are estimated as one parameter that can represent the level of performance proficiency (for ratees), severity (for raters), and difficulty (for traits, and for rating scale categories). Mathematically, a MFRM version of the rating scale model takes the following basic form (Linacre, 1990):

$$\ln[P_{nij k} / P_{nij k-1}] = B_n - D_i - C_j - F_k, \quad (2)$$

where $P_{nij k}$ denotes the probability of examinee n being rated k on item/task i by rater j , while $P_{nij k-1}$ refers to the probability of examinee n being rated $k - 1$ on item/task i by rater j . B_n represents level of performance proficiency for examinee n , and D_i means difficulty of item/task i . Rater parameter C_j denotes severity of rater j , and F_k refers to difficulty of scale category k relative to scale category $k - 1$ (i.e., thresholds).

When the rating category thresholds are not assumed to be equal across all categories and for all raters, a MFRM version of the partial credit model may be defined based on the adaptation of Equation (1) as below:

$$\ln[P_{nij k} / P_{nij k-1}] = B_n - D_i - C_j - F_{ikj}, \quad (3)$$

where $P_{nij k}$ denotes the probability of examinee n being rated k on item/task i by rater j , while $P_{nij k-1}$ refers to the probability of examinee n being rated $k - 1$ on item/task i by rater j . B_n represents level of performance proficiency for examinee n , and D_i means difficulty of item/task i . Rater parameter C_j denotes severity of rater j . F_{ikj} still represents difficulty of scale category k relative to scale category $k - 1$ but is free here to vary across item/task i and rater j (Eckes, 2015).

A partial credit model is specified based on the assumption that each rater interprets and uses each rubric element/dimension in their own individual ways. Thus, the partial credit model is a more complex model than the rating scale model and allows for the estimation of additional parameters for both raters and rubric element thresholds (Bond & Fox, 2015; Eckes, 2015; Myford & Wolfe, 2003).

The MFRM analysis allows researchers to evaluate the impact of each facet on the measurement process by estimating its unique parameter (e.g., level of severity for each rater), and then to compute the overall probability of any examinee performing on any item/task for any score category threshold and for any rater, after accounting for the estimated parameters of all facets (Bond & Fox, 2007). It is in this sense that MFRM is fully capable of modeling various facets in the assessment setting, estimating their effects on ratings, and placing them on the same logit scale for comparison. Each facet is calibrated from the potentially ordinal raw ratings (as rating scales are often used in rater-mediated performance assessments), and all facets (examinee, task, rater, etc.) are placed on a single common linear scale called a variable or facets map. Thus, MFRM treats each rating as a function of the interaction between examinee ability, task difficulty, criterion difficulty, rater severity, and possibly the effects of other external, measurement-irrelevant factors (Barkaoui, 2013; McNamara, 1996).

Myford and Wolfe (2003) concisely summarize the benefits of a MFRM approach in detecting and controlling for rater effects compared to other traditional CTT methods. Similar to an ANOVA-based approach, MFRM can be applied to investigate group-level rater effects (i.e., main effects), as well as rater-effect interactions. However, unlike ANOVA, the MFRM analysis does not allow possible interaction effects to contaminate

main effects, making the interpretation of main effects difficult. Furthermore, the MFRM approach is not only capable of detecting main effects and interaction effects but is also very effective in identifying individual-level rater effects, an important methodological advantage that can be utilized for the diagnosis and intervention of rater effects in rater-mediated performance assessment.

For each element of each facet, the MFRM analysis produces a measure (a logit estimates of the calibration), a standard error (information about the precision of that logit estimate), and fit indices (information about how well the observed scores associated with this parameter fit the expected scores of the measurement model). Besides these individual level statistical indicators, MFRM also provides several group-level statistical indicators useful for detecting pervasive trends in the data (e.g., separation statistics, fixed effect chi-square tests, summary fit statistics). Details are further discussed in **CHAPTER III: METHOD** regarding how to obtain and interpret these MFRM-based statistical indicators.

CHAPTER III

METHOD

Introduction

Recent increasing use of rater-mediated performance assessments (RMPA) for teaching and learning in mathematics classrooms (e.g., classroom observation protocols) calls for rigorous research to be conducted regarding the instrument validation, interpretation of assessment results (e.g., rater-assigned scores/ratings for teachers under observation), extended usage across educational contexts, and implications for facilitating future math teacher learning and training. The involvement of various rater effects/bias in such performance assessments further complicates the issues such as how to detect and control for measurement errors originated from construct-irrelevant variance sources (i.e., rater, examinee, test, and other external factors). Traditional approaches under the classical test theory (CTT) framework (e.g., factor analysis such as EFA and CFA, content validity, internal consistency such as Cronbach's alpha, interrater agreement and reliability, ect.) are proven theoretically and methodologically limited in effectively handling these issues. Consequently, rater-mediated performance assessments for teaching mathematics have mostly been used as a type of formative rather than summative assessment measures, since the numeric results yielded from the RMPA

process (e.g., rater-assigned ratings) cannot be reliably compared across teaching contexts as effective indicators of more or less teaching proficiency.

Thus, the purpose of this study is to evaluate the rating quality obtained from a K-16 math classroom observation protocol (MCOP²) under a MFRM framework for rater effects detection and control.

The remainder of this chapter is divided into eight sections. First, the **Research Questions** are identified, followed by the explanation about the **Research Design**. The **Participants** are then described. Next, the **Instrumentation** (i.e., MCOP²) is discussed in detail. The **Data Analysis** addresses description of the data analysis plan for each of the seven research questions. Then **Ethical Standards** reviews principles of research procedures and behaviors with respect to human subjects' protection. This chapter ends with a brief **Summary**.

Research Questions

This study seeks to answer seven research questions regarding how to control the construct-irrelevant measurement errors of the MCOP² protocol using a MFRM analysis. The specific research questions are repeated here as follows from **CHAPTER I**, for the convenience of the readers.

1. To what extent do the observed rating data obtained from the MCOP² instrument fit the MFRM modeling?
2. To what extent does the MCOP² observation protocol separate observed teachers into distinct levels of proficiency?
3. To what extent do raters differ in terms of the relative severity with which they rate observed teachers?

4. To what extent do raters consistently rate the teaching performance of observed teachers?
5. To what extent do raters consistently rate the teaching performance of observed teachers across the MCOP² items?
6. To what extent can the score levels of the MCOP² items be distinguished, without certain score levels being either underused or overused?
7. To what extent are the rater behaviors associated with the professional background characteristics (i.e., in-service vs. pre-service teachers, schools, and teaching grade levels) of the observed teachers?

Research Design

This study essentially consists of a validation study of the MCOP² protocol within a Rasch framework using the MFRM analysis, including the investigation of dimensionality, examinee fit, item fit, rater fit, overall data-model-fit, as well as possible interactions between any of the modeled facets/factors. The key lies in systematically calibrating the measures of all the involved facets (e.g., test item, examinee, raters, and other external factors) on a common continuum scale, so that the construct-irrelevant measurement errors (especially rater bias) can be effectively identified and accounted for. The calibrated teacher ratings/scores after the MFRM analysis can theoretically be compared with confidence across different classroom teaching contexts.

Participants

This section describes the demographic and/or professional characteristics of the three groups of human subjects involved in this study: teachers of the math classrooms observed (i.e., ratees) in the MCOP² sample data used in the study, and raters recruited to

rate the math teachers' teaching practices based on the MCOP² rubric.

Rates

The MCOP² rating data used in this study draws from two secondary data sources with the permission of the data owners: the first MCOP² sample consists of a cross-sectional (i.e., one-time classroom observation) dataset collected by the MCOP² development and research team at University of Alabama in 2016 for their final MCOP² validation study (Gleason, Livers, & Zelkowski, 2017); while the second MCOP² sample is comprised of longitudinal data compiled over a period of three years by the teacher educators at University of Kentucky from 2017 to 2020. For the convenience of writing, the first secondary dataset is hereby referred to as Sample AL, and the second as Sample KY.

Sample AL results from observations of 40 elementary, 53 secondary, and 36 tertiary mathematics classrooms in the southeastern United States. The classrooms observed at each grade level include math teachers with experience ranging from 0 to 40 years, a mixture of gender matching national norms for each grade band, and a mixture of direct and dialogical instruction in the lessons (Gleason, Livers, & Zelkowski, 2017). Sample AL include classroom observations of both in-service teachers ($n = 101$) and preservice teachers ($n = 28$).

Sample KY consists of observations of 108 K-16 mathematics classrooms in the neighboring school districts surrounding the Lexington, Kentucky area, involving 30 preservice teachers enrolled in a pedagogical methods course near the end of their teacher education programs. The MCOP²-based classroom observations in Sample KY were conducted mainly as a type of formative assessment associated with the methods course

to evaluate and guide the student teachers' professional learning in general classroom instruction and implementation of specific pedagogical strategies. Therefore, most of the student teachers provided four classroom observations each, which were rated by the same rater (i.e., student teachers' field study supervisors) usually spanning a period of two to three months in a semester. No other demographic information (e.g., gender, ethnicity, etc.) was provided for the preservice teachers in Sample KY (College of Education, UK, 2020).

Raters

For Sample AL, a total of five raters were asked to observe and rate the math teachers' classroom teaching performance. Prior to observing classes, the raters were arranged to analyze five different classroom videos to determine the interrater reliability of the MCOP² instrument.

Gleason, Livers, and Zelkowski (2017) note that the five raters vary in their educational and professional backgrounds. Two of the raters hold doctorates in mathematics education (one elementary-focused and the other secondary-focused), one rater holds a doctorate in mathematics with heavy involvement in the mathematics education community (both elementary and secondary), one rater works as a mathematics specialist with secondary teachers and has taught at both the secondary and postsecondary levels, and the fifth rater is a graduate student in mathematics with minimal background in education other than teaching some introductory college math classes.

All raters received the detailed descriptions of the items with the rubric prior to observing classes and asked some clarification questions prior to the observations.

However, no formal training on the use of the instrument occurred, simulating the probable future uses.

While for Sample KY, a total of seven raters were employed to observe and rate the math teachers' classroom teaching performance. No formal or informal rater training was documented to have been arranged for the raters prior to observing classes.

The seven raters in Sample KY all served as student teacher supervisors and/or university-based faculty members (e.g., instructors of a pedagogical methods course) during the data collection period. They also come from various educational and professional backgrounds. However, no further detailed information is available in Sample KY to describe each rater's demographic and/or professional profile (College of Education, University of Kentucky, 2020).

Just like the raters in Sample AL, all seven raters in Sample KY also received the detailed descriptions of the items with the rubric prior to observing classes.

Mathematics Classroom Observation Protocol for Practices (MCOP²)

The Mathematics Classroom Observation Protocol for Practices (MCOP²) is designed to be implemented in K-16 mathematics classrooms to measure the activities occurring in a mathematics classroom during a single lesson. Based on the confirmatory factor analyses findings (Gleason, Livers, & Zelkowski, 2017), the MCOP² measures two primary constructs (i.e., teacher facilitation and student engagement) with a total of sixteen items with full descriptions (the content validity of the 16 MCOP² items are supported by the feedback of 164 experts in mathematics education). The factorial structure of the MCOP² is depicted in Figure 1. The double arrows between the two theoretical constructs indicate the correlation of these two factors. The model also

includes residual error terms to account for unknown measurement errors in the model.

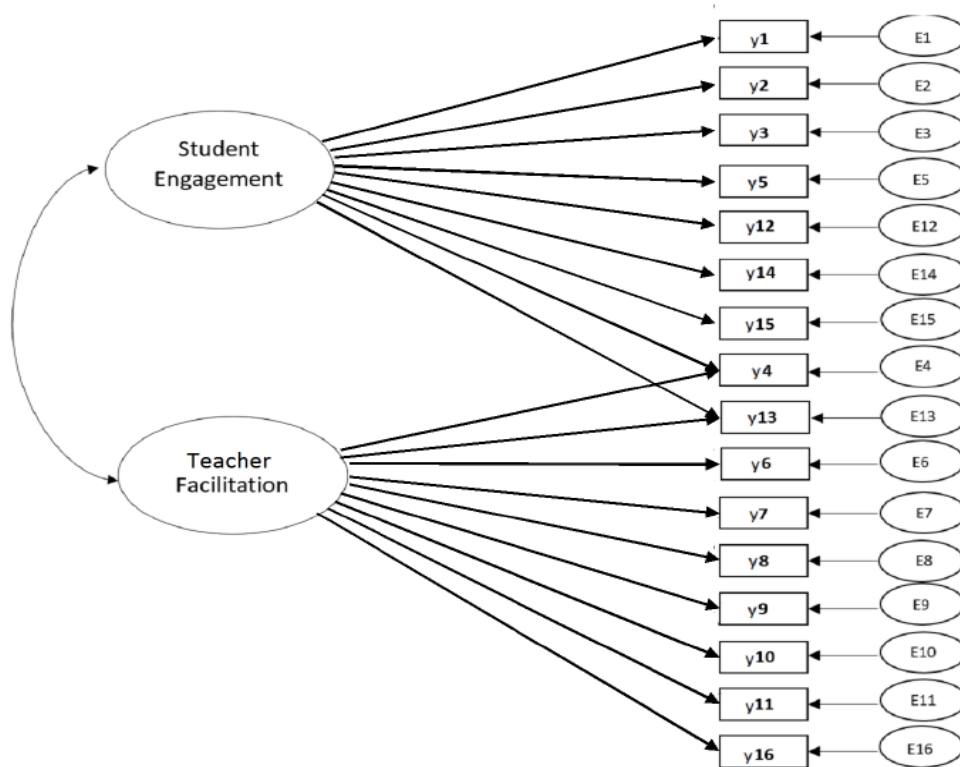


Figure 1. The MCOP2 Theoretical Model

Operationally regarding the scoring guidelines derived from the factor structure shown above, the MCOP² measures two distinct factors of Teacher Facilitation and Student Engagement through two subscales of 9 items each (Gleason, Livers, & Zelkowski, 2015). It is worth noting that the MCOP² is not designed to get a single score of a classroom.

The Teacher Facilitation subscale (Cronbach alpha of 0.850) measures the role of the teacher as the one who provides structure for the lesson and guides the problem-solving process and classroom discourse. To calculate the score for the Teacher Facilitation subscale, one would add the scores for items 4, 6-11, 13, and 16. While the Student Engagement subscale (Cronbach alpha of 0.897) measures the role of the student

in the classroom and their engagement in the learning process. To calculate the score for the Student Engagement subscale, one would add the scores for items 1-5 and 12-15 (Gleason, Livers, & Zelkowski, 2015). Figure 2 outlines the MCOP² scoring roadmap.

Item	Student Engagement	Teacher Facilitation
1	X	
2	X	
3	X	
4	X	X
5	X	
6		X
7		X
8		X
9		X
10		X
11		X
12	X	
13	X	X
14	X	
15	X	
16		X

Figure 2. The MCOP² Scoring Roadmap

In addition, other psychometric properties of the MCOP² within the CTT framework, such as interrater reliability (IRR), have also been calculated with a panel of five raters of various backgrounds without any formal training. This results in the intra-class correlation (ICC) of 0.669 for the Teacher Facilitation Sub-scale and 0.616 for the Student Engagement Sub-scale, indicating acceptable interrater reliability for using MCOP² in classroom observations (Gleason et al., 2017).

Finally, as Gleason, Livers, and Zelkowski (2015) recommend in the MCOP² Descriptors Manual, it is important to note that in using the MCOP² for classroom observation and teaching performance assessments, if one desires to measure the overall activities of a class, the form should be used to measure at least three different class

settings. An important item to remember is that while all of the items in the observation protocol are desired qualities of a mathematics classroom, not all of them are expected to be observed during a single lesson. It is expected that this instrument be used in a formative manner on single observations. Summatively, 3-6 observations are ideal in evaluating classroom instruction (p. 1).

Data Analysis

The two secondary MCOP² sample datasets (i.e., Sample AL and Sample KY) are combined as a final empirical study sample ($n = 237$) involving observations of 237 K-16 classrooms of different subjects, grade bands, and schools in the southeastern region of the United States, 159 in-service and preservice math teachers from various demographic and professional backgrounds, and 12 independent raters who are university instructors/researchers of mathematics teacher education, K-16 school teacher leaders, and/or student teaching supervisors.

Since the two sample datasets (i.e., Sample AL and Sample KY) are disconnected (i.e., there is zero rater-ratee overlap between the two samples), an **anchoring** technique is employed to link the two sample datasets so that the measures estimated in the combined sample MFRM analysis for each facet (especially the rater facet) are directly comparable: first, the MFRM analysis is performed only for Sample AL; and then the estimated measures for the raters in Sample AL are used for anchor values in the ensuing MFRM analysis for the combined sample dataset (Linacre, 2012). Unlike the **group-anchoring** technique, the anchor values for Sample AL raters do not need to sum up to zero. When such element-anchoring technique is used, Linacre (2012) strongly recommends that at least one facet is unanchored and non-centered, or the analysis will

be over-constrained, and will not estimate correctly. If the subsets are linked/connected after the element-anchoring procedures, the FACETS program will yield relevant diagnostic information such as “Subset connection O.K.”; otherwise, system warnings will be given (e.g., “Warning! There may be 2 disjoint subsets”).

MFRM analysis is applied to the study sample data to address Research Questions 1-7 for the purpose of (a) validating the MCOP² protocol within the Rasch framework and (b) calibrating the measures of all involved facets to account for any construct-irrelevant variances. All MFRM analyses in this study are implemented using the software program Facets, version 3.83.3 (Linacre, 2020).

Analyses Plan for Research Question One

Research Question 1 (i.e., To what extent do the observed rating data obtained from the MCOP² instrument fit the MFRM modeling?) is evaluated by testing the MFRM modeling assumptions, including local independence, unidimensionality, overall model fit, rater fit, and item fit.

Local independence. Local independence (LID) refers to the assumption that item responses are independent from one another after controlling for the construct of interest (DeMars, 2010). Therefore, there should not be any correlation between two items after controlling for the underlying trait. In other words, the items should only be correlated through the latent trait that the test is measuring (Lord and Novick, 1968). However, this LID assumption is almost always violated to various extents in empirical applications. In the case of significant item residuals correlations, the items in a test can be regarded as locally dependent on each other, or there might exist to a secondary dimension in the measurement not accounted for by the main dimension trait.

Violations of local independence (LD) are problematic because they may influence parameter estimates (Li, Li, & Wang, 2010; Smith, 2005) as well as inflate reliability estimates (Marais & Andrich, 2008; Wainer & Thissen, 1996; Wang, Cheng, & Wilson, 2005), since locally dependent items always cause substantial information loss for IRT modeling (Chen & Thissen, 1997).

Among the variety of methods for identifying LD that have been proposed in the related literature, the most widely used approach is based on Yen's Q_3 (1984, 1993) statistics through computing item residuals (observed item responses minus their expected values), and then correlating these residuals. Thus, in practice, LD is detected through observing the correlation matrix of item residuals based on estimated item and person parameters, and residual correlations above a certain cut-off value are pinpointed as the items that appear to be locally dependent.

Although no single critical cut-off value of Q_3 statistics is appropriate across all situations, simulation studies show that the Q_3 critical value appears to be reasonably stable around a value of 0.2 above the average residual correlation (Marais, 2013). That is to say, any item residual correlation that is 0.2 above the average residual correlation would appear to indicate LD, and any residual correlation of independent items that is 0.3 above the average correlation would seem unlikely (Christensen, Makransky, & Horton, 2017).

The Yen's Q_3 (1984, 1993) statistics for the MCOP² data used in this study can be calculated and investigated as part of the Principal Component Analysis of Residuals (PCAR) conducted in the Winsteps software program, version 4.7.0 (Linacre, 2020), where Table 23.99 (i.e., Largest residual correlations for items) can be obtained for

pairwise, item-level residual correlations by specifying the command of “PRCOMP = R” in the control file.

Unidimensionality. Unidimensionality is related to local independence and refers to the assumption that all assessment items measure only one, common construct (Bandalos, 2018; DeMars, 2010). Unidimensionality is evaluated by conducting a Principal Components Analysis (PCA) on the standardized residuals (PCAR) following the analysis of a basic 4-facet MFRM analysis (i.e., ratees + MCOP² items + raters + classrooms) in Facets. The PCAR was conducted using the Winsteps software program, version 4.7.0 (Linacre, 2020).

Standardized residuals were estimated as

$$Z_{nij} = \frac{x_{nij} - e_{nij}}{\sqrt{w_{nij}}} \quad (4)$$

where x_{nij} is the observed rating for student n on element i assigned by rater j ; e_{nij} is the expected rating for student n on element i assigned by rater j , given the model; and w_{nij} is the variability of the observed rating around its expected rating, given the model, otherwise known as model variance (Eckes, 2015).

The expected rating may be further defined as

$$e_{nij} = \sum_{k=0}^m k p_{nij k} \quad (5)$$

where k is a rating and $p_{nij k}$ is the probability of student n obtaining rating k on element i from rater j , given a specified MFRM model (Eckes, 2015). In the same fashion, the model variance may be further defined as

$$w_{nij} = \sum_{k=0}^m (k - e_{nij})^2 p_{nij k} \quad (6)$$

which can be used to calculate the square root of the model variance, the statistical

information contributed by a particular rating (Myford & Wolfe, 2003).

The procedures of conducting a PCAR analysis in Winsteps are as follows: (a) a basic 4-facet MFRM analysis (i.e., ratees + MCOP² items + raters + classrooms) is carried out in Facets to produce the measures of all the four facets, and (b) a rectangular data output file is exported from Facets into Winsteps, containing the MCOP² items as its columns and “ratees + raters” combined as its rows for a PCAR analysis in the Rasch framework.

PCAR analyses are used to evaluate whether there are systematic patterns in the item-level standardized residuals. If there are patterns in the residuals, a secondary dimension (i.e., a contrast) may be present. It is assumed that all items should be loaded on the first contrast of the Rasch dimension, and the PCAR specifically tests whether any items group on secondary contrasts. Each contrast has an associated eigenvalue, and the eigenvalues represent the number of items that make up the respective contrast. If eigenvalues for secondary contrasts are less than 2.0 (indicating there are fewer than two elements on the secondary contrasts), the unidimensionality assumption is met.

However, if eigenvalues for any of the secondary contrasts are greater than 2.0, it is recommended to further examine the disattenuated correlations between the person measures on the suspect cluster of items and the person measures on the other items. If the correlations are greater than 0.70, the suspect cluster of items is probably only measuring a secondary strand of the main Rasch dimension and should not be considered as a different dimension. By contrast, disattenuated correlations less than 0.30 or even negative values indicate that the suspect cluster of items is measuring something different than the construct of interests, and multidimensionality may become an issue (Linacre,

2012).

Overall model fit. To evaluate the overall model fit of the MFRM analysis, the absolute values of the standardized residuals are examined. Standardized residuals represent the number of standard deviations the observed score/rating deviates from the expected score/rating. For instance, standardized residuals of $|2.0|$ indicate that the observed score deviates by two standard deviations from the expected score. Thus, the related model-fit evaluation standard is that standardized residuals greater than $|2.0|$ indicate highly unexpected scores, and they should be expected to appear less than 5% of the time in data that fit well with the chosen MFRM model (Bond & Fox, 2015; Eckes, 2015; Myford & Wolfe, 2003; Wright & Masters, 1982). In this study, data are deemed to have good overall model-fit in the MFRM analysis, if fewer than 5% of the standardized residuals appear greater than or equal to $|2.0|$.

Rater fit and item fit. Mean Square outfit and Mean Square infit statistics (also referred to as MSU and MSW) are calculated and investigated (Bond & Fox, 2015; Eckes, 2005; Engelhard, 1994, 2002; Myford & Wolfe, 2003) to evaluate rater fit or item fit (Bond & Fox, 2015; Eckes, 2005; Engelhard, 1994, 2002; Myford & Wolfe, 2003).

For raters, the unweighted mean square (MS_U) index (i.e., MnSq outfit statistics) refers to an average of raters' squared standardized residuals for all examinees and items; while the weighted mean square (MS_W) values (i.e., MnSq infit statistics) are weighted by statistical information, resulting in differential weighting of ratings. Specifically, ratings assigned in score levels further from the examinees' ability are weighted less heavily than ratings assigned to the other score levels, as less information is contributed to the model by these extreme scores (Bond & Fox, 2015; Eckes, 2015). Similarly, the

MnSq outfit and infit statistics for items are calculated as a (unweighted or weighted) average of the items' squared standardized residuals for all examinees and raters, respectively.

MnSq outfit and infit indices range from 0 to positive infinity, with values of 1.0 indicating perfect fit of the data to the model (Linacre, 2003). Values less than 1.0 indicate that the observed ratings are closer to the model-implied ratings than would be predicted by the model (i.e., overfit of the model), and values greater than 1.0 indicate that the observed ratings are less similar to the model-implied ratings than would be predicted by the model (i.e., underfit of the model) (Eckes, 2015; Linacre, 2003).

Various benchmarks have been proposed for acceptable fit based on MnSq outfit and infit indices. Linacre (2003) proposes that outfit and infit values between 0.5 and 1.5 can indicate acceptable fit. However, Bond and Fox (2015) suggest that narrower limits between 0.7 and 1.3 are appropriate. Since the MCOP² ratings are often used for relatively low stakes performance assessments, the MnSq outfit and infit values between 0.5 and 1.5 are considered acceptable.

After all the above assumptions are evaluated for the MFRM modeling, data can be analyzed to address each research question. In all analyses, facets are oriented such that greater logits for examinee ability represent higher ability than lower logits; greater rater logits, however, indicate higher severity level in rating than lower logits, while greater item logits suggest higher difficulty level than lower logits. The average logits of the rater and item facets are centered to 0 so that the average examinee ability measures can be freely estimated.

Analyses Plan for Research Question Two

Research Question 2 (i.e., To what extent does the MCOP² observation protocol separate observed teachers into distinct levels of proficiency?) is addressed by examining the examinee facet in the MFRM analysis.

First, the MFRM analysis conducted using Facets yields (a) a measure of the examinee ability parameter on a logit scale for each individual teacher together with (b) a SE that indicates the uncertainty of (i.e., error associated with) that parameter estimate. The examinee ability measures are examined for the overall range/spread to determine how varied the teachers' teaching practices are based on the MCOP² assessments in this study sample. In addition, the average examinee ability measure can also be calculated as the average proficiency/effectiveness of the observed teachers. A relatively low (close to 0) or even negative mean examinee ability measure would suggest the MCOP² assessment is slightly too difficult for this sample of observed teachers. Whereas a relatively low SE value is desired, as it indicates low measurement errors associated with the examinee ability measures and high level of precision in estimating these measures.

The Separation Index for the examinee facet indicates the number of teaching proficiency levels among the observed math teachers, while the Reliability of Separation indicates the degree to which the MFRM analysis reliably distinguishes between different levels of math teaching proficiency. Fixed χ^2 tests the null hypothesis that all the observed teachers are equal in their math teaching proficiency/effectiveness. Thus, a significant fixed χ^2 value with $p < 0.50$ would indicate that the teachers are not equal in their teaching performances.

A high separation index indicates that the variance among the observed math teachers is substantially larger than the error of estimates; and that the MCOP² ratings are

highly capable of separating the teachers into a number of statistically distinct levels or strata in terms of the math teaching proficiency being measured. A higher reliability statistic indicates that the same ranking of the observed teachers in terms of their teaching proficiency would be more likely to obtain if their classes were to be observed and rated again based on the MCOP² protocol.

In sum, high examinee separation and reliability indices suggest that the assessment distinguishes between examinees in terms of the ability being measured, indicating a high level of replicability of examinee placement across other tasks or tests that measure the same construct (Bond & Fox, 2007). This means greater confidence can be placed in the consistency of score-based inferences.

Analyses Plan for Research Question Three

Research Question 3 (i.e., To what extent do raters differ in terms of the relative severity with which they rate observed teachers?) is evaluated by examining the rater facet in the MFRM analysis.

First, the fixed χ^2 for the rater facet is evaluated as a global test of whether leniency/severity differs across raters. The fixed effect χ^2 is estimated to evaluate the null hypothesis that there are no differences in rater severity after controlling for measurement error. A statistically significant χ^2 ($p < .05$) suggests that at least two raters are statistically significantly different in their leniency/severity measures (Myford & Wolfe, 2004).

Next, the rater separation index and reliability of rater separation are evaluated for the rater facet. The rater separation index is estimated, representing the number of statistically significantly different levels of rater leniency/severity (Myford & Wolfe, 2004). A small rater separation index is desirable, as smaller values indicate fewer

statistically distinct levels of rater leniency/severity compared to larger values (Myford & Wolfe, 2004).

The rater reliability of separation is also estimated for raters, reflecting how reliably raters can be separated along the severity continuum (Myford & Wolfe, 2003). A low rater reliability of separation is desired, suggesting that raters have similar leniency/severity measures and thus cannot be reliably separated along the ability continuum (Myford & Wolfe, 2003; Myford & Wolfe, 2004).

In addition, individual raters' leniency/severity measures are evaluated via visual inspection with a Wright map, also known as a vertical ruler or variable map (Bond & Fox, 2015; Eckes, 2015). The Wright map offers a visual depiction of raters' leniency/severity and the rank-ordering of raters by their leniency/severity measures. Ideally, raters should be clustered close around a logit score of 0 (i.e., average leniency/severity) on the Wright Map. If raters are dispersed across the logit continuum, it suggests that raters differ widely in their level of leniency/severity. Raters with leniency/severity measures greater than 0, and thus located higher above the center logit value at 0 on the Wright map, are regarded more severe than the average rater. By contrast, raters with severity measures less than 0, and thus located lower below the center value at 0 on the Wright map, are considered more lenient than the average rater (Eckes, 2015; Linacre, 2017).

In sum, rater leniency/severity are evaluated overall via the fixed chi square, rater separation index, and rater reliability of separation. Each of these global indices indicate the degree to which raters differ in their leniency/severity. After assessing rater leniency/severity differences globally, individual raters (anonymously coded as Rater 1,

Rater 2, Rater 3, etc.) are then evaluated visually via the Wright map.

Analyses Plan for Research Question Four

Research Question 4 (i.e., To what extent do raters consistently rate the teaching performance of observed teachers?) is evaluated by investigating possible interactions between raters and observed teachers using the MFRM analysis.

Rater fit statistics indicate the degree to which (a) a rater is internally self-consistent across examinees, items, and other factors, and (b) is able to implement the rating scale to make distinctions among examinees' performances (Bond & Fox, 2007; Weigle, 1998). Rater fit statistics close to the expected value of 1.0 suggests that a rater uses the rating scale consistently and thus maintains his/her personal level of severity across examinees, items, and other factors (i.e., intra-rater agreement). By contrast, rater misfit could indicate (a) that the rater exhibits more variation in their ratings than expected, (b) that their ranking of the examinees in terms of their measured latent ability is not reliable, and (c) that they are unable to use the rating scale consistently across items and examinees. The ratings of misfit raters tend to be "noisy", probably due to a tendency to overuse the extreme scale levels. Rater misfit can be detected by evaluating the rater outfit and infit indices for which acceptable values range from 0.5 to 1.5. A rater outfit and/or infit statistics greater than 1.5 would be considered to suggest rater misfit (Linacre, 2002; McNamara, 1996).

By contrast, rater overfit indicates that the rater shows less than expected variation in their ratings, even after controlling for measurement errors. The ratings of overfit raters tend to be "muted", probably because they are being unusually consistent or overly cautious in using the upper and lower levels of the rating scale (i.e., a central

tendency) (McNamara, 1996; Myford & Wolfe, 2000). Rater overfit can also be identified by evaluating the rater outfit and infit indices for which acceptable values range from 0.5 to 1.5. A rater outfit and/or infit statistics less than 0.5 would be considered to suggest rater overfit (Linacre, 2002).

Rater misfit is a more serious threat to general test validity than overfit or examinee misfit because it indicates divergent behavior from the norm on the part of the raters, and its effect on all other facet measure estimates can be strong (Bonk & Ockey, 2003). This is also why Rasch models do not adjust examinee scores/ratings as they can in the case of rater severity (Bonk & Ockey, 2003, p. 101; Myford & Wolfe, 2000, 2004).

Analyses Plan for Research Question Five

Research Question 5 (i.e., To what extent do raters consistently rate the teaching performance of observed teachers across the MCOP² items?) is evaluated by investigating possible interactions between raters and the MCOP² items using the MFRM analysis.

The 4-facet MFRM model (i.e., ratees + MCOP² items + raters + classrooms) is modified to include an interaction term between the rater facet and the MCOP² item facet to evaluate this research question. When evaluating interactions in the MFRM framework (also referred to as bias analysis in Facets language), interactions may be tested in an exploratory or confirmatory manner (Eckes, 2015). Exploratory interaction analyses are appropriate with no a priori hypotheses about the nature of the interactions. Since no a priori hypotheses exist about possible interactions between the MCOP² items and rater leniency/severity, an exploratory interaction analysis is conducted.

MFRM-based bias analysis in Facets investigates whether a particular aspect of

the assessment setting elicits a consistently biased pattern of scores/ratings. As McNamara (1996) put it, “The basic idea in bias analysis is to further analyze the residuals to see if any further sub-patterns emerge.” (p. 141) After estimating the main effects respectively for the rater severity (across all tasks), MCOP² item difficulty (across all raters), and examinee ability (across all items and raters), the MFRM analysis estimates the most likely score for each examinee with a given rater on a specific task, if the rater’s rating behavior remains consistent across all MCOP² items. These individual examinee scores are totaled across all examinees to produce a total *expected* score given by each rater on each item. This *expected* total score is then compared to the *observed* total score for all the examinees on the same item.

If the observed score for a given MCOP² item is higher than the expected score, this item seems to have elicited more lenient behavior than usual on the part of the raters. Fit statistics of the bias analysis summarize for each rater, item, and examinee the extent to which the differences between expected and observed values are within a normal range (expressed in standard deviations from the mean fit statistics).

MFRM-based bias analysis in Facets outputs a file (i.e., Table 13) that provides detailed statistical information to identify significantly biased rater-by-item interactions. Specifically, Table 13 reports the following statistics among others (Kondo-Brown, 2002; Lynch & McNamara, 1998; McNamara, 1996): (a) Observed Score (observed total raw score for this criterion-rater combination), (b) Expected Score (predicted total raw score for this criterion-rater combination), (c) Observed-Expected Average (the average difference between the observed and expected scores), (d) Bias (extent of any discrepancy between the average of the observed and expected values expressed as

logits), (e) Z-score (likelihood of this discrepancy occurring by chance), and (f) Mean Square Fit (fit tells us how consistent this pattern of bias is across all the test-takers involved on this criterion with this rater) (Barkaoui, 2013; Linacre, 2002).

All Z-scores should ideally be equal to zero. Z-score values larger than +2 or less than -2 indicate significantly biased interactions. Positive Z-score values indicate that the rater is more severe on that particular item, while negative z-values suggest that the rater is more lenient when rating that criterion. While with respect to the mean square fit indices for the biased interactions, infit mean square values within the range of two standard deviations around the mean of infit indicate that raters are consistent in the identified patterns of bias across all examinees (Barkaoui, 2013; McNamara, 1996).

McNamara (1996) and Kondo-Brown (2002) both recommend that only biased interactions with Z-values equal to or higher than the absolute value of 2, plus MnSq infit values within the range of two standard deviations around the mean of infit should be considered.

Analyses Plan for Research Question Six

Research Question 6 (i.e., To what extent can the score levels of the MCOP² items be distinguished, without certain score levels being either underused or overused?) is evaluated by examining both the graphic indicators (i.e., Item Characteristic Curves, and Item Information Functions) and the statistical indicators (i.e., item category ordering for individual raters, and rater fit indices).

Scale functioning analysis assesses the quality of the rating scale by examining how the scale levels/categories are functioning and whether the thresholds indicate a hierarchical pattern to the rating scale (Bond & Fox, 2007; Davidson, 1991; North, 2003).

Descriptive statistics such as counts and percentages of scores in each category are first examined. Bond and Fox (2007) suggest that, as a rule of thumb, each category should be assigned to at least 10 ratings/observations to allow scale diagnostics (Linacre, 2003).

Next, Probability Category Curves (PCCs), Item Characteristic Curves (ICCs), Item Information Functions (IFFs), and Category Information Curves (CICs) are also examined to determine possible overuse or underuse of specific categories. For PCCs, thresholds with flat curves are problematic. Davidson (1991) points out that such scale-steps are “operationally worthless” as they are never the most probable rater scale-step choice on any point along overall test-taker ability (p. 159). Thus, he suggests three ways to address this problem: (a) rewriting the level descriptors to clarify what the level is intended to measure, (b) removing that step from the scale if it is not needed, and/or (c) providing rater training to explain the meaning of the underused step.

For CICs, the wider the curves (capturing a wider range of values), the more popular the category would be, signifying overuse. For IIFs, the more dissimilar the shapes (sizes) of curves are, the more evidence there would be that the curves are conveying different amounts of information.

The results of MFRM-based scale functioning analysis in Facets are all included in a Facets output file (i.e., Table 6). Table 6 includes a variety of diagnostic information to examine scale functioning.

For instance, Column 4 in Table 6 reports the (observed) average examinee ability measure associated with each category. This is computed by averaging the examinee ability measures (in logits) for all examinees in the sample who are assigned that

particular score. These measures are expected to increase monotonically in size as the latent ability being measured increases, indicating that, on average, those with higher ability will be assigned to the higher scores (Bond & Fox, 2007; Linacre, 2003). If a score level violates the monotonicity pattern, it will be automatically flagged.

Column 6 reports the outfit mean square index for each category. The expected value of this index is 1.0, indicating that the observed and expected examinee ability measures are equal. The larger the difference between the observed and expected measures, the larger the outfit mean-square index will be. An outfit mean-square index greater than 2.0 suggests that a rating in that level for one or more classroom observations may not be contributing to meaningful measurement of the latent trait (Linacre, 1999).

The last two columns in Table 6 report step- or threshold-calibrations, representing difficulties estimated for choosing one response category over another (Davidson, 1991; Linacre, 2002). Bond and Fox (2007) explain that “threshold distances should indicate that each step defines a distinct position on the variable” and that they should be neither too close together nor too far apart on the logit scale. As a rule of thumb, “thresholds should increase by at least 1.4 logits, to show distinction between, but not more than 5 logits, so as to avoid large gaps in the variable” (Bond & Fox, 2007, p. 163).

Analyses Plan for Research Question Seven

Research Question 7 (i.e., To what extent are the rater behaviors associated with the professional background characteristics (i.e., in-service vs. pre-service teachers, study sites, and teaching grade levels) of the observed teachers?) is evaluated by examining possible interactions between raters and the facets indicating observed teachers’

professional background in the MFRM analysis.

For each of the three external facets (i.e., in-service vs. pre-service teachers, study sites, and teaching grade levels), the original 4-facet MFRM model (i.e., ratees + MCOP² items + raters + classrooms) is modified to include an interaction term between the rater facet and the particular external facet to implement a MFRM-based bias analysis in Facets, respectively.

The data analysis plans for these three MFRM-based bias analyses follows the same procedures and decision-making guidelines as detailed in the previous **Analyses Plan for Research Question Five**.

Ethical Standards

Because this study involved human subjects, the University of Kentucky (UK) Institutional Review Board (IRB) clearance is required. After the approval process was finalized (for acquiring and using the MCOP² secondary data sources), data acquisition and analyses proceeded as described above (see the relevant sections in **CHAPTER III: METHOD**). Written permissions from the original data owners were acquired before any data analyses, and adherence to the rules of privacy safeguarding participant information was followed as required by law.

Protocol for research on human subjects, per the Institutional Review Board (IRB) at the University of Kentucky research department, was strictly followed. The researcher of this study had complied with all requirements related thereto. After permission was gained, the IRB approval letter was filed and approved (see Appendix A).

CHAPTER IV

RESULTS

Introduction

The purpose of this study was to evaluate a math classroom observation protocol (MCOP²) under a Rasch measurement framework for calibrating rater assessment of math teachers' classroom instructional performance, featuring the Many-Facet Rasch Model (MFRM) for rater effects control. Gleason, Zelkowski, and their colleagues (2016, 2017, 2018) conducted several validation studies under the CTT framework for the 16-item Mathematics Classroom Observation Protocol for Practices (MCOP²). Among which, exploratory/confirmatory factor analysis and interrater reliability analysis were performed on the MCOP² raw data for internal structure analysis and rater effects control, respectively. However, the methodological limitations of the CTT approach for rater-mediated assessments were discussed above, such as causing unintended interpretations of a scoring rubric (Eckes, 2008), biased ratings due to power dynamics among raters (Hoyt & Kerns, 1999), or the need for costly and time-consuming training programs that often fail to produce a high degree of rater agreement (Barrett, 2001). Thus, it is highly necessary to use the MFRM modeling technique for furthering the investigation and evaluation of the MCOP² validity and reliability, including dimensionality analysis, item-level analysis, rater effects control, and ratee and rater ability level calibration.

The previous three chapters introduced the key concepts of rater-mediated performance assessment and the Many-Facet Rasch Model (MFRM), reviewed the literature related to earlier MCOP² validation and empirical studies, MFRM modeling,

interrater reliability, and multi-facet calibration techniques for controlling rater effects, and outlined the methodology utilized in the current study. This chapter presents the results of the MFRM-based data analysis that pertain to three parts: (a) evaluation of the overall model fit of the sampled MCOP² rating data under the MFRM framework; (b) examination of the psychometric properties of the rater, ratee, and item facets under the MFRM framework, respectively; (c) contingent upon acceptable overall model fit, investigation of possible interaction bias among the key facets under the MFRM framework (e.g., raters, ratees, and their professional background characteristics). Descriptive statistics are discussed first, followed by the analysis results of each of the seven empirical research questions.

Research Questions

This study investigated the following seven questions:

1. To what extent do the observed rating data obtained from the MCOP2 instrument fit the MFRM modeling? This question is evaluated by testing the MFRM model assumptions, including local independence, unidimensionality, overall model fit, rater fit, and item fit.
2. To what extent does the MCOP2 observation protocol separate observed teachers into distinct levels of proficiency? Such a separation is evaluated by examining the examinee facet in the MFRM analysis.
3. To what extent do raters differ in terms of the relative severity with which they rate observed teachers? This question is evaluated by examining the rater facet in the MFRM analysis.
4. To what extent do raters consistently rate the teaching performance of

observed teachers? This question is evaluated by investigating possible interactions between raters and observed teachers using the MFRM analysis.

5. To what extent do raters consistently rate the teaching performance of observed teachers across the MCOP2 items? This question is evaluated by examining investigating possible interactions between raters and the MCOP2 items using the MFRM analysis.
6. To what extent can the score levels of the MCOP2 items be distinguished, without certain score levels being either underused or overused? This question is evaluated by examining both the graphic indicators (i.e., Item Characteristic Curves, and Item Information Functions) and the statistical indicators (i.e., item category ordering for individual raters, and rater fit indices).
7. To what extent are the rater behaviors associated with the professional background characteristics (i.e., in-service vs. pre-service teachers, schools, and teaching grade levels) of the observed teachers? This question is evaluated by examining possible interactions between raters and the facets indicating observed teachers' professional background in the MFRM analysis.

The results of the statistical analyses related to each of these questions are presented in the order of the research questions. The implications are discussed in the next chapter.

Descriptive Statistics

The population for this study is formed by all pre- and in-service teachers who teach math in P-12 classrooms. The specific study sample drawn from this population is composed of 129 pre- and/or in-service math teachers from the neighboring school

districts around the University of Alabama (i.e., Sample AL) and thirty pre-service math teachers from the University of Kentucky (i.e., Sample UK) whose teaching performances were observed and rated according to the MCOP² rubrics in the P-12 classrooms across the elementary, secondary, and post-secondary levels. All the 159 math teachers in the combined sample were observed and rated by a single rater who had received formal or informal training on how to observe and give scores on the sixteen MCOP² items.

Since Sample AL contains cross-sectional data collected at a single time point, each of the 129 math teachers was observed and rated only once by one of the four AL raters, and each observation record is unique on the rater, ratee, and classroom facet. In contrast, the thirty pre-service math teachers in Sample UK were enrolled in their respective pedagogy courses (e.g., SEM435 or SEM746) near the end of their teacher education program, and their student teaching performances had been observed and rated by their university faculty supervisors (seven raters in total) using the MCOP² protocol three to four times over the time span of about four months. In order to combine Sample AL and Sample UK for overall cross-sectional data analysis, only the chronologically most recent observation record was retained for each of the thirty UK pre-service teachers in the combined sample (For example, a pre-service teacher was observed and rated four times throughout the course of SEM435 on February 1st, February 7th, April 10th, and April 15th, 2019, only his/her MCOP² observation ratings on April 15th, 2019 was retained and included in the combined study sample). Consequently, the final combined study sample contains 2,534 valid responses based on the sixteen MCOP² items, unique for each of the 159 pre- and/or in-service math teachers. Missing data ($n =$

13) accounted for less than 1% of all ratings.

As shown in Table 1, the demographic/background features of the math teachers ($n = 159$) in the combined study sample are detailed by the four demographic variables, namely, Study Site, MCOP² Raters, Classroom Grade Level, and Service Type (i.e., Pre-Service or In-Service). One important difference to note concerns the specification of the classroom grade level for each math teacher between the AL and UK samples: because no information was provided for the thirty pre-service math teachers concerning their classroom grade levels in the original UK sample, the numeric value “99” were filled in for the UK observations to indicate missing or unspecified values.

Table 1

Descriptive Statistics for Demographic Variables in the Combined Sample (N = 159)

Variable	Response	N	%
Study Site	Sample AL	129	81
	Sample UK	30	19
Raters	AL Raters (1-4)	4	36
	UK Raters (5-11)	7	64
Classroom Grade Level ^a	1 (LE: Lower Elementary)	27	17
	2 (UE: Upper Elementary)	13	8
	3 (MS: Middle School)	12	7
	4 (HS: High School)	25	16
	5 (Sec: Secondary)	16	10
	6 (UG: Tertiary)	36	23
	99 (Unspecified)	30	19
Pre- or In-Service	1 (In-Service Teachers)	101	64
	2 (Pre-Service Teachers)	58	36

Notes. ^aRegarding Classroom Grade Level, Sample UK fails to provide any specific information; thus, the numeric value “99” were filled in for the UK observations to indicate missing or unspecified values.

The subscale total ratings on Student Engagement (including nine items) and Teacher Facilitation (including nine items), as well as the MCOP² total scores (including sixteen items) were examined for all the 159 math teachers as a whole and for teachers in each of the two original study samples (i.e., Sample AL and Sample UK), respectively. The full range of ratings, from 0 = *most unsatisfactory performance* to 3 = *most satisfactory performance*, were used for all the items in the above-listed subscales, although the specific descriptors on the four rating levels are unique for each of the sixteen items based on the MCOP² rubric. Table 2 provide a range of descriptive statistics for the raw ratings on the total MCOP² protocol and for each of its two subscales.

Table 2

Descriptive Statistics for Participants' MCOP² Raw Scores in the Respective AL, UK, & Combined Samples

Descriptive Statistics	Sample								
	AL			UK			Combined		
	<i>SE</i> ^a (<i>N</i> = 127)	<i>TF</i> ^b (<i>N</i> = 127)	<i>Total</i> ^c (<i>N</i> = 125)	<i>SE</i> (<i>N</i> = 28)	<i>TF</i> (<i>N</i> = 29)	<i>Total</i> (<i>N</i> = 28)	<i>SE</i> (<i>N</i> = 155)	<i>TF</i> (<i>N</i> = 156)	<i>Total</i> (<i>N</i> = 153)
<i>Mean</i>	1.55	1.56	1.56	1.69	1.33	1.51	1.57	1.52	1.55
<i>Median</i>	1.56	1.56	1.50	1.67	1.33	1.50	1.56	1.44	1.50
<i>Mode</i>	1.56	1.33	1.00	1.67	1.33	1.50	1.56	1.33	1.00
<i>SD</i>	.72	.62	.60	.48	.48	.42	.69	.60	.57
<i>Min</i>	.33	.44	.44	.78	.33	.69	.33	.33	.44
<i>Max</i>	3.00	3.00	3.00	2.56	2.33	2.44	3.00	3.00	3.00
<i>Range</i>	2.67	2.56	2.56	1.78	2.00	1.75	2.67	2.67	2.56

Notes. ^athe total score of the Student Engagement Subscale; ^bthe total score of the Teacher Facilitation Subscale; ^cthe average score of all 16 items in the MCOP² instrument.

As shown in Table 2, comparing the raw ratings of Sample AL and Sample UK

math teachers' on the two MCOP² subscales (i.e. Student Engagement and Teacher Facility), three tendencies were worth noting: (a) the pre-service math teachers in Sample UK were rated significantly lower on Teacher Facilitation than the math teachers in Sample AL ($M\ diff = .236, t = 2.251, p = .029, Cohen's\ d = .395$); (b) on average, the pre-service math teachers in Sample UK received much lower ratings on Teacher Facilitation ($M = 1.33, SD = .48$) than their ratings on Student Engagement ($M = 1.69, SD = .48$); and (c) the raw ratings of the pre- or in-service teachers in Sample AL were about equal on the two subscales of Teacher Facilitation ($M = 1.56, SD = .62$) and Student Engagement ($M = 1.55, SD = .72$).

These differences in the mean comparisons of the MCOP² raw scores may lead to interesting interpretations from the psychometric perspective: if the MCOP² protocol is deemed valid and reliable, the significant differences between the Teacher Facilitation ratings received by the pre-service math teachers in Sample UK and those received by the math teachers in Sample AL (78% are in-service math teachers) may reflect the extent to which the MCOP² protocol can distinguish math teachers' true levels of teaching effectiveness across study samples. However, because the CTT approach of calculating interrater reliability is sample sensitive and cannot effectively control for various rater effects, the possibility cannot be eliminated that such mean differences may be attributed to differences in rater severity/leniency across study samples.

Analyses for Research Question One

Research Question 1 (i.e., To what extent do the observed rating data obtained from the MCOP² instrument fit the MFRM modeling?) was evaluated by testing the MFRM modeling assumptions, including local independence, unidimensionality, overall

model fit, rater fit, and item fit.

Local Independence

Local independence (LID) refers to the assumption that item responses are independent from one another after controlling for the construct of interest (DeMars, 2010). This LID assumption is, however, almost always violated to various extents in empirical applications. The most widely used method for identifying LD is based on Yen's Q_3 (1984, 1993) statistics through computing item residuals (observed item responses minus their expected values), and then correlating these residuals.

The Yen's Q_3 (1984, 1993) statistics for the MCOP² data used in this study can be calculated and investigated as part of the Principal Component Analysis of Residuals (PCAR) conducted in the Winsteps software program, version 4.7.0 (Linacre, 2020), where Table 23.99 (i.e., Largest residual correlations for items) can be obtained for pairwise, item-level residual correlations by specifying the command of "PRCOMP = R" in the control file.

Simulations show that any item residual correlation that is 0.2 above the average residual correlation would appear to indicate LD, and any residual correlation of independent items that is 0.3 above the average correlation would seem unlikely (Christensen, Makransky, & Horton, 2017; Marais, 2013).

Table 3 below shows that (a) the 16-item MCOP² scale indicate serious LD issues, with 5 pairs of item residual correlation well above the average residual correlation (.24 to .38 above the average Q_3 .17); (b) the 9-item Student Engagement subscale suggests slight LD problems with 3 pairs of item residual correlation notably above the average residual correlation (.21 to .26 above the average Q_3 .11); and (c) no

LD-related concerns are identified for the 9-item Teacher Facilitation subscale where none of the pairs of item residual correlation is 0.2 above the average Q_3 .11.

Table 3

Summary of the Local Independence (Yen's Q_3) Statistics for the MCOP² Protocol, Student Engagement Subscale, and Teacher Facilitation Subscale

Yen's Q_3	MCOP ² Scale (16 items)		Yen's Q_3	Student Engagement (9 items)		Yen's Q_3	Teacher Facilitation (9 Items)	
<i>Avg Q_3 = .17</i>	<i>Pairs of Correlated Items</i>		<i>Avg Q_3 = .11</i>	<i>Pairs of Correlated Items</i>		<i>Avg Q_3 = .10</i>	<i>Pairs of Correlated Items</i>	
.55**	Item 1	Item 5	.37*	Item 3	Item 12	.26	Item 11	Item 16
.50**	Item 3	Item 12	.36*	Item 1	Item 5	.22	Item 13	Item 16
.45*	Item 12	Item 15	-.32*	Item 14	Item 15	-.19	Item 4	Item 6
-.42*	Item 3	Item 6	.25	Item 12	Item 15	.19	Item 8	Item 11
.41*	Item 1	Item 15	-.22	Item 2	Item 5	-.18	Item 10	Item 13

Note. *.0.2 to 0.3 above the average item residual correlation; **.0.3 or greater above the average item residual correlation.

Unidimensionality

Unidimensionality is related to local independence and refers to the assumption that all assessment items measure only one, common construct (Bandalos, 2018; DeMars, 2010). Unidimensionality is evaluated by conducting a Principal Components Analysis (PCA) on the standardized residuals (PCAR) following the analysis of a basic 4-facet MFRM analysis (i.e., ratees + MCOP² items + raters + classrooms) in Facets. The PCAR was conducted using the Winsteps software program, version 4.7.0 (Linacre, 2020).

It is assumed that all items should be loaded on the first contrast of the Rasch dimension, and the PCAR specifically tests whether any items group on secondary contrasts. Each contrast has an associated eigenvalue, and the eigenvalues represent the number of items that make up the respective contrast. If eigenvalues for secondary contrasts are less than 2.0 (indicating there are fewer than two elements on the secondary contrasts), the unidimensionality assumption is met.

However, if eigenvalues for any of the secondary contrasts are greater than 2.0, it is recommended to further examine the disattenuated correlations between the person measures on the suspect cluster of items and the person measures on the other items. If the correlations are greater than 0.70, the suspect cluster of items is probably only measuring a secondary strand of the main Rasch dimension and should not be considered as a different dimension. By contrast, disattenuated correlations less than 0.30 or even negative values indicate that the suspect cluster of items is measuring something different than the construct of interests, and multidimensionality may become an issue (Linacre, 2012).

Table 4

Summary of the PCA Statistics for the MCOP² Protocol, Student Engagement Subscale, and Teacher Facilitation Subscale

PCA Statistics	MCOP ² Scale (16 items)			Student Engagement (9 items)			Teacher Facilitation (9 Items)		
	Eigenvalue	Observed	Expected	Eigenvalue	Observed	Expected	Eigenvalue	Observed	Expected
<i>Total Variance</i>	28.99	100%	100%	21.30	100%	100%	17.92	100%	100%
<i>Variance by Measure</i>	12.99	44.8%	44.6%	12.30	57.7%	57.7%	8.92	49.8%	49.2%
<i>Variance by Persons</i>	5.77	19.9%	19.8%	7.03	33.0%	33.0%	4.52	25.2%	24.9%
<i>Variance by Items</i>	7.22	24.9%	24.8%	5.27	24.7%	24.7%	4.40	24.5%	24.2%
<i>Total Unexplained Variance</i>	16.00	55.2%	55.4%	9.00	42.3%	42.3%	9.00	50.2%	50.8%
<i>Unexplained 1st Contrast</i>	3.75*	12.9%	-	2.01*	9.4%	-	1.66	9.3%	-
<i>Unexplained 2nd Contrast</i>	1.67	5.8%	-	1.56	7.3%	-	1.40	7.8%	-
<i>Unexplained 3rd Contrast</i>	1.51	5.2%	-	1.24	5.8%	-	1.23	6.9%	-
<i>Unexplained 4th Contrast</i>	1.25	4.3%	-	1.19	5.6%	-	1.20	6.7%	-
<i>Unexplained 5th Contrast</i>	1.12	3.9%	-	1.08	5.1%	-	1.02	5.7%	-

Note. *Eigenvalues for secondary contrasts are greater than 2.0, indicating there are more than two elements on the secondary contrasts, and the unidimensionality assumption is violated.

As illustrated in Table 4 above, the PCA results showed that (a) the 16-item MCOP² protocol used as one single scale failed to uphold the unidimensionality assumption, as the residual variances of more than two items (eigenvalue = 3.75) clustered on a different dimension in addition to the variances explained by the MCOP² measure; (b) the unidimensionality assumption was better met for the 9-item Student Engagement subscale, with just about two item residuals loaded on a dimension other than the latent trait measured (eigenvalue = 2.01); and since less than two item residuals (eigenvalue = 1.66) were strongly correlated to form any contrast/factor apart from the variances explained by the measure, the 9-item Teacher Facilitation subscale successfully met the unidimensionality assumption.

To further investigate the unidimensionality issues revealed in the above PCA analyses, the disattenuated correlations were also examined between the person measures on the suspect cluster of items and the person measures on the other items for the 16-item MCOP² protocol and the 9-item Student Engagement subscale, respectively. It was found that for the 16-item MCOP² protocol, the person measure disattenuated correlations between the 1st and 3rd cluster of items fell between the cut-off value range of 0.30 - 0.70 ($r = 0.48$), suggesting that the cluster of items on the suspect 1st contrast were measuring a secondary strand of the main Rasch dimension probably warranting separate investigation. However, for the 9-item Student Engagement subscale, all the disattenuated correlations were well above the upper bound of the cut-off value range (0.70), indicating that the suspect cluster of items was only measuring an insignificant secondary strand of the latent trait of interests and should not be considered as a different dimension (Linacre, 2012).

Overall Model Fit

To evaluate the overall model fit of the MFRM analysis, the absolute values of the standardized residuals were examined for the 9-item Student Engagement subscale and the 9-item Teacher Facilitation subscale, respectively. In this study, data are deemed to have good overall model-fit in the MFRM analysis, if fewer than 5% of the standardized residuals appear greater than or equal to $|2.0|$ and about 0.3% or less of standardized residuals are greater than or equal to $|3.0|$ (Linacre, 2004).

In the 9-item Student Engagement subscale dataset analyzed with the MFRM, there were a total of 1,423 valid responses, of which 58 (4.08%) were associated with (absolute) standardized residuals greater than or equal to 2, and 4 responses (0.28%) associated with (absolute) standardized residuals of greater than or equal to 3. Whereas based on the MFRM analysis of the 9-item Teacher Facilitation subscale dataset, among the total 1,428 valid responses, 50 (3.50%) were associated with (absolute) standardized residuals greater than or equal to 2, and 6 responses (0.42%) associated with (absolute) standardized residuals of greater than or equal to 3. Taken together, these results were indicative of a satisfactory overall model fit for both subscales.

Additional methods for assessing the fit of the MFRM to the MCOP² data (e.g., rater fit statistics) were presented later in the following sections.

Rater Fit and Item Fit

Mean Square outfit and Mean Square infit statistics (also referred to as MSU and MSW) were calculated and investigated (Bond & Fox, 2015; Eckes, 2005; Engelhard, 1994, 2002; Myford & Wolfe, 2003) to evaluate rater fit and/or item fit (Bond & Fox, 2015; Eckes, 2005; Engelhard, 1994, 2002; Myford & Wolfe, 2003). Linacre (2003)

proposes that outfit and infit values between 0.5 and 1.5 can indicate acceptable fit. Since the MCOP² ratings are often used for relatively low stakes performance assessments, the MnSq outfit and infit values between 0.5 and 1.5 were considered acceptable. The overfitting raters would have muted ratings that suggested a central tendency or, alternatively, a halo effect (see Engelhard, 2002; Myford & Wolfe, 2004). While the underfitting raters would suggest their ratings show off-target deviations/noise from the way the measure is intended to be used, and thus are unproductive (or even degrading in case of a serious extent of rater underfit) for construction of measurement.

Table 5.1

Percentages of Rater Mean-Square Fit Statistics for the Student Engagement Subscale and the Teacher Facilitation Subscale

Fit Range	Student Engagement		Teacher Facilitation	
	Infit	Outfit	Infit	Outfit
fit < 0.50 (overfit)	9%	9%	0%	0%
0.50 ≤ fit ≤ 1.50	73%	73%	91%	91%
fit > 1.50 (misfit)	18%	18%	9%	9%

Table 5.1 above displayed percentages of rater fit values falling into overfit, acceptable fit, and misfit categories, using the relatively wide range of upper and lower control limits ($0.50 \leq \text{fit} \leq 1.50$). Based on the infit values, one rater showed overfit (9%) and two raters fell into the misfit category (18%), when using the 9-item Student Engagement subscale. Since the percentage of raters showing acceptable fit (73%) for using the Student Engagement subscale fell well below 90%, it was concluded that the

raters were internally inconsistent in using the 4-point rating scale, or the raters might not have used the Student Engagement rating scale appropriately.

In comparison, when using the 9-item Teacher Facilitation subscale, more raters fell into the desirable category between 0.50 and 1.50 (91%) meaning that the number of overfitting and underfitting raters was minimal (only one out of the eleven raters showed underfit). Since the percentage of raters showing acceptable fit was above 90%, it could be concluded that the raters were internally consistent and used the 4-point Teacher Facilitation rating scale appropriately.

Table 5.2

Percentages of Item Mean-Square Fit Statistics for the Student Engagement Subscale and the Teacher Facilitation Subscale

Fit Range	Student Engagement		Teacher Facilitation	
	Infit	Outfit	Infit	Outfit
fit < 0.50 (overfit)	0%	0%	0%	0%
0.50 ≤ fit ≤ 1.50	89%	100%	89%	100%
fit > 1.50 (misfit)	11%	0%	11%	0%

Table 5.2 above displayed percentages of item fit values falling into overfit, acceptable fit, and misfit categories, using the relatively wide range of upper and lower control limits ($0.50 \leq \text{fit} \leq 1.50$). Based on the infit and outfit values, only one item showed underfit (11%) for both the 9-item Student Engagement subscale and the 9-item Teacher Facilitation subscale. Since the percentage of items showing acceptable fit (89%) for both the subscale were very close to 90%, it could be concluded that the nine

items on either the Student Engagement or the Teacher Facilitation subscale were internally consistent and can be used to measure the latent traits of interests appropriately.

Analyses for Research Question Two

Research Question 2 (i.e., To what extent does the MCOP² observation protocol separate observed teachers into distinct levels of proficiency?) was addressed by examining the examinee facet in the MFRM analysis.

Figures 3.1 and 3.2 displayed variable maps (also referred to as Wright maps) visualizing the calibrations of raters, ratees, items, and the 4-point rating scales for the Student Engagement and Teacher Facilitation data, respectively. The item facets had been centered and therefore, constrained to have a mean element of zero. However, the measures for the ratee facets were freed to float because extreme values had been included in the analyses (both maximum-possible and minimum-possible scores) as suggested by Linacre (2011). It should be noted, though, that the extreme scores would make no difference to the estimates of the other elements, or to their fit statistics, and are usually preferred to be included in an analysis.

Figures 3.1 and 3.2 below illustrated the Wright maps with calibrated rater, ratee, item, and rating scale facets for the Student Engagement and Teacher Facilitation subscales, respectively. It revealed that (a) for the Student Engagement subscale, the variability across ratees in their level of proficiency seemed substantial, with their proficiency estimates forming a 7.61-logit range (-2.06 ~ 5.55); and (b) the similar pattern was also present for the Teacher Facilitation subscale, with the ratees' level of proficiency falling into the logit range between -2.35 and 5.39.

Measr	Math Teachers	Raters	Items	Scale
6	.	+	+	(3)
	.			
5	+	+	+	
	.			
4	+	+	+	
	*			
	.			
3	***.	+	+	
	.			
	*			
	****.			---
2	+	+	+	
	**			
	*			
	*			
	***		SE_Item5	
1	***.	+	SETF_Item4	2
	**	UA Rater4		

	*****.	UK Rater2	SE_Item2	
	*****.	UK Rater6		
* 0 *	***.	* UA Rater1 UK Rater1	* SE_Item15	* --- *
	****	UK Rater3 UK Rater4 UK Rater7	SE_Item1 SE_Item12	
	**.	UA Rater2 UA Rater3 UK Rater5	SE_Item14	
	***.		SETF_Item13	

-1	*****.	+	+	1

	**		SE_Item3	

	.			
-2	+	+	+	

-3	+	+	+	(0)

Figure 3.1
The Student Engagement Subscale Wright Map

Measr	Math Teachers	Raters	Items	Scale
5	.			(3)
	*			
4	*			
	.			
3				
	*			
	.			---
	.			
2	.			
	*			
	*			

	**			
	**.			
1	***.	UK Rater2	TF_Item7	2
	*			
	****		SETF_Item4	
	*.	UK Rater7		
	****		TF_Item8 TF_Item9	
	**.	UK Rater1 UK Rater3 UK Rater4		
0	****.	UA Rater1 UA Rater3 UA Rater4		---
	*****.	UA Rater2	TF_Item11 TF_Item16	
	*****	UK Rater5 UK Rater6	TF_Item6	
	*****.			
	*		SETF_Item13	

-1	****.			1
	****.		TF_Item10	
	*			

	*.			
-2				
	.			---
	*			
-3				(0)

Figure 3.2

The Teacher Facilitation Subscale Wright Map

These findings were further supported by the statistics given in Tables 6.1 and 6.2. Specifically, for the Student Engagement data, the standard deviation of the estimated proficiencies of the ratees was 1.46. The *RMSE* value for the ratee proficiency estimates was 0.55 (the highest of all three facets), indicating that these ratee measures were estimated with a relatively high error component. Comparing the ratee facet with the other two facets, the item facet showed the lowest *RMSE* (0.12), indicating that these item measures were estimated with a particularly low error component. Such a difference was probably because the estimation of the item measures was based on a much larger number of observations. It could also explain why the item facet received the highest value (out of the 3 facets) of separation ratio ($G = 6.04$, as compared to 0.49 and 2.35 for rater and ratee facets respectively). The chi-square statistic testing the hypothesis that all ratees had the same proficiency was highly significant, suggesting that all ratees did not share the same proficiency level (after allowing for measurement error). The ratee separation index (H) estimates that, within this sample of ratees there were about 3 (3.47) statistically distinct strata of proficiency. The separation ratio (G) for the ratee facet was 2.35, indicating that the true standard deviation of ratee proficiency measures was about 2 times greater than their standard error of measurement. The separation reliability of the ratee proficiency estimates was 0.85. For ratees, this reliability statistic provided information about how well one could differentiate among the ratees in terms of their levels of proficiency. The ratee separation reliability indicated how different the ratee proficiency measures were. Since the purpose of most performance assessments is to differentiate ratees in terms of their proficiency as well as possible, a high value of separation statistic is desired. It seemed that for Student Engagement, this statistic was

relatively high, implying that the Student Engagement subscale could differentiate well among the ratees (i.e., math teachers) in terms of their classroom teaching performance to engage students.

Table 6.1

Summary of the MFRM Analysis Statistics for the Student Engagement Subscales

MFRM Statistics	Rater	Ratee	Item
<i>M</i> (measure)	-.02	.26	0.00 ^a
<i>SD</i> (measure)	.36	1.46	.76
<i>M SE</i>	.10	.30	.01
<i>RMSE</i>	.32	.55	.12
Adj. (true) <i>SD</i>	.16	1.29	.75
χ^2	124.1***	794.5***	320.4***
<i>df</i>	10	158	8
Separation ratio (<i>G</i>)	.49	2.35	6.04
Separation (strata) index (<i>H</i>)	.99	3.47	8.39
Separation reliability (<i>R</i>)	.19	.85	.97

Note. ^aThe item facet was constrained to have a mean estimate of zero for the Rasch model-based analysis. *M SE* = Mean-square measurement error. *RMSE* = Root mean-square measurement error. *** $p < .001$.

Similarly, as shown in Table 6.2 below, for the Teacher Facilitation data, the standard deviation of the estimated proficiencies of the ratees was 1.37. The *RMSE* value for the ratee proficiency estimates was 0.52 (the highest of all three facets), indicating that these ratee measures were estimated with a relatively high error component. The chi-square statistic testing the hypothesis that all ratees had the same proficiency was highly significant, suggesting that all ratees did not share the same proficiency level (after allowing for measurement error). The ratee separation index (*H*) estimates that, within this sample of ratees there were about 3 (3.39) statistically distinct strata of proficiency.

The separation ratio (G) for the ratee facet was 2.29, indicating that the true standard deviation of ratee proficiency measures was about 2 times greater than their standard error of measurement. The separation reliability of the ratee proficiency estimates was 0.84. For ratees, this reliability statistic provided information about how well one could differentiate among the ratees in terms of their levels of proficiency. The ratee separation reliability indicated how different the ratee proficiency measures were. It seemed that for Teacher Facilitation, this statistic was relatively high, implying that the Teacher Facilitation subscale could differentiate well among the ratees (i.e., math teachers) in terms of their teacher facilitation performance.

Table 6.2

Summary of the MFRM Analysis Statistics for the Teacher Facilitation Subscale

MFRM Statistics	Rater	Ratee	Item
M (measure)	.12	.09	0.00 ^a
SD (measure)	.37	1.37	.63
$M SE$.08	.27	.01
$RMSE$.28	.52	.12
Adj. (true) SD	.24	1.19	.61
χ^2	27.4***	730.1***	244.6***
df	10	158	8
Separation ratio (G)	.85	2.29	5.22
Separation (strata) index (H)	1.47	3.39	7.29
Separation reliability (R)	.42	.84	.96

Note. ^aThe item facet was constrained to have a mean estimate of zero for the Rasch model-based analysis. $M SE$ = Mean-square measurement error. $RMSE$ = Root mean-square measurement error. *** $p < .001$.

Analyses for Research Question Three

Research Question 3 (i.e., To what extent do raters differ in terms of the relative severity with which they rate observed teachers?) was evaluated by examining the rater facet in the MFRM analysis.

Rater leniency/severity were evaluated overall via the fixed chi square, rater separation index, and rater reliability of separation. Each of these global indices indicated the degree to which raters differ in their leniency/severity. After assessing rater leniency/severity differences globally, individual raters (anonymously coded as Rater 1, Rater 2, Rater 3, etc.) were then evaluated visually via the Wright maps.

Examining the Wright maps in Figures 3.1 and 3.2, it was noted that the variability across raters in the level of the severity with which items were rated was not substantial for both the Student Engagement and Teacher Facilitation subscale. The rater severity estimates showed a relatively narrow spread of 1.24 logits and 1.36 logits for Student Engagement and Teacher Facilitation, respectively. This finding was also supported by examining the relevant statistics in Tables 6.1 and 6.2. The mean estimated severity of all the eleven raters was -0.02 with a standard deviation of 0.36 for Student Engagement and 0.12 with a standard deviation of 0.37 for Teacher Facilitation. (The rater facets were not centered and therefore, were not constrained to have a mean element measure of 0.) The *RMSE* values were 0.32 and 0.28 for Student Engagement and Teacher Facilitation respectively, indicating that these rater measures were estimated with a relatively low error component.

The fixed chi-square statistics testing the hypothesis that all raters have the same severity were highly significant for both the subscales, indicating that at least two raters

were statistically significantly different in their leniency/severity measures (Myford & Wolfe, 2004). However, the rater separation indices (H) showed that within this group of eleven raters there was only 1 (0.49 for Student Engagement and 0.85 for Teacher Facilitation) statistically distinct strata of severity. The separation ratios (G) for the rater facets were 0.99 for Student Engagement and 1.47 for Teacher Facilitation, indicating that the true standard deviations of rater severity measures were only about 1 time greater than their standard error of measurement. The reliability statistics of rater separation also attested to a relatively low dissimilarity degree in rater severity (0.19 for Student Engagement and 0.42 for Teacher Facilitation). Low rater separation reliability (such as in this study) is generally desirable as this would indicate that raters were approaching the ideal of being interchangeable. The rater separation reliability should not be confused with interrater reliability (which is the index of how similar raters are with respect to their severity). Rater separation reliability is an index of how different severity measures are based on Rasch modeling. The results of this study showed that the estimated mean severity of the Sample AL raters was 0.035 compared with the mean severity for Sample UK raters of -0.059 when using the Student Engagement subscale, indicating that Sample AL raters were about 0.1 logit more severe than the Sample UK raters (e.g., the most severe rater on Student Engagement was identified as Rater 4 from Sample AL). In contrast, when using the Teacher Facilitation subscale, the estimated mean severity of the Sample AL raters was -0.042 compared with the mean severity for Sample UK raters of 0.209, suggesting that Sample AL raters were about 0.3 logit more lenient than the Sample UK raters (e.g., the most severe rater on Teacher Facilitation was identified as Rater 2 from Sample UK).

Analyses for Research Question Four

Research Question 4 (i.e., To what extent do raters consistently rate the teaching performance of observed teachers?) was evaluated by investigating possible interactions between raters and observed teachers (i.e., rater fit indices) using the MFRM analysis.

Rater fit statistics indicate the degree to which (a) a rater is internally self-consistent across examinees, items, and other factors, and (b) is able to implement the rating scale to make distinctions among examinees' performances (Bond & Fox, 2007; Weigle, 1998). Rater fit statistics close to the expected value of 1.0 suggests that a rater uses the rating scale consistently and thus maintains his/her personal level of severity across examinees, items, and other factors (i.e., intra-rater agreement).

FACETS program provides two types of mean-square statistics that are indicative of data-model fit for each rater, namely, rater infit and rater outfit. The infit statistic is usually sensitive to an accumulation of unexpected ratings. On the other hand, the outfit statistic is sensitive to individual unexpected ratings. Both the infit and the outfit statistics can range from 0 to infinity and have an expected value of 1 (Linacre, 2002; Myford & Wolfe, 2003).

Rater misfit (i.e., judged upon the rater outfit statistics) is considered a more serious threat to general test validity than rater overfit (i.e., judged upon the rater infit statistics) or examinee misfit because it indicates divergent behavior from the norm on the part of the raters, and its effect on all other facet measure estimates can be strong (Bonk & Ockey, 2003).

Referring back to Table 5.1 in the previous **Analyses for Research Question One** which displayed percentages of rater fit values falling into overfit, acceptable fit,

and misfit categories, one rater showed overfit (9%) and two raters fell into the misfit category (18%), when using the 9-item Student Engagement subscale. Since the percentage of raters showing acceptable fit (73%) for using the Student Engagement subscale fell well below 90%, it was concluded that the raters were internally inconsistent in using the 4-point rating scale, or the raters might not have used the Student Engagement rating scale appropriately. All misfitting and/or overfitting raters were identified coming from Sample UK as Rater 7, 1 and 3.

In comparison, when using the 9-item Teacher Facilitation subscale, more raters fell into the desirable category between 0.50 and 1.50 (91%) meaning that the number of overfitting and underfitting raters was minimal (only one out of the eleven raters showed underfit). Since the percentage of raters showing acceptable fit was above 90%, it could be concluded that the raters were internally consistent and used the 4-point Teacher Facilitation rating scale appropriately. The one misfitting rater was identified as Rater 1 from Sample UK.

Analyses for Research Question Five

Research Question 5 (i.e., To what extent do raters consistently rate the teaching performance of observed teachers across the MCOP² items?) was evaluated by investigating possible interactions between raters and the MCOP² items using the MFRM analysis.

Since no a priori hypotheses exist about possible interactions between the MCOP² items and rater leniency/severity, an exploratory interaction analysis was conducted. MFRM-based bias analysis in FACETS investigates whether a particular aspect of the assessment setting elicits a consistently biased pattern of scores/ratings. As McNamara

(1996) put it, “The basic idea in bias analysis is to further analyze the residuals to see if any further sub-patterns emerge.” (p. 141)

If the observed score for a given MCOP² item is higher than the expected score, this item seems to have elicited more lenient behavior than usual on the part of the raters. Fit statistics of the bias analysis summarize for each rater, item, and examinee the extent to which the differences between expected and observed values are within a normal range (expressed in standard deviations from the mean fit statistics).

MFRM-based bias analysis in Facets outputs a file (i.e., Table 13) that provides detailed statistical information to identify significantly biased rater-by-item interactions. Specifically, Table 13 reports the following statistics among others (Kondo-Brown, 2002; Lynch & McNamara, 1998; McNamara, 1996): (a) Observed Score (observed total raw score for this criterion-rater combination), (b) Expected Score (predicted total raw score for this criterion-rater combination), (c) Observed-Expected Average (the average difference between the observed and expected scores), (d) Bias (extent of any discrepancy between the average of the observed and expected values expressed as logits), (e) *Z*-score (or *t* statistics) (likelihood of this discrepancy occurring by chance), and (f) Mean Square Fit (fit tells us how consistent this pattern of bias is across all the test-takers involved on this criterion with this rater) (Barkaoui, 2013; Linacre, 2002).

All *Z*-scores (or *t* statistics) should ideally be equal to zero. *Z*-score values (or *t* statistics) larger than +2 or less than -2 indicate significantly biased interactions. Positive *Z*-score values (or *t* statistics) indicate that the rater is more severe on that particular item, while negative *Z*-values (or *t* statistics) suggest that the rater is more lenient when rating that criterion. While with respect to the mean square fit indices for the biased

interactions, infit mean square values within the range of two standard deviations around the mean of infit indicate that raters are consistent in the identified patterns of bias across all examinees (Barkaoui, 2013; McNamara, 1996).

McNamara (1996) and Kondo-Brown (2002) both recommend that only biased interactions with Z -values equal to or higher than the absolute value of 2, plus MnSq infit values within the range of two standard deviations around the mean of infit should be considered.

To investigate whether each rater maintained a uniform level of severity across the nine items on the Student Engagement subscale, or whether particular raters gave ratings on some items more severely or leniently than expected, a two-way interaction analysis of Raters by Items was performed. Similarly, interaction analyses (i.e., Raters by Sites, Raters by Service Types, and Raters by Grade Levels) to test for patterns of unexpected ratings related to particular study sites, service types, and classroom grade levels were also performed.

Table 7.1 below listed the total number of combinations of facet elements considered in each interaction analysis, the percent of absolute t -scores equal to or greater than 2, minimum and maximum t -values along with their degrees of freedom, the means and standard deviations of the bias sizes, fixed chi-square statistics, as well as the percentages of variances in the Student Engagement data explained by the bias terms.

Regarding the rater by item interaction, about one fifth of the combinations (21.28%) yielded statistically significant t -scores. This means that some raters tended to alternate between more severe ratings on one item and more lenient ratings on another item. Furthermore, for these significantly biased interactions, the majority of the

associated infit mean square values fell within the range of two standard deviations around the mean of infit, indicating that raters appeared consistent in the identified patterns of bias across all ratees (Barkaoui, 2013). This relatively high percentage of significant rater by item biased interactions altogether contributed to 8.75% of the total raw variances in the Student Engagement data.

Table 7.1

Summary Statistics of the Interaction Analysis for the Student Engagement Subscale

Statistic	Rater by Item	Rater by Site	Rater by Service Type	Rater by Grade Level
<i>N</i> combinations	94	11	12	0
% large <i>t</i> -scores ^a	21.28	0.00	0.00	-
Min- <i>t</i> (<i>df</i>)	-3.46(35)**	-.02(303)	-.02(107)	-
Max- <i>t</i> (<i>df</i>)	4.74(35)***	.01(359)	.02(203)	-
<i>M</i>	-.02	0.00	0.00	-
<i>SD</i>	94	0.00	0.00	-
χ^2 (<i>df</i>)	274.00(94)***	0.00(11)	0.00(12)	-
<i>Variance by Bias</i>	8.75%	0.00%	0.00%	-

Note. ^aPercentage of absolute *t*-scores equal or greater than 2.00

Figure 4.1 below plotted the individual rater by item biased interactions. As highlighted in yellow, five significant biased interactions (i.e., rater absolute measure equal to or greater than 2 logits above the mean rater measure) were noted involving Rater 1, Rater 2, and Rater 7 from Sample UK, as well as Rater 2 and Rater 4 from Sample AL. Specifically, Raters 1, 2, and 7 from Sample UK and Rater 4 from Sample AL rated more leniently than expected on SE_Item3, SETF_Item4, SE_Item2, and SE_Item5, respectively; while Rater 2 from Sample AL rated more severely than expected on SE_Item3.

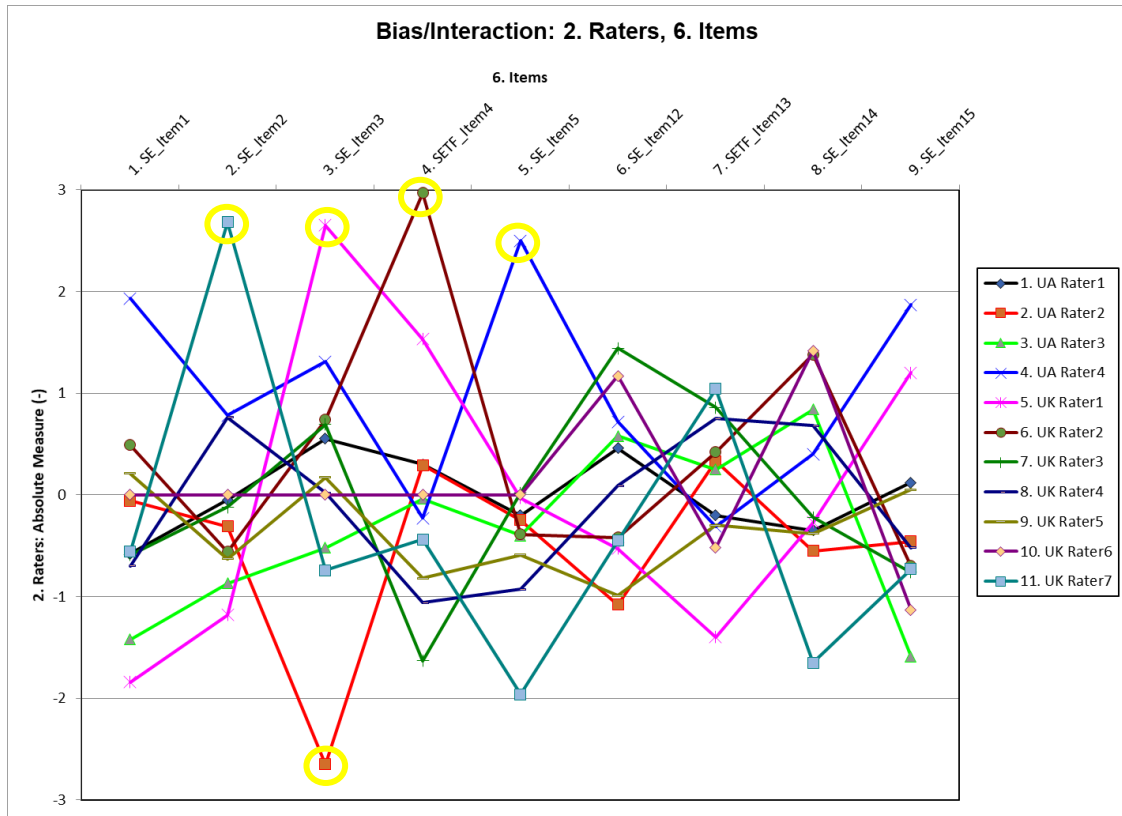


Figure 4.1

Plot Illustrating the Rater by Item Bias Interactions for the Student Engagement Subscale

Similarly, Table 7.2 below listed the total number of combinations of facet elements considered in each interaction analysis, the percent of absolute t -scores equal to or greater than 2, minimum and maximum t -values along with their degrees of freedom, the means and standard deviations of the bias sizes, fixed chi-square statistics, as well as the percentages of variances in the Teacher Facilitation data explained by the bias terms.

Regarding the rater by item interaction, a relatively lower percentage of the combinations (13.27%) yielded statistically significant t -scores compared to those produced for the Student Engagement subscale, suggesting that fewer raters tended to alternate between more severe ratings on one item and more lenient ratings on another item. Furthermore, for these significantly biased interactions, the majority of the

associated infit mean square values fell within the range of two standard deviations around the mean of infit, indicating that raters appeared consistent in the identified patterns of bias across all ratees. The slightly lower percentage of significant rater by item biased interactions still contributed to 7.06% of the total raw variances in the Teacher Facilitation data.

Table 7.2

Summary Statistics of the Interaction Analysis for the Teacher Facilitation Subscale

Statistic	Rater by Item	Rater by Site	Rater by Service Type	Rater by Grade Level
<i>N</i> combinations	98	11	12	0
% large <i>t</i> -scores ^a	13.27	0.00	0.00	-
Min- <i>t</i> (<i>df</i>)	-3.16(15)**	-.01(321)	-.02(106)	-
Max- <i>t</i> (<i>df</i>)	3.80(12)**	0.00(62)	.00(62)	-
<i>M</i>	-.03	0.00	0.00	-
<i>SD</i>	.96	0.00	0.00	-
χ^2 (<i>df</i>)	162.8(98)***	0.00(11)	0.00(12)	-
<i>Variance by Bias</i>	7.06%	0.00%	0.00%	-

Note. ^aPercentage of absolute *t*-scores equal or greater than 2.00

Figure 4.2 below plotted the individual rater by item biased interactions for Teacher Facilitation. Highlighted in yellow, nine significant biased interactions (i.e., rater absolute measure equal to or greater than 2 logits above the mean rater measure) were noted involving Rater 1, Rater 2, Rater 6, and Rater 7 from Sample UK. Specifically, Raters 2, 6, and 7 from Sample UK rated more leniently than expected on SETF_Item4, TF_Item9, TF_Item16, TF_Item10, and TF_Item6, respectively; while Rater 1 from Sample UK rated more severely than expected on TF_Item7.

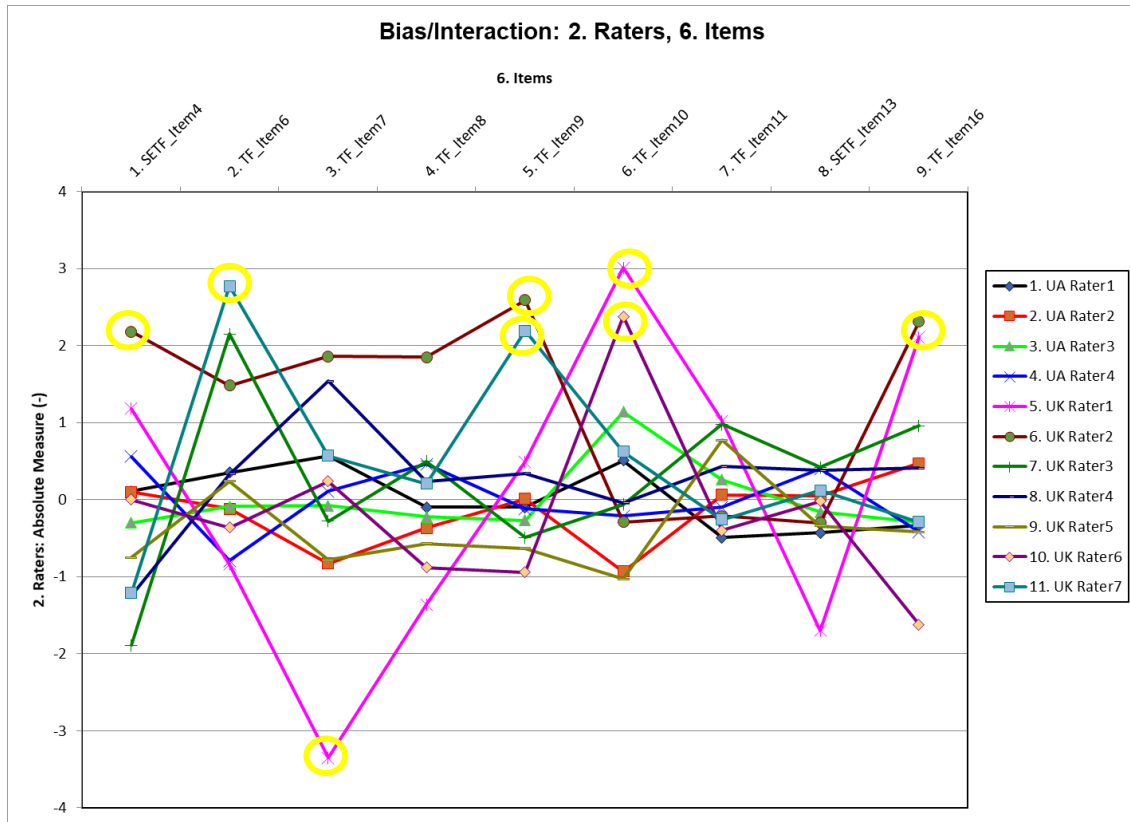


Figure 4.2

Plot Illustrating the Rater by Item Bias Interactions for the Teacher Facilitation Subscale

Analyses for Research Question Six

Research Question 6 (i.e., To what extent can the score levels of the MCOP² items be distinguished, without certain score levels being either underused or overused?) was evaluated by examining both the graphic indicators (i.e., Item Characteristic Curves, and Item Information Functions) and the statistical indicators (i.e., item category ordering for individual raters, and rater fit indices).

Descriptive statistics such as counts and percentages of scores in each category are first examined. Bond and Fox (2007) suggest that, as a rule of thumb, each category should be assigned to at least 10 ratings/observations to allow scale diagnostics (Linacre, 2003).

DATA				QUALITY CONTROL			RASCH-ANDRICH	EXPECTATION	MOST	RASCH-	Cat		
Category	Counts	Used	Cum. %	Avge Meas	Exp. Meas	OUTFIT MnSq	Thresholds	Measure at	PROBABLE	THURSTONE	PEAK		
Score	Total		%				measure	S.E.	Category	-0.5	from	Thresholds	Prob
0	226	226	16% 16%	-1.84	-1.84	1.0			(-3.08)	low	low	100%	
1	448	448	32% 48%	-.60	-.60	1.0	-1.91	.09	-1.00	-2.19	-1.91	-2.03	56%
2	456	456	32% 80%	.66	.66	1.0	-.01	.07	1.00	.00	-.01	.00	56%
3	293	284	20% 100%	2.29	2.28	1.0	1.91	.09	(3.09)	2.20	1.91	2.03	100%

(Mean)----- (Modal)--- (Median)-----

Figure 5.1

Summary Statistics of the Rating Scale Functioning for the Student Engagement Subscale

As shown in the above Figure 5.1, the Student Engagement 4-point scale functioning was examined based on a variety of diagnostic information, and the following findings were noted: (a) counts and percentages of scores in each of the four categories (highlighted in yellow) confirmed that each rating level had well above the minimum cut-off number (i.e., 10) of ratings/observations (ranging from 226 to 456 observations) to allow scale diagnostics (Linacre, 2003); (b) the (observed) average examinee ability measure associated with each category (highlighted in green) appeared to increase monotonically in size as the latent trait being measured increases, indicating that, on average, those with higher ability would be assigned to the higher scores (Bond & Fox, 2007; Linacre, 2003); (c) the outfit mean square index for each of the four categories (highlighted in blue) were all observed to be the ideal value 1.0, indicating that the observed and expected ratee ability measures were equal; and finally (d) step- or threshold-calibrations were reported in the highlighted red box representing difficulties estimated for choosing one response category over another, and they showed step increase as expected between 1.4 and 5 logits (Bond & Fox, 2007, p. 163).

DATA				QUALITY CONTROL			RASCH-ANDRICH	EXPECTATION	MOST	RASCH-	Cat	
Category	Counts	Used	Cum. %	Avge Meas	Exp. Meas	OUTFIT MnSq	Thresholds Measure	S.E.	Measure at category	PROBABLE from	THURSTONE Thresholds	PEAK Prob
0	224	224	16% 16%	-1.47	-1.3	.9			(-2.92)	low	low	100%
1	473	473	33% 49%	-.56	-.5	.9	-1.73	.08	-.97 -2.07	-1.73	-1.89	52%
2	488	488	34% 84%	.54	.4	1.1	-.11	.07	.94 -.04	-.11	-.07	56%
3	243	234	16% 100%	1.79	1.9	1.1	1.84	.09	(3.01) 2.12	1.84	1.95	100%

Figure 6.1

Summary Statistics of the Rating Scale Functioning for the Teacher Facilitation Subscale

The above Figure 6.1 displayed highly similar diagnostic information regarding the Teacher Facilitation 4-point scale functioning: (a) counts and percentages of scores in each of the four categories (highlighted in yellow) confirmed that each rating level had well above the minimum cut-off number (i.e., 10) of ratings/observations (ranging from 224 to 488 observations) to allow scale diagnostics (Linacre, 2003); (b) the (observed) average examinee ability measure associated with each category (highlighted in green) appeared to increase monotonically in size as the latent trait being measured increases, indicating that, on average, those with higher ability would be assigned to the higher scores (Bond & Fox, 2007; Linacre, 2003); (c) the outfit mean square index for each of the four categories (highlighted in blue) were all observed to be equal or very close to (e.g., 0.9) the ideal value 1.0, indicating that the observed and expected ratee ability measures were equal; and finally (d) step- or threshold-calibrations were reported in the highlighted red box representing difficulties estimated for choosing one response category over another, and they showed step increase as expected within the 1.4 to 5 logits range (Bond & Fox, 2007, p. 163).

Figures 5.2 and 6.2 above displayed a graphical representation of the Student Engagement and Teacher Facilitation subscale rating scales and the way they were used by the raters. From the graphs, it was clear that the raters were using all the categories of the two rating scales (0 through 3). The horizontal axis represented the ratee proficiency in logits and the vertical axis (from 0 to 1) represented the ratees' probability of being scored on a certain rating level. The scale category probability curves are labeled as 0, 1, 2, and 3, since both the SE and TD subscales used a 4-point rating scale.

It is important to discern whether there is a separate peak for each rating scale category probability curve, and whether the curves appear as an evenly spaced series of hills (Park, 2004). Each separate peak of a scale category curve indicates that, for ratees in a specific portion of the ratee proficiency distribution, that category is the most likely rating for their teaching performances. The absence of a separate peak would mean that the category is never the most probable rating for any clearly designated portion of the ratee proficiency distribution. As Davidson (1991) points out, such flat scale-steps are “operationally worthless” as they are never the most probable rater scale-step choice on any point along overall ratee ability (p. 159).

Examining Figures 5.2 and 6.2, the probability curves for the 4 ratings on both the SE and TF subscales were represented by a fairly evenly spaced series of hills. For each rating category there was a clearly designated portion of the ratee proficiency distribution for which that category would be the most probable rating given. Categories 0 and 3, as compared to the other categories, however, seems to be relatively underused.

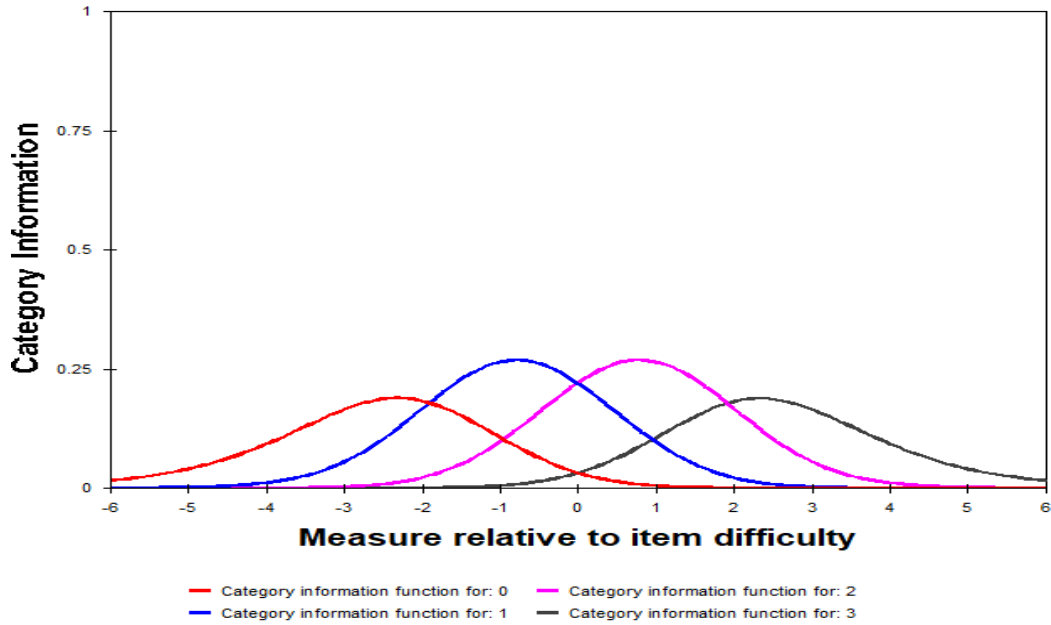


Figure 7.1

The Student Engagement Subscale Category Information Function (CIF)

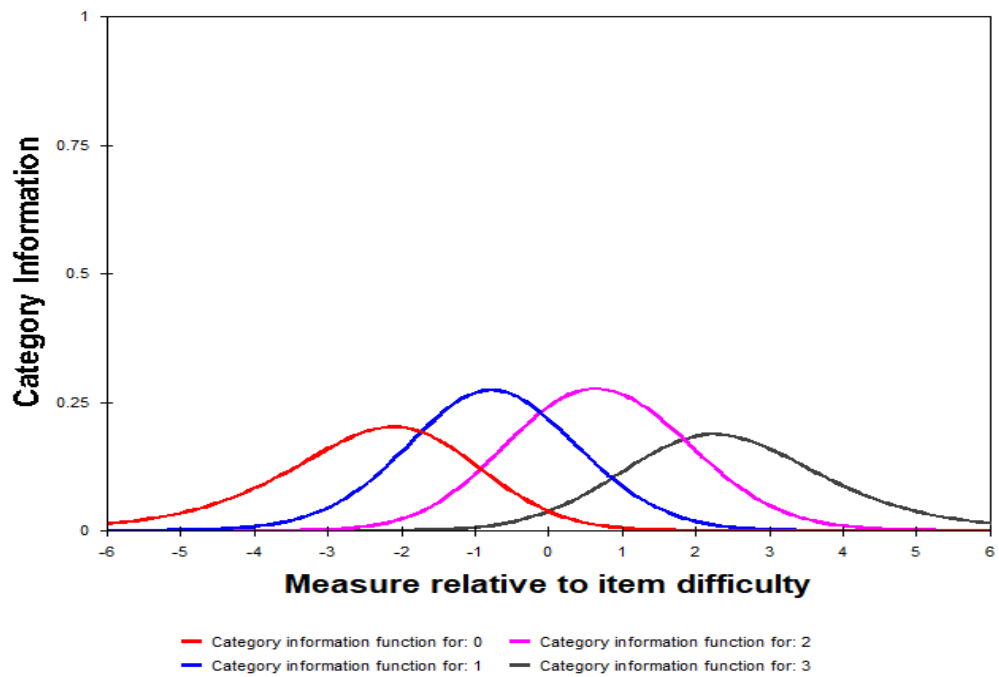


Figure 7.2

The Teacher Facilitation Subscale Category Information Function (CIF)

Figures 7.1 and 7.2 above showed a graphical representation of category information functions (CICs) for the Student Engagement and Teacher Facilitation subscales, respectively. For CICs, the wider the curves (capturing a wider range of values), the more popular the category would be, signifying overuse. It was found that for both the subscales, categories 1 and 2 gave the most information as they displayed the highest peaks (at the expense of the neighboring categories 0 and 3). Ideally, all these curves for all the categories should be of an equal height and spacing.

Figures 8.1 and 8.2 below presented a graphical representation of the item information functions (IIFs) for the Student Engagement and Teacher Facilitation subscales, respectively. For IIFs, the more dissimilar the shapes (sizes) of curves are, the more evidence there would be that the curves are conveying different amounts of information. The peaks occur where the categories intersect and where the item is doing best in discriminating between test taker proficiencies. For the MFRM analysis, any item would be most informative for ratees whose ability is equal to the difficulty level of the item. As shown in Figures 7.1 and 7.2, on both the SE and TF scales, the general pattern seemed to suggest that the items tended to give the most information (where the peaks were located) when the ratees' ability levels fell into the range between -1 and 1 logits.

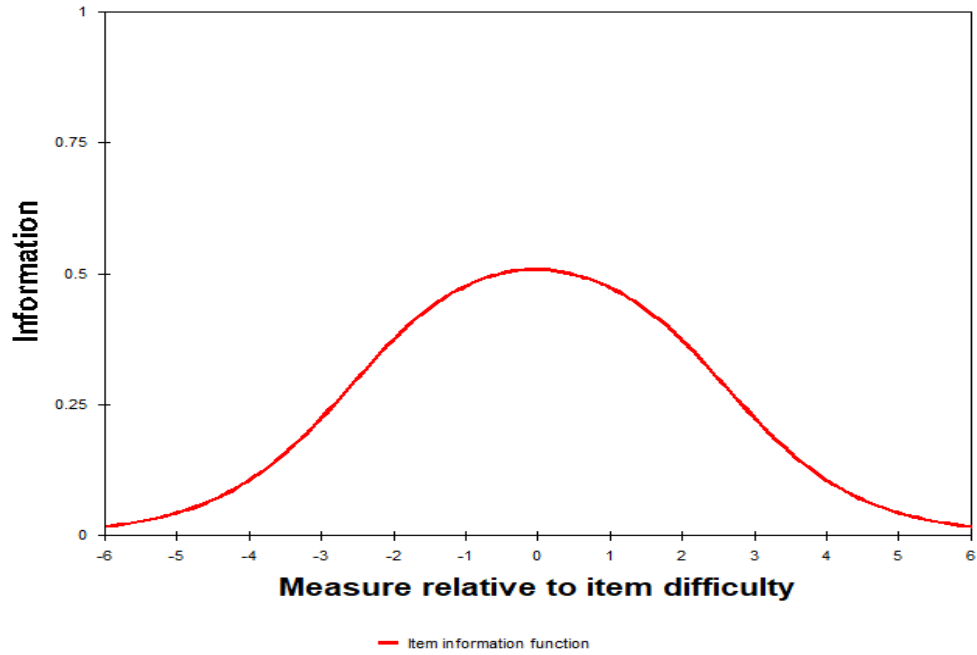


Figure 8.1

The Student Engagement Subscale Item Information Functions (IIFs)

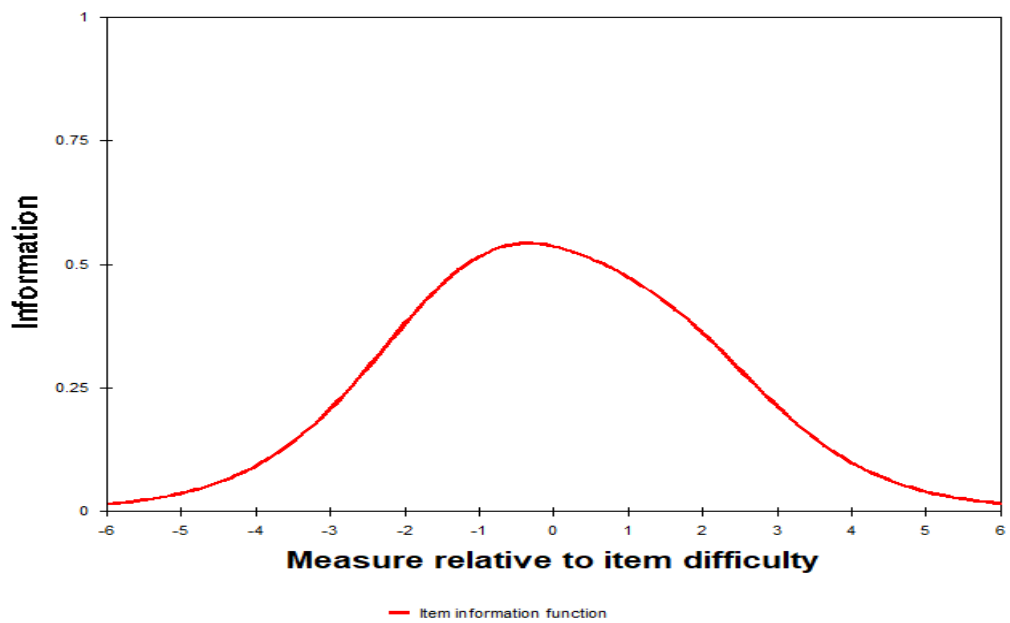


Figure 8.2

The Teacher Facilitation Subscale Item Information Function (IIF)

Analyses for Research Question Seven

Research Question 7 (i.e., To what extent are the rater behaviors associated with the professional background characteristics (i.e., in-service vs. pre-service teachers, study sites, and teaching grade levels) of the observed teachers?) was evaluated by examining possible interactions between raters and the facets indicating observed teachers' professional background in the MFRM analysis.

For each of the three external facets (i.e., in-service vs. pre-service teachers, study sites, and teaching grade levels), the original 4-facet MFRM model (i.e., ratees + MCOP² items + raters + classrooms) was modified to include an interaction term between the rater facet and the particular external facet to implement a MFRM-based bias analysis in Facets, respectively.

These three MFRM-based interaction analyses were performed following the same procedures and decision-making guidelines as detailed in the previous **Analyses for Research Question Five**.

Referring back to Tables 8.1 and 8.2 under **Analyses for Research Question Five**, a variety of the biased interaction information were presented including the total number of combinations of facet elements considered in each interaction analysis, the percent of absolute *t*-scores equal to or greater than 2, minimum and maximum *t*-values along with their degrees of freedom, the means and standard deviations of the bias sizes, fixed chi-square statistics, as well as the percentages of variances in the Student Engagement and/or Teacher Facilitation data explained by the bias terms.

However, the results showed that none of the interaction combinations in the three types of bias analyses yielded statistically significant *t*-scores. This means that raters did

not alternate between more severe ratings and more lenient ratings as a function of Sites, Service Types, or Classroom Grade Levels. Thus, all these interactions contributed to 0.00% of the total raw variances in the Student Engagement and/or Teacher Facilitation data and their effects on the overall MFRM analysis should be ignored.

Since the findings suggested that the MCOP² raters were not biased (i.e., either inappropriately increasing or decreasing their scores) towards certain types of candidates as related to their Sites, Service Types, or Classroom Grade Levels, the MFRM analysis calibration could successfully provide fair average scores for each ratee after adjusting for the rater effects and the above-mentioned three types of contextual factors in the observed raw ratings.

Table 8.1 below presented descriptive statistics of the observed and fair scores on Student Engagement the math teachers received across the various groups defined by their Sites, Service Types, or Classroom Grade Levels. An overall high correlation ($r = 0.987$) was observed between the observed and fair scores, suggesting a strong positive linear relationship between the observed and fair scores given to the math teachers in terms of Student Engagement. However, it was also evident that the MFRM-calibrated fair average scores substantially changed the raw score cross-group mean differences in Student Engagement after controlling for the contextual effects of Sites (i.e., from $-.13$ to $-.06$), Service Types (i.e., from $-.07$ to $.06$), or Classroom Grade Levels (e.g., raised the Tertiary Level raw score of $.88$ to 1.23 in its fair average score form).

Table 8.1

Ratees' Observed and Fair Scores on Student Engagement by Sites, Service Types, or Classroom Grade Levels

	<i>N</i>	Observed Scores <i>M(SD)</i>	Fair Scores <i>M(SD)</i>
Site			
Sample AL	129	1.56 (.72)	1.59 (.65)
Sample UK	30	1.69 (.46)	1.65 (.45)
Service Type			
In-Service	101	1.55 (.77)	1.62 (.68)
Pre-Service	58	1.62 (.50)	1.56 (.49)
Grade Levels			
Lower Elementary	27	1.88 (.61)***	1.74 (.66)**
Upper Elementary	13	1.88 (.57)***	1.74 (.62)
Middle School	12	1.81 (.78)***	1.83 (.80)
High School	25	1.63 (.65)***	1.64 (.66)
Secondary	16	1.87 (.11)***	1.75 (.48)**
Tertiary	36	.88 (.09)***	1.23 (.56)**
Unspecified	30	1.69 (.46)***	1.65 (.45)**

Note. ** $p < 0.01$ in cross-group mean comparison; *** $p < 0.001$ in cross-group mean comparison.

Similar trends were also observed in the MFRM analysis of the Teacher Facilitation data. An overall high correlation ($r = 0.985$) was observed between the observed and fair scores, suggesting a strong positive linear relationship between the observed and fair scores given to the math teachers regarding Teacher Facilitation. Again, it was noted that the MFRM-calibrated fair average scores substantially changed the raw score cross-group mean differences in Teacher Facilitation after controlling for the contextual effects of Sites (i.e., from .23 to .12), Service Types (i.e., from .27 to .20), or Classroom Grade Levels (e.g., decreased the Tertiary Level raw score of 1.52 to 1.46 in its fair average score form).

Table 8.2

Rates' Observed and Fair Scores on Teacher Facilitation by Sites, Service Types, or Classroom Grade Levels

	<i>N</i>	Observed Scores <i>M(SD)</i>	Fair Scores <i>M(SD)</i>
Site			
Sample AL	129	1.57 (.63)*	1.49 (.65)
Sample UK	30	1.34 (.48)*	1.37 (.36)
Service Type			
In-Service	101	1.62 (.64)**	1.54 (.66)*
Pre-Service	58	1.35 (.51)**	1.34 (.46)*
Grade Levels			
Lower Elementary	27	1.62 (.56)	1.49 (.58)
Upper Elementary	13	1.70 (.80)	1.59 (.84)
Middle School	12	1.40 (.70)	1.34 (.72)
High School	25	1.63 (.70)	1.58 (.73)
Secondary	16	1.52 (.59)	1.46 (.61)
Tertiary	36	1.52 (.56)	1.46 (.58)
Unspecified	30	1.34 (.48)	1.37 (.36)

Note. * $p < 0.05$ in cross-group mean comparison; ** $p < 0.01$ in cross-group mean comparison; *** $p < 0.001$ in cross-group mean comparison.

Summary

This study evaluated the rating quality obtained from a K-16 math classroom observation protocol (MCOP²) under a MFRM framework for the detection and control of rater effects and the effects of other potential construct-irrelevant factors during the rating processes. The data analyses (Research Question 1) testing the model-data fit of the MCOP² rating data to the MFRM analysis framework yielded results that (a) the CTT

factor analysis findings from the previous validation studies were further confirmed regarding the 2-factor structure for the 16-item MCOP² protocol; (b) raters appeared more internally consistent in using the 4-point rating scale appropriately for Teacher Facilitation than for Student Engagement; and (c) the nine items on both the Student Engagement and Teacher Facilitation subscales showed overall acceptable model-data fit, indicating that all subscale items were able to provide meaningful information on the latent trait being measured.

To investigate how well the ratings data for the two MCOP² subscales (i.e., Student Engagement and Teacher Facilitation) differentiate raters and ratees (Research Questions 2-3), the respective MFRM analyses suggested: (a) ratees measured by both the MCOP² subscales were separated into about 3 statistically distinct strata in terms of their performance on Student Engagement and Teacher Facilitation, respectively (Research Question 2); and (b) in contrast, raters using the two MCOP² subscales showed insubstantial cross-rater variability, although at least two raters were identified as significantly different from each other in the level of their severity/leniency (Research Question 3). These findings are further explored in **Chapter V**.

MFRM analysis was also used to study 2-way interactions between raters and ratees (Research Question 4), raters and items (Research Question 5), as well as raters and other contextual (construct-irrelevant) characteristics (Research Question 7). Regarding the rater-ratee interaction, the results showed that (a) raters were *internally inconsistent* in using the Student Engagement 4-point rating scale, or some raters might not have used the Student Engagement rating scale appropriately; and (b) for the Teacher Facilitation Subscale, raters were *internally consistent* and used its 4-point rating scale

appropriately.

While with respect to the rater-item interaction, it was found that (a) *not all* raters maintained a uniform level of severity across the nine items on the Student Engagement subscale, and the identified significant rater-item biases altogether contributed to 8.75% of the total raw variances in the Student Engagement data; while (b) when using for the Teacher Facilitation Subscale, a slightly *lower percentage of* significant rater by item biased interactions were identified, which still contributed to 7.06% of the total raw variances in the Teacher Facilitation data.

Based on the findings evaluating the interactions between raters and ratee background characteristics such as Study Sites (i.e., Sample AL vs. Sample UK), Service Types (i.e., In-Service Teachers vs. Pre-Service Teachers), and Classroom Grade Levels, raters are *not biased* towards certain types of ratees (i.e., math teachers under observation), either inappropriately increasing or decreasing their scores when using the two MCOP² subscales (i.e., Student Engagement & Teacher Facilitation).

The quality of the 4-point Likert scale functioning used in the MCOP² protocol for Student Engagement and Teacher Facilitation respectively was also systematically evaluated in the MFRM analysis for category ordering, fit indices, and/or possible underuse/overuse of some categories over others (Research Question 6). The findings highlighted that raters were using all the categories of the 4-point rating scale (0 through 3) in the expected/intended ranking order for Student Engagement and Teacher Facilitation respectively, with categories 0 and 3 appearing slightly relatively underused as compared to the other categories. Again, these findings are further discussed in **Chapter V**.

CHAPTER V

DISCUSSION AND CONCLUSION

The Study in Brief

In this chapter, the results obtained in this study were revisited and interpreted. First, a brief review of the purpose and rationale of the study was presented to provide the overarching research background for the following sections. The **Discussion** section was written following the same sequence of the seven research questions addressed in the **RESULTS** section. Next, the strengths and limitations of the current study were discussed in detail. Finally, implications for future research were explored.

This research sought to use the Many-Facet Rasch Model (MFRM) analysis for a systematic re-examination of the psychometric properties of a math classroom observation protocol (MCOP²), in which raw ratings of math teachers' classroom instructional performance were able to be calibrated after controlling for rater effects and other construct-irrelevant factors (i.e., math teachers' background characteristics). The findings of this study were expected to (a) address the methodological limitations displayed in the previous MCOP² validation studies where factor analysis was conducted, and interrater reliability statistics were calculated under the classical test theory (CTT) framework; and (b) transform and calibrate the MCOP² raw ratings of the math teachers on a common Rasch scale to produce observation scores that could be compared across rates, raters, classrooms, and study samples, especially in self- and/or peer-performance

assessments.

Discussion

This section discusses the meanings and connotations of the main findings for each of the seven research questions plus descriptive statistics of the study samples based on the data analysis results described in Chapter IV. Such interpretations were tied back to the research literature reviewed in Chapter II, with reference to the related theoretical and empirical studies as deemed necessary.

Sample Characteristics

The results concerning descriptive statistics first defined the population for this study as all pre- and in-service teachers who teach math in P-12 classrooms. While the two study samples drawn from this population were specified as 129 pre- and/or in-service math teachers from the neighboring school districts around the University of Alabama (i.e., Sample AL) and thirty pre-service math teachers from the University of Kentucky (i.e., Sample UK) whose teaching performances were observed and rated according to the MCOP² rubrics in the P-12 classrooms across the elementary, secondary, tertiary, and/or post-secondary levels. All the 159 math teachers in the combined sample were observed and rated by a single rater who had received formal or informal training on how to observe and give scores on the sixteen MCOP² items. The demographic background features of the math teachers ($n = 159$) in the combined study sample were defined by the four variables, namely, Study Site, MCOP² Raters, Classroom Grade Level, and Service Type (i.e., Pre-Service or In-Service).

Next, the descriptive statistics analysis highlighted various degrees of uneven distribution among the ratees (i.e., the math teachers under observation) grouped by the

four background variables. For example, only 19% of the participants in the combined study sample came from Sample UK, compared to 81% from Sample AL; Sample AL had fewer raters ($n = 4$) than Sample UK ($n = 7$); and more in-service math teachers ($n = 101$) were represented than pre-service teachers ($n = 58$) in the final combined sample, etc. Such varying background characteristics of the participants in the MCOP² samples reflected to a certain extent the true contextual complexities with which classroom observations (such as MCOP²) were typically implemented, including but not limited to school climate, teacher and teaching characteristics, the grade level, teaching topic, classroom dynamics, student academic achievements, and student demographic characteristics (Bell, Dobbelaer, & Klette, 2018; Grossman, Cohen, & Brown, 2014). Thus, it was vital to test the validity and reliability assumptions of the classroom observation protocols across different observation contexts, so that the observation ratings could be compared meaningfully in self- and/or peer-performance assessments longitudinally over the time and/or simultaneously with other classrooms (Mikeska, Holtzman, McCaffrey, Liu, & Shattuck, 2019). Without such sample-independent validation of the observation protocols, direct comparisons of the raw observation ratings could be very hard to interpret, and the resulting conclusions might be invalid, or even misleading (Gage & Needels, 1989; Medley, Coker, & Soar, 1984; Waxman, Tharp, & Hilberg, 2004).

To better understand the above-mentioned methodological concerns empirically, the raw MCOP² ratings obtained for Student Engagement and Teacher Facilitation respectively in the combined study sample in this research were directly compared for possible statistically significant cross-group differences by Study Site, MCOP² Raters,

Classroom Grade Level, and Service Type (i.e., Pre-Service or In-Service).

Such comparisons yielded findings in three aspects: (a) the pre-service math teachers in Sample UK were rated significantly lower on Teacher Facilitation than the math teachers in Sample AL; (b) on average, within Sample UK, the pre-service math teachers received much lower ratings on Teacher Facilitation than their ratings on Student Engagement; and (c) within Sample AL, the raw ratings of the pre- or in-service teachers were about equal on the two subscales of Teacher Facilitation and Student Engagement.

These differences in the mean comparisons of the MCOP² raw scores might lead to interesting interpretations from the psychometric perspective. For example, if the MCOP² protocol was deemed valid and reliable across Sample UK and Sample AL, the significant cross-group differences in the math teachers' Teacher Facilitation ratings might reflect the extent to which the MCOP² protocol could distinguish math teachers' true levels of teaching effectiveness across study samples. Because an overwhelming 78% of the rates in Sample AL were identified as in-service teachers, compared to Sample UK containing 100% pre-service math teachers, one would expect that compared to the pre-service teachers, the in-service teachers would be more experienced in teaching math and thus should perform notably better in facilitating student learning in their classrooms.

However, because the CTT approach of calculating interrater reliability is sample sensitive and cannot effectively control for various rater effects, the possibility could not be eliminated that such cross-sample differences might be mainly attributed to differences in rater severity/leniency levels and/or other rater bias across the study samples (Hilberg, Waxman, & Tharp, 2004; Ho & Kane, 2013).

Research Questions 1

The first research question investigated the overall model-data fit of the MCOP² ratings to the MFRM model, systematically evaluating the MFRM-based assumptions such as local independence, unidimensionality, overall model fit, rater fit, and item fit.

Specifically, testing the first two assumptions (i.e., local independence and unidimensionality) would provide further empirical evidence under the MFRM framework for the internal factorial structure and internal consistency of the 16-item MCOP² protocol and the two suggested subscales (i.e., Student Engagement and Teacher Facilitation) respectively. While the examination of the overall model fit, rater fit, and item fit of the MCOP² ratings were expected to offer unique MFRM-based diagnostic information on how the dynamic combination of raters, ratees, and the MCOP² items function in observing and assessing the P-12 math classrooms in terms of Student Engagement and Teacher Facilitation.

First, the findings regarding the local independence tests suggested that (a) the 16-item MCOP² scale indicated serious local dependency (LD) issues, with 5 pairs of item residual correlation well above the average residual correlation; (b) the 9-item Student Engagement subscale suggested slight LD problems with 3 pairs of item residual correlation notably above the average residual correlation; and (c) no LD-related concerns were identified for the 9-item Teacher Facilitation subscale where none of the pairs of item residual correlation is 0.2 above the average Q_3 .11.

The manifestation of LD-related issues in an instrument implied that apart from the variance explained by the latent construct of interest in the item responses, the remaining (i.e., residual) variances of some items were clustered on one or more possible

independent secondary factor(s) (Christensen, Markransky, & Horton, 2017; DeMars, 2010). To put it simply, strong evidence for LD concerns in various types of Rasch analysis warrants further investigation of multi-dimensionality problems. Thus, the LD-related findings for the 16-item MCOP² protocol strongly indicated the 16 items together measured more than one latent construct, generally consistent with the previous CTT factor analysis that resulted in a two-factor model for 16-item MCOP² scale (Gleason & Cofer, 2014). While mixed results were yielded for the two established 9-item subscales (i.e., Student Engagement and Teacher Facilitation) in terms of local independence, with the Student Engagement subscale showing slight LD concerns. Similar problems for the Student Engagement subscale were not identified or mentioned in the previous MCOP² validation studies (Gleason & Cofer, 2014; Gleason, Livers, & Zelkowski, 2017). However, since the extent of the item residual clustering on Student Engagement was not alarmingly notable (all less than 0.3 above the average item residual correlation), the manifested local dependency might be due to random noise in the MCOP² data, not necessarily indicating the existence of a secondary dimension apart from the latent construct of Student Engagement.

Second, with regard to the MFRM-based unidimensionality analyses (i.e., Principal Components Analysis on the standardized residuals), results highlighted that (a) the 16-item MCOP² protocol used as one single scale failed to uphold the unidimensionality assumption, as the residual variances of more than two items (eigenvalue = 3.75) clustered on a different dimension in addition to the variances explained by the MCOP² measure; (b) the unidimensionality assumption was better met for the 9-item Student Engagement subscale, with just about two item residuals loaded on

a dimension other than the latent trait measured (eigenvalue = 2.01); and (c) since less than two item residuals (eigenvalue = 1.66) were strongly correlated to form any contrast/factor apart from the variances explained by the measure, the 9-item Teacher Facilitation subscale successfully met the unidimensionality assumption.

To further evaluate whether the item residuals really clustered on a secondary dimension apart from the construct(s) of interest, the disattenuated correlations were also examined between the person measures on the suspect cluster of items and the person measures on the other items for the 16-item MCOP² protocol and the 9-item Student Engagement subscale, respectively. It was found that for the 16-item MCOP² protocol, the person measure disattenuated correlations between the 1st and 3rd cluster of items fell between the cut-off value range of 0.30 - 0.70 ($r = 0.48$), suggesting that the cluster of items on the suspect 1st contrast were measuring a secondary strand of the main Rasch dimension probably warranting separate investigation. However, for the 9-item Student Engagement subscale, all the disattenuated correlations were well above the upper bound of the cut-off value range (0.70), indicating that the suspect cluster of items was only measuring an insignificant secondary strand of the latent trait of interests and should not be considered as a different dimension (Linacre, 2013).

These unidimensionality findings again appeared to be in accordance with Gleason and his colleagues' (2014, 2017) MCOP² validation studies where factor analysis was conducted and yielded a two-factor model for the 16 MCOP² items. However, compared to the previous CTT factor analysis approach, the current MFRM-based dimensionality analysis had unique methodological advantages and thus offered more meaningful diagnostic information and more valid recommendations concerning the

MCOP² psychometric properties.

As Boone (2016) clearly outlined, CTT factor analysis was useful in describing the sample-dependent data with all its variety and intricacies to “evaluate the strength of the inferences drawn from instruments and to compute respondents’ (e.g., student, teacher) performances”; while Rasch dimensionality analysis (e.g., PCA of item residuals) was a sample-independent, prescriptive approach allowing researchers to identify subtle departures in the data from the ideal by fitting the Rasch model to the data in the process of constructing instruments (p. 1). Specifically, factor analysis might be able to identify different factors where clusters of certain items were loaded on but provided little help in the decision as to whether these factors could hang together to measure one overall latent construct. Additionally, factor analysis also tended to assign items in different difficulty strata to different factors, which often gave rise to misleading findings. In other words, inter-item correlations and item loadings in factor analysis could be affected by item difficulties, where the factor analysis of a pool of items containing both easy and difficult items could mistakenly produce two factors, even if all the items were supposed to measure one construct (Duncan, 1984). Therefore, the MFRM-based dimensionality analysis results in the current study provided strong psychometric evidence in support of the developers’ recommendation on the proper use of MCOP² protocol: the MCOP² was not designed to get a single score of a classroom; instead, it was used to measure two distinct (unidimensional) factors of Teacher Facilitation and Student Engagement through two subscales of 9 items each (Gleason, Livers, & Zelkowski, 2015).

The overall model-data fit was thus evaluated for the Student Engagement and

Teacher Facilitation subscales, respectively. In this study, data were deemed to have good overall model-fit in the MFRM analysis, if fewer than 5% of the standardized residuals appeared greater than or equal to $|2.0|$ and about 0.3% or less of standardized residuals are greater than or equal to $|3.0|$ (Linacre, 2004). The results indicated a satisfactory overall model fit for both subscales based on these criteria.

Further, the overall rater fit and item fit were examined for the two subscales, respectively. Mean Square outfit and Mean Square infit statistics (also referred to as MSU and MSW) were calculated and investigated (Bond & Fox, 2015; Eckes, 2005; Engelhard, 1994, 2002; Myford & Wolfe, 2003) to evaluate rater fit and/or item fit (Bond & Fox, 2015; Eckes, 2005; Engelhard, 1994, 2002; Myford & Wolfe, 2003). Linacre (2003) proposes that outfit and infit values between 0.5 and 1.5 can indicate acceptable fit. Results showed that one rater showed overfit (9%) and two raters fell into the misfit category (18%), when using the 9-item Student Engagement subscale. Since the percentage of raters showing acceptable fit (73%) for using the Student Engagement subscale fell well below 90%, it was concluded that the raters were internally inconsistent in using the 4-point rating scale, or the raters might not have used the Student Engagement rating scale appropriately. In comparison, when using the 9-item Teacher Facilitation subscale, more raters fell into the desirable category between 0.50 and 1.50 (91%) meaning that the number of overfitting and underfitting raters was minimal (only one out of the eleven raters showed underfit). Since the percentage of raters showing acceptable fit was above 90%, it could be concluded that the raters were internally consistent and used the 4-point Teacher Facilitation rating scale appropriately.

According to Eckes (2009), rater fit refers to the extent to which a given rater is

associated with unexpected ratings, summarized over ratees and items. Thus, overfitting raters would have muted ratings that suggested a central tendency or, alternatively, a halo effect (Engelhard, 2002; Myford & Wolfe, 2004). The underfitting/misfitting raters would suggest their ratings show off-target deviations/noise from the way the measure was intended to be used, and thus were unproductive (or even degrading in case of a serious extent of rater underfit/misfit) for construction of measurement.

The systematic diagnosis of rater fit in the MFRM analysis can be used to effectively address the methodological limitations noted in the standard approach of calculating interrater reliability (IRR) statistics to identify rater variability. First, a variety of existing IRR indices conceptualize interrater reliability differently (Bramley, 2007; Hayes & Krippendorff, 2007; LeBreton & Senter, 2008). For example, two broad classes of IRR indices (consensus indices and consistency indices) are widely used in the CTT validation studies of rater-mediated performance assessments (Stemler & Tsai, 2008). Specifically, a *consensus index* of IRR (also called *interrater agreement*) refers to the extent to which independent raters provide the identical rating of a particular person or object (absolute correspondence of ratings); whereas a *consistency index* of IRR refers to the extent to which independent raters provide the same relative ordering or ranking of the persons or objects being rated (relative correspondence of ratings) (Eckes, 2009). Whether one type of IRR index is chosen over the other, or both indices were reported in research related to rater-mediated performance assessments, high IRR statistics do not equal accurate ratings because (a) it is theoretically and empirically possible to observe low interrater consensus and high interrater consistency at the same time (and vice versa), and (b) even when the interrater consensus and consistency indices show the same trend

(i.e., both low or both high), the possibility of inaccurate ratings still cannot be eliminated as neither consensus and consistency indices could diagnose raters' use of the rating scale (e.g., overuse and/or underuse of certain response categories) or individual raters' severity/leniency levels (Eckes, 2009, 2011, 2012).

Second, in most social science measurement research, ordinal data are obtained from Likert-type scales, and too often researchers treat raw scores that are ordinal by nature (e.g., numeric values 0 to 3 are assigned to the response category levels in order of *strongly disagree*, *disagree*, *agree*, and *strongly agree*) as interval data in various statistical tests and analyses (including the calculation of interrater reliability coefficients). As Wright and Linacre (1989) rightfully point out, raw scores are NOT measures, since ordinal raw scores are limited due to “inequality of the units” counted as well as the resulting non-linearity in its distributions with strong ceiling and floor effects (Thorndike, 1904). Rasch modeling for ordinal observation raw scores can solve these problems by (a) confirming that raw scores can indeed be used for measuring the latent variable additively where a higher score indicates more of the latent variable than a lower score, and (b) transforming the non-linear ordinal raw scores into equal interval logits illustrating how much more of the latent variable one more score-point indicates at different locations along the latent variable (Wright & Stone, 1979).

In Gleason and his colleagues' MCOP² validation study (2017), a two-way mixed, absolute agreement intraclass correlation (ICC) were computed to assess the degree that raters provided consistent MCOP² ratings of the classrooms across subjects. The resulting single-measure ICCs for the student engagement subscale (0.669) and the teacher facilitation subscale (0.616) both fell within the “good” range (Cicchetti, 1994),

indicating a high degree of agreement among raters on both subscales (Gleason et al., 2017, p. 8). However, as mentioned above, these high ICC coefficients alone are not sufficient to support the assumption of accurate ratings. Moreover, such high ICC indices can hardly be replicated across different samples with different rating design (unlike the fully crossed model in the validation study where all raters rated all classrooms) and with different proportions and mechanisms of missing data for computing the ICC coefficients.

With respect to the overall item fit analysis, Results showed that based on the infit and outfit values, only one item showed underfit (11%) for both the 9-item Student Engagement subscale and the 9-item Teacher Facilitation subscale. Since the percentage of items showing acceptable fit (89%) for both the subscale were very close to 90%, it could be concluded that the nine items on either the Student Engagement or the Teacher Facilitation subscale were internally consistent and can be used to measure the latent traits of interests appropriately.

Although both the Student Engagement and Teacher Facilitation subscale items showed high internal consistency as the Cronbach's alpha values were both greater than 0.85 in Gleason and his colleagues (2017)'s validation studies, it does not sufficiently support the claims that these two subscales can be used for "effectively measuring differences at the group level, or at the individual level with at least three observations" due to a series of major methodological limitations related to the CTT internal consistency analysis methods (Gleason et al., 2017, p. 7).

Most reliability coefficients (e.g., Cronbach's alpha) are based on correlational statistical models of group-level information that treats individual items on a scale as

separate variables. Thus, the computation of Cronbach's alpha coefficient incorporates a single standard error estimated from that proportion of the variance not attributable to a common factor, assumably the latent construct of interest (Fisher Jr, Elbaum, & Coulter, 2010). In practice, Cronbach's alpha is more often used as a measure of a scale's internal consistency than as an estimate of reliability. However, significant methodological problems exist for Cronbach's alpha to be used in both circumstances. When used as a measure of internal consistency, Sijtsma (2009) posits that alpha is actually unrelated to the internal structure of a scale: since a 1-factor scale can have any alpha value as shown in numerous empirical studies, and vice versa, different scales of varying factorial composition may have the same alpha value, it would be safe to conclude that the alpha value is not indicative of unidimensionality and provides little psychometric information regarding a scale's internal structure. Similarly, when used as an estimate of reliability (i.e., repeatability of individual test performance described by the individual's propensity distribution), alpha statistics based on a single test administration cannot reflect the accuracy of individuals' test performance, because according to Molenaar (2004), "a single-administration sample of test scores does not contain information about the individuals' propensity distributions unless both types of distributions—between individuals as in single-administration data and within individuals as in propensity distributions—obey restrictive distributional properties" (as cited in Sijtsma, 2009, p. 117).

In contrast, measurement models of individual-level response processes employ individual-level error estimates (such as MFRM and other Rasch-based models), not correlational group-level residual variance estimates (as in the case of computing

Cronbach's alpha). The individual-level measurement errors are statistically equivalent to sampling confidence intervals. Measurement errors and confidence intervals both decline at the same rate with larger numbers of item responses per person, or larger numbers of person responses per item, which leads to improved measurement precision (Fisher Jr, Elbaum, & Coulter, 2010). Consequently, the MFRM analysis of item fit in the current study provided unique systematic diagnosis (including item difficulty, item functioning, item information, item measurement precision and reliability, item measurement invariance, various interaction effects between items and other facets, etc.) of the Student Engagement and Teacher Facilitation subscales at the individual item and scale/test level to evaluate the psychometric properties of each individual item in measuring the latent construct of interest.

To sum up, the findings for Research Question 1 laid the foundation for addressing the following research questions by analyzing the internal structure and consistency of the MCOP² items as a measurement of Student Engagement and Teacher Facilitation in K-16 math classrooms under the MFRM framework. Specifically, key results related to local independence, unidimensionality, overall model-data fit, overall rater fit and item fit were presented and interpreted, respectively. In addition, the methodological advantages of each aspect of the above-mentioned MFRM analysis in comparison with its corresponding CTT method were discussed in detail, highlighting the unique contributions of the current study to the MCOP²-related validation and empirical research.

Research Question 2

Research Question 2 (i.e., To what extent does the MCOP² observation protocol

separate observed teachers into distinct levels of proficiency?) was addressed by examining the examinee facet in the MFRM analysis. Variable maps (also referred to as Wright maps) were first examined closely visualizing the calibrations of raters, ratees, items, and the 4-point rating scales for the Student Engagement and Teacher Facilitation data, respectively. In addition, a series of MFRM analysis statistics related to the ratee facet were also investigated: standard deviations (SDs) of the estimated ratees' proficiencies, *RMSE* values for the ratee proficiency estimates indicating measurement errors, chi-square statistics testing the hypothesis that all ratees had the same proficiency, ratee separation index (*H*) estimates indicating the number of statistically distinct strata of measured ratee proficiency, separation ratio (*G*) estimates indicating that how many times greater the true standard deviation of ratee proficiency measures were than their standard error of measurement, and finally, separation reliability of the ratee proficiency estimates indicating how different the ratee proficiency measures were.

Results suggested (a) for both the Student Engagement and Teacher Facilitation subscales, the variability across ratees in their level of proficiency seemed substantial, with their proficiency estimates forming a wide range covering roughly seven logits; (b) the chi-square statistics were highly significant for both subscales, suggesting that overall, ratees significantly differ in terms of their proficiency level (after allowing for measurement error); (c) the ratee separation index (*H*) estimates showed that within this sample of ratees, there were about 3 statistically distinct strata of proficiency for both Student Engagement and Teacher Facilitation; and (d) The separation reliability statistics of the ratee proficiency estimates for both subscales were considerably high (i.e., 0.85 and 0.84 for Student Engagement and Teacher Facilitation, respectively), implying that

both subscales could differentiate very well among the ratees in terms of their levels of proficiency.

Taken together all the above-listed findings about the ratee facet, the MFRM analysis provided compelling (both visual and statistical information) psychometric evidence that both the 9-item Student Engagement and Teacher Facilitation subscales could effectively measure and differentiate the math teachers' performances along their respective latent constructs roughly into three proficiency level groups: those who fell below the expected performance standards, just meet the standards, and exceed the standards. Individual math teachers varied greatly within a wide, 7-logit range based on their MCOP² performance ratings.

Research Question 3

Research Question 3 (i.e., To what extent do raters differ in terms of the relative severity with which they rate observed teachers?) was evaluated by examining the rater facet in the MFRM analysis. Each of the global indices (i.e., the fixed chi square, rater separation index, and rater reliability of separation) were first examined that indicated the degree to which raters differed in their leniency/severity. After assessing rater leniency/severity differences globally, individual raters (anonymously coded as Rater 1, Rater 2, Rater 3, etc.) were then evaluated visually via the Wright maps.

It was found that (a) the fixed chi-square statistics testing the hypothesis that all raters have the same severity were highly significant for both the Student Engagement and Teacher Facilitation subscales, indicating that at least two raters were statistically significantly different in their leniency/severity measures; (b) however, the variability across raters in the level of the severity with which items were rated was not substantial

(i.e., only 1 statistically distinct strata of rater severity) for both the subscales; (c) Sample AL raters were about 0.1 logit more severe than the Sample UK raters (e.g., the most severe rater on Student Engagement was identified as Rater 4 from Sample AL); and (d) when using the Teacher Facilitation subscale, Sample AL raters were about 0.3 logit more lenient than the Sample UK raters (e.g., the most severe rater on Teacher Facilitation was identified as Rater 2 from Sample UK).

It was important to note that for the combined study sample used in the current study, each of the ratees was observed and scored by only one rater, and none of the 11 raters' ratings overlapped on any of the ratees. Thus, unlike the fully crossed study sample used in Gleason et al. (2017) validation study to compute the interrater reliability index, each set of the ratings on the 16 MCOP² items in the current study represented a unique case by rater and by ratee, and the interrater absolute agreement was 0% since there was zero rater overlap on the ratees. This kind of rating designs seem extremely ill-structured - also referred to as ill-structured measurement designs (ISMDs) in the literature - and are usually shunned in measurement research; however, it was not uncommon in empirical administrations of many classroom observation protocols in self- and/or peer-performance assessments due to limited resources and/or time (Conway, Jako, & Goodman, 1995; Hoyt, 2000; McCloy & Putka, 2004; Putka, Le, & McCloy, 2008).

The traditional CTT approach of interrater reliability analysis was very limited in its capacity of handling such ISMDs which would only magnify the already existing methodological issues of the traditional IRR methods. By contrast, the MFRM approach showed great potentials in calibrating raters' ratings in a common reference framework

even when the rating design was ill-structured/incomplete and missing data were inevitably present. The key lies in the use of the anchoring method within the MFRM framework to manage the placement of raters in non-fully crossed rating design. The MFRM anchoring method can be applied under any incomplete rating design for a combined dataset that is sufficiently connected (with sufficient links among every element, such as ratee, rater, and item, included in an observation case) (Eckes, 2009; Engelhard, 1997; Linacre & Wright, 2002; Wright & Stone, 1979). Group anchoring is a Rasch anchoring technique widely used in the literature, which is to set the average measure of the groups within one facet, such as raters, test-takers, or tasks, to zero logits (Linacre, 2012, 2017). The basic assumption underlying Rasch group anchoring is that the elements within that group-anchored facet (e.g., individual ratees or raters) are essentially exchangeable (Wind & Stager, 2019). Since the primary purpose of this study was to examine how well the MCOP² as a measurement could differentiate math teachers' performances while holding the rater effects stable, I chose to group-anchor the two groups (i.e., Sample AL and Sample UK raters) in the rater facet rather than those in the ratee facets.

The analysis highlighted that some raters significantly differ from each other in their levels of severity despite training. Specifically, Sample AL raters were 0.1 logit more severe than Sample UK raters on Student Engagement, and 0.3 logit more lenient than Sample UK raters on Teacher Facilitation. These differences might be attributed to several possible factors: (a) compared to Sample AL raters, the raters from Sample UK only received a limited amount of informal training on how to use MCOP², and did not go through a rigorous rating calibration process prior to the classroom observations; and

(b) Sample UK contained 100% pre-service math teachers observed and scored by their respective faculty supervisors during their student teaching, which might make the UK raters to rate more severely on Teacher Facilitation for the teacher training purposes. However, despite these rater severity level differences, low rater separation reliability (such as in this study) statistics were noted for both the Student Engagement and Teacher Facilitation subscales, which was generally desirable as this would indicate that raters were approaching the ideal of being interchangeable.

Research Question 4

Research Question 4 (i.e., To what extent do raters consistently rate the teaching performance of observed teachers?) was evaluated by investigating possible interactions between raters and observed teachers (i.e., rater fit indices) using the MFRM analysis. If the previous Research Question 3 addressed the interrater comparisons among the raters, this research question sought to investigate the intra-rater consistency and rating behaviors.

Rater fit statistics were first examined to understand the degree to which a rater (a) was internally self-consistent across examinees, items, and other factors, and (b) was able to implement the rating scale to make distinctions among ratees' performances (Bond & Fox, 2007; Weigle, 1998). Rater fit statistics close to the expected value of 1.0 suggested that a rater used the rating scale consistently and thus maintained his/her personal level of severity across ratees, items, and other factors (also referred to as intra-rater agreement).

Results showed that (a) some raters (i.e., Sample UK Rater 7, 1 and 3) were internally inconsistent in using the 4-point rating scale, or these raters might not have

used the Student Engagement rating scale appropriately; and (b) most raters (except for Sample UK Rater 1) were internally consistent and used the 4-point Teacher Facilitation rating scale appropriately.

Most of the CTT intra-rater reliability calculation methods are indirect and inaccurate estimates of the rating quality of the average rater in a sample or of individual raters. Thus, intra-rater reliability can be reported as a single index for a whole assessment project or for each of the raters in isolation. The single average intra-rater reliability index for a group of raters was often indexed by an average of the individual rater reliabilities, by an intra-class-correlation (ICC) or by an index of generalizability of the retesting facet that referred to the whole group of raters but not to individual raters. Whereas an individual rater's intra-rater reliability was usually reported as Cohen's kappa statistic, or as a correlation coefficient between two readings of the same set of essays (Shohamy et al., 1992). This type of intra-rater reliability is mathematically equivalent to the test-retest reliability of a single test-form. However, using these inter-rater reliability indices may bias the estimate of measurement error upwards or downwards (Cohen, 2017; Oberle, 2018; Rossi, 2017).

To address the issue of rater errors, some researchers has recommended another CTT analysis approach based on Generalizability Theory (Brennan, 2001), where the effects of multiple sources of rating errors are simultaneously investigated. With respect to the intra-rater reliability, researchers can limit the Generalizability Theory (GT) analysis only to one source of measurement error that is caused by the inconsistency of each rater by him/herself (Cohen & Allalouf, 2016; Cohen, 2017). However, compared to both the correlation-based intra-rater reliability analysis and the GT approach, the

MFRM approach possesses several methodological benefits in analyzing the raters' rating behavior and internal consistency. As Kim and Wilson (2009) point out, while G theory provides a general summary for all raters involved (including an estimation of the relative influence of each facet on a measure and the reliability of a decision based on the data), MFRM is able to (a) diagnose the individual rater's rating behavior systematically, (b) provide as fair a measure as it is possible to derive from the data, and (c) present summary information (e.g., reliability indices, the main effects of each facet, as well as any possible interaction effects among the facets). Therefore, the MFRM analysis related to Research Question 4 made unique contribution in terms of providing direct and systematic insights into the individual MCOP2 raters' rating behavior and if such rating behavior was held consistent/stable across ratees without rater drifts.

Research Question 5

Research Question 5 (i.e., To what extent do raters consistently rate the teaching performance of observed teachers across the MCOP² items?) was evaluated by investigating possible interactions between raters and the MCOP² items using the MFRM analysis.

To investigate whether each rater maintained a uniform level of severity across the nine items on the Student Engagement subscale, or whether particular raters gave ratings on some items more severely or leniently than expected, a two-way interaction analysis of Raters by Items was performed. MFRM-based bias analysis in Facets output a file (i.e., Table 13) with detailed statistical information to identify significantly biased rater-by-item interactions. McNamara (1996) and Kondo-Brown (2002) both recommend that only biased interactions with Z-values equal to or higher than the

absolute value of 2, plus MnSq infit values within the range of two standard deviations around the mean of infit should be considered.

Results showed that for Student Engagement, (a) about one fifth of the rater by item interaction combinations (21.28%) yielded statistically significant *t*-scores, suggesting that some raters tended to alternate between more severe ratings on one item and more lenient ratings on another item; (b) the majority of these significantly biased rater by item interactions appeared consistent in the identified patterns of bias across all rates; and (c) this relatively high percentage of significant rater by item biased interactions altogether contributed to 8.75% of the total raw variances in the Student Engagement data.

While for Teacher Facilitation, it was found that (a) a relatively lower percentage of the rater by item interaction combinations (13.27%) yielded statistically significant *t*-scores compared to those produced for the Student Engagement subscale; (b) the majority of these significantly biased interactions appeared consistent in the identified patterns of bias across all rates; and (c) the slightly lower percentage of significant rater by item biased interactions still contributed to 7.06% of the total raw variances in the Teacher Facilitation data.

The biased individual-level rater by item interactions were plotted and could be directly examined visually: (a) for Student Engagement, Raters 1, 2, and 7 from Sample UK and Rater 4 from Sample AL rated more leniently than expected on SE_Item3, SETF_Item4, SE_Item2, and SE_Item5, respectively; while Rater 2 from Sample AL rated more severely than expected on SE_Item3; while (b) for Teacher Facilitation, Raters 2, 6, and 7 from Sample UK rated more leniently than expected on SETF_Item4,

TF_Item9, TF_Item16, TF_Item10, and TF_Item6, respectively; while Rater 1 from Sample UK rated more severely than expected on TF_Item7.

Taken together all the above-listed findings, it seemed that some raters tended to interpret the scoring rubric on certain MCOP² items quite differently from each other, leading to the final variations in their rating severity/leniency levels on these items. These rater by item biases contributed to an unignorable proportion of the variances in the rating responses for both Student Engagement and Teacher Facilitation, and they could be the major factors causing the occurrences of rater misfits/overfits, especially on the Student Engagement subscale as previously illustrated under **Research Question 3**.

Wigglesworth (1993) believed that bias analysis could reveal systematic sub-patterns of rater behavior, and this notion was illustrated and supported in this MFRM study. Although the rater by item bias patterns discussed above only affected some, not all raters, they still suggested the presence of factors other than the latent constructs measured (i.e., Student Engagement and Teacher Facilitation) which would influence rater judgment when using the two MCOP² subscales. The identification of systematic sub-patterns to these factors could offer very important practical implications for further rater training and warrants future investigation. The findings related to Research Question 5 also demonstrated the powerful potential of MFRM in pinpointing the sources of rater bias, and in making rater-mediated performance assessments fairer, more equitable, and more informative (O'Neill & Lunz, 1997; Schaefer, 2008; Wigglesworth, 1993).

Research Question 6

Research Question 6 (i.e., To what extent can the score levels of the MCOP²

items be distinguished, without certain score levels being either underused or overused?) was evaluated by examining both the graphic indicators (i.e., Item Characteristic Curves, and Item Information Functions) and the statistical indicators (i.e., item category ordering for individual raters, and rater fit indices).

The 4-point scale functioning was examined for Student Engagement and Teacher Facilitation respectively based on a variety of diagnostic information, and the following findings were noted: (a) counts and percentages of scores in each of the four categories confirmed that each rating level had well above the minimum cut-off number (i.e., 10) of ratings/observations (ranging from 224 to 488 observations) to allow scale diagnostics; (b) the (observed) average examinee ability measure associated with each category appeared to increase monotonically in size as the latent trait being measured increases, indicating that, on average, those with higher ability would be assigned to the higher scores; (c) the outfit mean square index for each of the four categories were all observed to be equal or very close to the ideal value 1.0, indicating that the observed and expected ratee ability measures were equal; and finally (d) step- or threshold-calibrations were reported representing difficulties estimated for choosing one response category over another, and they showed step increase as expected between 1.4 and 5 logits.

Furthermore, the Probability Category Curves (PCCs) displayed a graphical representation of the Student Engagement and Teacher Facilitation subscale rating scales and the ways they were used by the raters. From the graphs, it was clear that the raters were using all the categories of the two rating scales (0 through 3). The PCCs for both subscales were represented by a fairly evenly spaced series of hills. For each rating category there was a clearly designated portion of the ratee proficiency distribution for

which that category would be the most probable rating given. Categories 0 and 3, as compared to the other categories, however, seems to be relatively underused.

Based on a graphical representation of category information functions (CICs) for the Student Engagement and Teacher Facilitation subscales, it was found that for both the subscales, categories 1 and 2 gave the most information as they displayed the highest peaks (at the expense of the neighboring categories 0 and 3), supporting the related findings from the previous examination of the PCCs. Ideally, all these curves for all the categories should be of an equal height and spacing.

Additionally, a graphical representation of the item information functions (IIFs) for the Student Engagement and Teacher Facilitation subscales showed that the items tended to give the most information (where the peaks were located) when the ratees' ability levels fell into the range between -1 and 1 logits.

All these findings concurred with each other, together supporting the notion from different perspectives that the 4-point rating scales for Student Engagement and Teacher Facilitation functioned reliably with this combined group of raters, who, with different experiences and training, were largely able to use the MCOP² rating scales to assess K-16 math classrooms to a satisfactory standard (despite the presence of some misfitting raters).

Research Question 7

Research Question 7 (i.e., To what extent are the rater behaviors associated with the professional background characteristics, namely, in-service vs. pre-service teachers, study cites, and teaching grade levels, of the observed teachers?) was evaluated by examining possible interactions between raters and the facets indicating observed

teachers' professional background in the MFRM analysis.

For each of the three external facets (i.e., in-service vs. pre-service teachers, study sites, and teaching grade levels), the original 4-facet MFRM model (i.e., ratees + MCOP² items + raters + classrooms) was modified to include an interaction term between the rater facet and the particular external facet to implement a MFRM-based bias analysis in Facets, respectively.

A variety of the biased interaction information were examined including the total number of combinations of facet elements considered in each interaction analysis, the percent of absolute *t*-scores equal to or greater than 2, minimum and maximum *t*-values along with their degrees of freedom, the means and standard deviations of the bias sizes, fixed chi-square statistics, as well as the percentages of variances in the Student Engagement and/or Teacher Facilitation data explained by the bias terms.

The findings suggested that the MCOP² raters were not biased (i.e., either inappropriately increasing or decreasing their scores) towards certain types of candidates as related to their Sites, Service Types, or Classroom Grade Levels. Thus, the MFRM analysis calibration could successfully provide fair average scores for each ratee after adjusting for the rater effects and the above-mentioned three types of contextual factors in the observed raw ratings.

The impact of the MFRM calibration was then evaluated by investigating descriptive statistics of the observed and fair scores on Student Engagement and Teacher Facilitation the math teachers received across the various groups defined by their Sites, Service Types, or Classroom Grade Levels. Results showed that for both Student Engagement and Teacher Facilitation, (a) an overall high correlation (above 0.98) was

observed between the observed and fair scores, suggesting a strong positive linear relationship between the observed and fair scores given to the math teachers; and (b) the MFRM-calibrated fair average scores for ratees substantially changed the raw score cross-group mean differences after controlling for the rater effects, as well as the three contextual effects, namely, Sites, Service Types, and Classroom Grade Levels.

As Eckes (2009) noted, a key issue with observed average scores was that they tended to confound ratee proficiency and rater severity, as well other construct-irrelevant factors. For example, when a particular ratee's observed average score was notably lower than other ratees' observed averages, this could be because he/she got a more severe rater than the other raters, or because this ratee belonged to a group rated by inexperienced raters. Fair averages produced by the MFRM calibration analysis could effectively resolve this problem. Fair averages thus disentangled rater severity from ratee proficiency so that the MFRM calibrated scores could be compared across samples, raters, and other grouping variables with more confidence than raw scores, a major methodological benefit of using the MFRM approach to analyze rater-mediated performance assessment data (Engelhard & Myford, 2003; Johnson, Penny, & Gordon, 2009; Linacre, 2012; Weigle, 1998; Wright & Mok, 2004).

Implications

In typical rater-mediated performance assessments such as classroom observation protocols, the process of obtaining assessment data mediated by human raters is complex and indirect, and very vulnerable to a variety of measurement errors, such as rater variability/effects and other construct-irrelevant variances (Eckes, 2015). Consequently, the observation data obtained using these classroom observation protocols confound ratee

ability/proficiency with rater severity and other construct-irrelevant effects. To separate the proportion of data variances directly attributed to the latent construct measured (i.e., true ratee ability/proficiency) from the residual variances caused by any construct-irrelevant factors, new modeling and statistical approaches are needed for validation and empirical research in rater-mediated performance assessments, and they should be different from the traditional methods and techniques under the classical test (CTT) theory framework that only work with raw scores/observed ratings.

To meet such needs, this study employed the many-facet Rasch measurement (MFRM) approach to (a) re-evaluate the psychometric properties of a classroom observation protocol, namely, the Mathematics Classroom Observation Protocol for Practices (MCOP²), as a valid and reliable measurement; (b) to demonstrate how MFRM could be used as a more robust methodological approach to validate the two MCOP² subscales in terms of internal structure and internal consistency, as well as to detect potential deficiencies of rater effects in MCOP² assessments; and (c) to reveal the powerful potentials of MFRM in calibrating observation ratings for rater effects to be used in multiple-site, large-scale self- and/or peer-performance assessments.

Therefore, this study had two important implications for future validation and empirical research in rater-mediated performance assessments. First, the systematic comparison was conducted between MFRM and the traditional CTT validation methods (e.g., factor analysis, interrater reliability), and the methodological pros and cons for each method were first discussed in theory, and then demonstrated empirically in the various elements of the MFRM analysis performed for the two MCOP² subscales. For example, factor analysis under the CTT framework may assist in identifying clusters of items

which threaten the invariance of the measurement system, but it is indirect and inexact (in some cases, these methodological limitations may even lead to misleading conclusions) compared with Rasch-based identification of anomalies in the data (Boone, 2016). In contrast, Rasch modeling (including MFRM) identifies departures in the data for persons, items, and other facets from the ideal of unidimensional structure of a measure. These deviations are reported with fit statistics that can guide the improvement of the instrument at the individual item level and point out possible flaws in the data. Under the Rasch framework, the most widely used technique for identifying multi-dimensionality in the data is Principal Component Analysis of Item Residuals (PCAR), which can be viewed as a form of Rasch-based factor analysis, but methodologically superior to its CTT counterpart (Linacre, 2009, 2012, 2014).

Second, the empirical implication of this MFRM study provided systematic diagnostic information to evaluate the psychometric properties of the two MCOP² subscales as a valid and reliable measure of Student Engagement and Teacher Facilitation, respectively. Psychometric evidence was examined for unidimensionality, overall model-data fit, rater fit, item fit, rating scale functioning, as well as rater bias across items and across groups of ratees as defined by Study Sites, Teacher Service Types, and Classroom Grade Levels.

In terms of the internal structure, the MFRM analysis results showed that (a) the overall MFRM findings appeared consistent with the conclusions reached in the previous CTT factor analysis (Gleason et al., 2017), namely, the nine items on the two MCOP² subscales were able to uphold the unidimensionality assumptions respectively for Student Engagement and Teacher Facilitation; (b) raters seemed more internally consistent in

using the 4-point rating scale appropriately for Teacher Facilitation than for Student Engagement; and (c) the nine items on both the Student Engagement and Teacher Facilitation subscales showed overall acceptable model-data fit, indicating that all subscale items were able to provide meaningful information on the latent trait being measured.

Furthermore, regarding how well the ratings data obtained on the two MCOP² subscales could reliably differentiate raters and ratees, the related MFRM findings highlighted that (a) ratees measured by both the MCOP² subscales were separated into about 3 statistically distinct strata in terms of their performance on Student Engagement and Teacher Facilitation; while (b) in contrast, raters using the two MCOP² subscales showed insubstantial cross-rater variability, although at least two raters were identified as significantly different from each other in the level of their severity/leniency. These findings, together with the findings on interaction analyses and rating scale functioning, strongly supported the general notion that both MCOP² subscales were highly reliable rater-mediated performance measures across raters, ratees, and study samples.

However, rater bias analyses yielded mixed results: although no significant rater bias was identified across the groups of ratees as defined by Study Sites, Service Types, and Classroom Grade Levels, rater bias on certain items from both subscales seemed to constitute a substantial proportion of the total variance in the MCOP² data. This implied a type of intra-rater inconsistency, where some raters tended to rate more severely than other raters on certain items while more leniently on some other items. MFRM analysis were able to provide detailed diagnostic information, both statistically and graphically, for targeted revision/rewording of the subscale item descriptors and/or enhanced training

for specific raters.

Limitations

In addition to the five major potential limitations listed in **Chapter I** (see pp. 16-18) including issues related to generalizability, replicability, data quality, small sample size, and secondary data sources, two additional limitations were noted upon the completion of the current study.

The first limitation concerned rater types. In the current study, although the eleven raters involved came from two study sites and had distinct experiences and forms of training in using the MCOP² subscales, they were all university faculty, independent researchers, or teacher educators. Thus, they could only conduct classroom observation and provide ratings on the MCOP² subscales from the supervisor's and/or the third-party perspective, which, according to the related research literature, is markedly different from self- and/or peer-performance assessment in terms of rater severity and rating behavior (Aryadoust, 2015; Farrokhi, Esfandiari & Dalili 2011; Farrokhi, Esfandiari & Schaefer 2012; Karakaya, 2015). For further investigation of the validity and reliability of the MCOP² subscales to be used in self- and/or peer-assessments, it would be critical to include a considerable number of self- and peer-raters in the future MFRM analyses of the MCOP² rating data.

The second limitation is related to the lack of the MFRM-based investigation of the MCOP² rating data over time and across parallel individual classroom contexts per teacher. The combined sample examined in the current study involved only cross-sectional data. However, as recommended by the MCOP² developers (Gleason et al., 2014), a single time MCOP²-based observation should only be used for formative

assessment; while for summative assessment, a minimum of 3 to 6 classroom observations should be conducted and recorded for each mathematics teacher. MFRM analysis of longitudinal observation data and parallel observation data would provide important insights into (a) the performance of the MCOP² subscales in measuring the ratees' changes over time, and (b) the ways the frequency of classroom observations could impact the validity and reliability of MCOP²-based summative assessment/evaluation for K-16 math teachers.

Future Research

Corresponding to the limitations discussed in the previous sections, three suggestions are proposed for future research on MFRM studies of MCOP² rating data.

First, future researchers might want to include self- and peer-raters in their MFRM modeling so that the rater severity and rating behaviors of these two types of raters can be calibrated and compared with other types of raters (e.g., supervisors, faculty mentors, school administrators, and internal and external evaluators and/or researchers) in a Rasch-based common reference framework.

Second, future researchers could expand the MCOP² rating datasets to include more study samples from diverse backgrounds (e.g., different types of schools, school districts, states, and math teachers trained in different teacher education/preparation programs) for further MFRM analysis to pinpoint any potential significantly biased interactions among raters, ratees, items, and contextual factors. For example, if significantly biased interactions are identified between some items and some contextual factors describing the ratees' personal or professional backgrounds, it would be considered differential item functioning (DIF) under the MFRM framework which

warrants further investigation.

Finally, future researchers could also attempt using/modeling the MCOP²-based ratings from the cognitive perspective, where a selected DCM model can be applied to provide detailed diagnostic information on individual teachers' weaknesses and strengths in terms of their mastery of cognitive attributes/skills necessary to perform effective classroom teaching. A dearth of research literature highlighting individual teachers' cognitive diagnosis exists in the field of rater-mediated teaching performance assessment; and in the limited number of studies exploring the underlying cognitive attribute/skills that facilitate the development of teacher proficiency in classroom instruction, researchers almost exclusively choose various qualitative methods (e.g., interviews and focus groups) to glean feedback from teacher educators, in-service and pre-service teachers, school administrators, or policy-makers (Leong, 2015; Wasserman & Ham, 2013; Wilson, 2005). While these exploratory studies may offer valuable opinions from different stakeholders within the teaching profession, their conclusions have never been validated psychometrically with real data, and thus lack the theoretical and empirical grounds for wide application in teacher assessment, learning, and training.

APPENDIX A: IRB APPROVAL LETTER



University of
Kentucky

Office of Research Integrity
IRB, RDRC

EXEMPTION CERTIFICATION

IRB Number: 63852

TO: Chunling Niu, EdD
College of Social Work
PI phone #: 2705350618

PI email: chunling.niu@uky.edu

FROM: Chairperson/Vice Chairperson
Nonmedical Institutional Review Board (IRB)

SUBJECT: Approval for Exemption Certification

DATE: 1/25/2021

On 1/25/2021, it was determined that your project entitled "*RATE TO COMPARE MATHEMATICS TEACHING: RECALIBRATE THE MATHEMATICS CLASSROOM OBSERVATION PROTOCOL FOR PRACTICES (MCOP2) FOR RATER EFFECTS CONTROL*" meets federal criteria to qualify as an exempt study.

Because the study has been certified as exempt, you will not be required to complete continuation or final review reports. However, it is your responsibility to notify the IRB prior to making any changes to the study. Please note that changes made to an exempt protocol may disqualify it from exempt status and may require an expedited or full review.

The Office of Research Integrity will hold your exemption application for six years. Before the end of the sixth year, you will be notified that your file will be closed and the application destroyed. If your project is still ongoing, you will need to contact the Office of Research Integrity upon receipt of that letter and follow the instructions for completing a new exemption application. It is, therefore, important that you keep your address current with the Office of Research Integrity.

For information describing investigator responsibilities after obtaining IRB approval, download and read the document "[PI Guidance to Responsibilities, Qualifications, Records and Documentation of Human Subjects Research](#)" available in the online Office of Research Integrity's [IRB Survival Handbook](#). Additional information regarding IRB review, federal regulations, and institutional policies may be found through [ORT's web site](#). If you have questions, need additional information, or would like a paper copy of the above mentioned document, contact the Office of Research Integrity at 859-257-9428.

see blue.

405 Kinkead Hall | Lexington, KY 40506-0057 | P: 859-257-9428 | F: 859-257-8995 | www.research.uky.edu/ori/

An Equal Opportunity University

APPENDIX B: MATHEMATICS CLASSROOM OBSERVATION PROTOCOL FOR PRACTICES

Mathematics Classroom Observation Protocol for Practices (MCOP²)

) Students engaged in exploration/investigation/problem solving.

SE	Description	Comments
3	Students regularly engaged in exploration, investigation, or problem solving. Over the course of the lesson, the majority of the students engaged in exploration/investigation/problem solving.	
2	Students sometimes engaged in exploration, investigation, or problem solving. Several students engaged in problem solving, but not the majority of the class.	
1	Students seldom engaged in exploration, investigation, or problem solving. This tended to be limited to one or a few students engaged in problem solving while other students watched but did not actively participate.	
0	Students did not engage in exploration, investigation, or problem solving. There were either no instances of investigation or problem solving, or the instances were carried out by the teacher without active participation by any students.	

) Students used a variety of means (models, drawings, graphs, concrete materials, manipulatives, etc.) to represent concepts.

SE	Description	Comments
3	The students manipulated or generated two or more representations to represent the same concept, and the connections across the various representations, relationships of the representations to the underlying concept, and applicability or the efficiency of the representations were explicitly discussed by the teacher or students, as appropriate.	
2	The students manipulated or generated two or more representations to represent the same concept, but the connections across the various representations, relationships of the representations to the underlying concept, and applicability or the efficiency of the representations were not explicitly discussed by the teacher or students.	
1	The students manipulated or generated one representation of a concept.	
0	There were either no representations included in the lesson, or representations were included but were exclusively manipulated and used by the teacher. If the students only watched the teacher manipulate the representation and did not interact with a representation themselves, it should be scored a 0.	

) Students were engaged in mathematical activities.

SE	Description	Comments
3	Most of the students spend two-thirds or more of the lesson engaged in mathematical activity at the appropriate level for the class. It does not matter if it is one prolonged activity or several shorter activities. (Note that listening and taking notes does not qualify as a mathematical activity unless the students are filling in the notes and interacting with the lesson mathematically.)	
2	Most of the students spend more than one-quarter but less than two-thirds of the lesson engaged in appropriate level mathematical activity. It does not matter if it is one prolonged activity or several shorter activities.	
1	Most of the students spend less than one-quarter of the lesson engaged in appropriate level mathematical activity. There is at least one instance of students' mathematical engagement.	
0	Most of the students are not engaged in appropriate level mathematical activity. This could be because they are never asked to engage in any activity and spend the lesson listening to the teacher and/or copying notes, or it could be because the activity they are engaged in is not mathematical – such as a coloring activity.	

) Students critically assessed mathematical strategies.

SE	TF	Description	Comments
3	3	More than half of the students critically assessed mathematical strategies. This could have happened in a variety of scenarios, including in the context of partner work, small group work, or a student making a comment during direct instruction or individually to the teacher.	
2	2	At least two but less than half of the students critically assessed mathematical strategies. This could have happened in a variety of scenarios, including in the context of partner work, small group work, or a student making a comment during direct instruction or individually to the teacher.	
1	1	An individual student critically assessed mathematical strategies. This could have happened in a variety of scenarios, including in the context of partner work, small group work, or a student making a comment during direct instruction or individually to the teacher. The critical assessment was limited to one student.	
0	0	Students did not critically assess mathematical strategies. This could happen for one of three reasons: 1) No strategies were used during the lesson; 2) Strategies were used but were not discussed critically. For example, the strategy may have been discussed in terms of how it was used on the specific problem, but its use was not discussed more generally; 3) Strategies were discussed critically by the teacher but this amounted to the teacher telling the students about the strategy(ies), and students did not actively participate.	

Mathematics Classroom Observation Protocol for Practices (MCOP²)

5) Students persevered in problem solving.

SE	Description	Comments
3	Students exhibited a strong amount of perseverance in problem solving. The majority of students looked for entry points and solution paths, monitored and evaluated progress, and changed course if necessary. When confronted with an obstacle (such as how to begin or what to do next), the majority of students continued to use resources (physical tools as well as mental reasoning) to continue to work on the problem.	
2	Students exhibited some perseverance in problem solving. Half of students looked for entry points and solution paths, monitored and evaluated progress, and changed course if necessary. When confronted with an obstacle (such as how to begin or what to do next), half of students continued to use resources (physical tools as well as mental reasoning) to continue to work on the problem.	
1	Students exhibited minimal perseverance in problem solving. At least one student but less than half of students looked for entry points and solution paths, monitored and evaluated progress, and changed course if necessary. When confronted with an obstacle (such as how to begin or what to do next), at least one student but less than half of students continued to use resources (physical tools as well as mental reasoning) to continue to work on the problem. There must be a road block to score above a 0.	
0	Students did not persevere in problem solving. This could be because there was no student problem solving in the lesson, or because when presented with a problem solving situation no students persevered. That is to say, all students either could not figure out how to get started on a problem, or when they confronted an obstacle in their strategy they stopped working.	

6) The lesson involved fundamental concepts of the subject to promote relational/conceptual understanding.

TF	Description	Comments
3	The lesson includes fundamental concepts or critical areas of the course, as described by the appropriate standards, and the teacher/lesson uses these concepts to build relational/conceptual understanding of the students with a focus on the "why" behind any procedures included.	
2	The lesson includes fundamental concepts or critical areas of the course, as described by the appropriate standards, but the teacher/lesson misses several opportunities to use these concepts to build relational/conceptual understanding of the students with a focus on the "why" behind any procedures included.	
1	The lesson mentions some fundamental concepts of mathematics, but does not use these concepts to develop the relational/conceptual understanding of the students. For example, in a lesson on the slope of the line, the teacher mentions that it is related to ratios, but does not help the students to understand how it is related and how that can help them to better understand the concept of slope.	
0	The lesson consists of several mathematical problems with no guidance to make connections with any of the fundamental mathematical concepts. This usually occurs with a teacher focusing on procedure of solving certain types of problems without the students understanding the "why" behind the procedures.	

7) The lesson promoted modeling with mathematics.

TF	Description	Comments
3	Modeling (using a mathematical model to describe a real-world situation) is an integral component of the lesson with students engaged in the modeling cycle (as described in the Common Core State Standards).	
2	Modeling is a major component, but the modeling has been turned into a procedure (i.e. a group of word problems that all follow the same form and the teacher has guided the students to find the key pieces of information and how to plug them into a procedure.); <u>or</u> modeling is not a major component, but the students engage in a modeling activity that fits within the corresponding standard of mathematical practice.	
1	The teacher describes some type of mathematical model to describe real-world situations, but the students do not engage in activities related to using mathematical models.	
0	The lesson does not include any modeling with mathematics.	

Mathematics Classroom Observation Protocol for Practices (MCOP²)

8) The lesson provided opportunities to examine mathematical structure. (symbolic notation, patterns, generalizations, conjectures, etc.)

TF	Description	Comments
3	The students have a sufficient amount of time and opportunity to look for and make use of mathematical structure or patterns.	
2	Students are given some time to examine mathematical structure, but are not allowed adequate time or are given too much scaffolding so that they cannot fully understand the generalization.	
1	Students are shown generalizations involving mathematical structure, but have little opportunity to discover these generalizations themselves or adequate time to understand the generalization.	
0	Students are given no opportunities to explore or understand the mathematical structure of a situation.	

9) The lesson included tasks that have multiple paths to a solution or multiple solutions.

TF	Description	Comments
3	A lesson which includes several tasks throughout; or a single task that takes up a large portion of the lesson; with multiple solutions and/or multiple paths to a solution and which increases the cognitive level of the task for different students.	
2	Multiple solutions and/or multiple paths to a solution are a significant part of the lesson, but are not the primary focus, or are not explicitly encouraged; or more than one task has multiple solutions and/or multiple paths to a solution that are explicitly encouraged.	
1	Multiple solutions and/or multiple paths minimally occur, and are not explicitly encouraged; or a single task has multiple solutions and/or multiple paths to a solution that are explicitly encouraged.	
0	A lesson which focuses on a single procedure to solve certain types of problems and/or strongly discourages students from trying different techniques.	

10) The lesson promoted precision of mathematical language.

TF	Description	Comments
3	The teacher "attends to precision" in regards to communication during the lesson. The students also "attend to precision" in communication, or the teacher guides students to modify or adapt non-precise communication to improve precision.	
2	The teachers "attends to precision" in all communication during the lesson, but the students are not always required to also do so.	
1	The teacher makes a few incorrect statements or is sloppy about mathematical language, but generally uses correct mathematical terms.	
0	The teacher makes repeated incorrect statements or incorrect names for mathematical objects instead of their accepted mathematical names.	

11) The teacher's talk encouraged student thinking.

TF	Description	Comments
3	The teacher's talk focused on high levels of mathematical thinking. The teacher may ask lower level questions within the lesson, but this is not the focus of the practice. There are three possibilities for high levels of thinking: analysis, synthesis, and evaluation. Analysis : examines/ interprets the pattern, order or relationship of the mathematics; parts of the form of thinking. Synthesis : requires original, creative thinking. Evaluation : makes a judgment of good or bad, right or wrong, according to the standards he/she values.	
2	The teacher's talk focused on mid-levels of mathematical thinking. Interpretation : discovers relationships among facts, generalizations, definitions, values and skills. Application : requires identification and selection and use of appropriate generalizations and skills	
1	Teacher talk consists of " lower order " knowledge based questions and responses focusing on recall of facts. Memory : recalls or memorizes information. Translation : changes information into a different symbolic form or situation.	
0	Any questions/ responses of the teacher related to mathematical ideas were rhetorical in that there was no expectation of a response from the students.	

12) There were a high proportion of students talking related to mathematics.

SE	Description	Comments
3	More than three quarters of the students were talking related to the mathematics of the lesson at some point during the lesson.	
2	More than half, but less than three quarters of the students were talking related to the mathematics of the lesson at some point during the lesson.	
1	Less than half of the students were talking related to the mathematics of the lesson.	
0	No students talked related to the mathematics of the lesson.	

Mathematics Classroom Observation Protocol for Practices (MCOP²)

13) There was a climate of respect for what others had to say.

SE	TF	Description	Comments
3	3	Many students are sharing, questioning, and commenting during the lesson, including their struggles. Students are also listening (active), clarifying, and recognizing the ideas of others.	
2	2	The environment is such that some students are sharing, questioning, and commenting during the lesson, including their struggles. Most students listen.	
1	1	Only a few share as called on by the teacher. The climate supports those who understand or who behave appropriately. Or Some students are sharing, questioning, or commenting during the lesson, but most students are actively listening to the communication.	
0	0	No students shared ideas.	

14) In general, the teacher provided wait-time.

SE	TF	Description	Comments
3	3	The teacher frequently provided an ample amount of "think time" for the depth and complexity of a task or question posed by either the teacher or a student.	
2	2	The teacher sometimes provided an ample amount of "think time" for the depth and complexity of a task or question posed by either the teacher or a student.	
1	1	The teacher rarely provided an ample amount of "think time" for the depth and complexity of a task or question posed by either the teacher or a student.	
0	0	The teacher never provided an ample amount of "think time" for the depth and complexity of a task or question posed by either the teacher or a student.	

15) Students were involved in the communication of their ideas to others (peer-to-peer).

SE	TF	Description	Comments
3	3	Considerable time (more than half) was spent with peer to peer dialog (pairs, groups, whole class) related to the communication of ideas, strategies and solution.	
2	2	Some class time (less than half, but more than just a few minutes) was devoted to peer to peer (pairs, groups, whole class) conversations related to the mathematics.	
1	1	The lesson was primarily teacher directed and little opportunities were available for peer to peer (pairs, groups, whole class) conversations. A few instances developed where this occurred during the lesson but only lasted less than 5 minutes.	
0	0	No peer to peer (pairs, groups, whole class) conversations occurred during the lesson.	

16) The teacher uses student questions/comments to enhance conceptual mathematical understanding.

SE	TF	Description	Comments
3	3	The teacher frequently uses student questions/ comments to coach students, to facilitate conceptual understanding, and boost the conversation. The teacher sequences the student responses that will be displayed in an intentional order, and/or connects different students' responses to key mathematical ideas.	
2	2	The teacher sometimes uses student questions/ comments to enhance conceptual understanding.	
1	1	The teacher rarely uses student questions/ comments to enhance conceptual mathematical understanding. The focus is more on procedural knowledge of the task verses conceptual knowledge of the content.	
0	0	The teacher never uses student questions/ comments to enhance conceptual mathematical understanding.	

Additional Notes: Preservice or Inservice. Live or Video. #Students, Grade Level, topic/subject, date, other demographics, school, etc.

APPENDIX C: MATHEMATICS CLASSROOM OBSERVATION PROTOCOL FOR
PRACTICES: DESCRIPTORS MANUAL

Mathematics Classroom Observation Protocol for Practices: Descriptors Manual

Authors

Jim Gleason
Department of Mathematics
The University of Alabama
jgleason@ua.edu

Stefanie Livers
Department of Curriculum and Instruction
The University of Alabama
sdlivers@bamaed.ua.edu

Jeremy Zelkowski
Department of Curriculum and Instruction
The University of Alabama
jzelkowski@ua.edu

Citation:

Gleason, J., Livers, S.D., & Zelkowski, J. (2015). *Mathematics classroom observation protocol for practices: Descriptors manual*. Retrieved from <http://jgleason.people.ua.edu/mcop2.html>

Acknowledgements:

We would like to thank Tracy Weston, John Dantzler, and John Abby Khalilian for their assistance in developing some of the items and descriptors. We would also like to acknowledge the many anonymous reviewers of the items that gave helpful feedback on item and descriptor wording.

Published May18, 2015

Disclaimer: This instrument may be used for evaluative educational purposes with consent from the authors. Upon publication in its final form, the authors grant permission for use for research purposes to anyone, with appropriate citations.

Mathematics Classroom Observation Protocol for Practices: Descriptors Manual

The Mathematics Classroom Observation Protocol for Practices (MCOP²) is a K-16 mathematics classroom instrument designed to measure the degree of alignment of the mathematics classroom with the various standards set out by the corresponding national organization that focus on conceptual understanding in the mathematics classroom including:

- Common Core State Standards in Mathematics: Standards for Mathematical Practice (National Governors Association Center for Best Practices, Council of Chief State School Officers, 2010),
- Mathematical Association of America (MAA): CUPM Curriculum Guide (Barker, et al., 2004)
- American Mathematical Association of Two-Year Colleges (AMATYC): “Crossroads” (AMATYC, 1995) and “Beyond Crossroads” (AMATYC, 2006), and
- National Council of Teachers of Mathematics (NCTM): Process Standards (NCTM, 2000).

Recommended Uses

The MCOP² form is designed to measure the activities occurring in a mathematics classroom during a single lesson. However, if one desires to measure the overall activities of a class, the form should be used to measure at least three different class settings. **An important item to remember is that while all of the items in the observation protocol are desired qualities of a mathematics classroom, not all of them are expected to be observed during a single lesson. It is expected that this instrument be used in a formative manner on single observations. Summatively, 3-6 observations are ideal in evaluating classroom instruction.**

The MCOP² form is not designed to be used during a single lesson or day to evaluate the teaching and learning atmosphere of the mathematics classroom.

When completing the MCOP² form, it is essential that the descriptors outlined in this manual are followed to maintain the validity and reliability of the instrument.

Published May18, 2015

Disclaimer: This instrument may be used for evaluative educational purposes with consent from the authors. Upon publication in its final form, the authors grant permission for use for research purposes to anyone, with appropriate citations.

How to Score

The MCOP² measures two distinct factors of Teacher Facilitation and Student Engagement through two subscales of 9 items each. (The MCOP² is not designed to get a single score of a classroom.)

The Teacher Facilitation subscale (Cronbach alpha of 0.850) measures the role of the teacher as the one who provides structure for the lesson and guides the problem solving process and classroom discourse. To calculate the score for the Teacher Facilitation subscale, one would add the scores for items 4, 6-11, 13, and 16.

The Student Engagement subscale (Cronbach alpha of 0.897) measures the role of the student in the classroom and their engagement in the learning process. To calculate the score for the Student Engagement subscale, one would add the scores for items 1-5 and 12-15.

Item	Student Engagement	Teacher Facilitation
1	X	
2	X	
3	X	
4	X	X
5	X	
6		X
7		X
8		X
9		X
10		X
11		X
12	X	
13	X	X
14	X	
15	X	
16		X

Published May18, 2015

Disclaimer: This instrument may be used for evaluative educational purposes with consent from the authors. Upon publication in its final form, the authors grant permission for use for research purposes to anyone, with appropriate citations.

1) Students engaged in exploration/investigation/problem solving.

The role of exploration, investigation, and problem solving is central in teaching mathematics as a process. In order for students to develop a flexible use of mathematics, they must be allowed to engage in exploration, investigation, and/or problem solving activities which go beyond following procedures presented by the teacher. Furthermore, problem solving can be developed as a valuable skill in itself (Barker, et al., 2004) and a way of thinking (NCTM, 1989), rather than just as the means to an end of finding the correct answer. Student exploration may also promote a stance of mathematics as a discipline that can be explored, reasoned about, connected to other subjects, and one that ‘makes sense’ (Barker, et al., 2004).

Mathematically proficient students start by explaining to themselves the meaning of a problem and looking for entry points to its solution. They analyze givens, constraints, relationships, and goals. They make conjectures about the form and meaning of the solution and plan a solution pathway rather than simply jumping into a solution attempt. They consider analogous problems, and try special cases and simpler forms of the original problem in order to gain insight into its solution (National Governors Association Center for Best Practices, Council of Chief State School Officers, 2010).

If students are following a procedure established by the teacher, then it does not count as exploration/investigation/problem solving. Instead, students should be determining their own solution pathway without necessarily knowing that the path will lead to the desired result.

Score	Description
3	Students regularly engaged in exploration, investigation, or problem solving. Over the course of the lesson, the majority of the students engaged in exploration/investigation/problem solving.
2	Students sometimes engaged in exploration, investigation, or problem solving. Several students engaged in problem solving, but not the majority of the class.
1	Students seldom engaged in exploration, investigation, or problem solving. This tended to be limited to one or a few students engaged in problem solving while other students watched but did not actively participate.
0	Students did not engage in exploration, investigation, or problem solving. There were either no instances of investigation or problem solving, or the instances were carried out by the teacher without active participation by any students.

2) Students used a variety of means (models, drawings, graphs, concrete materials, manipulatives, etc.) to represent concepts.

In mathematics instruction it is common for the teacher to use various representations (models, drawings, graphs, concrete materials, manipulatives, graphing calculators, compass & protractor, i.e. tools for the mathematics classroom) to focus students' thinking on and develop their conceptions of a mathematical concept. It is also important for students to interact with and develop representations of mathematical concepts and not merely observe the teacher presenting such representations. Thus, this item is concerned with whether the students use representations to represent mathematical concepts. The representations can be student generated (a drawing or a graph) or provided by the teacher (manipulatives or a table), but it is the students that must then use the representation. Just because there is a representation in a lesson, if it is only used by the teacher while students watch (such as a graph on a PowerPoint slide), it is not considered to be used by students unless the students manipulate and interact with the representation.

Students' notes can count as a type of representation if the students themselves offer some sort of input. For instance, if a student corrects a teacher's mistake in a problem he or she is copying down then the notes are actually being manipulated by a student and should therefore count as a type of representation.

Score	Description
3	The students manipulated or generated two or more representations to represent the same concept, and the connections across the various representations, relationships of the representations to the underlying concept, and applicability or the efficiency of the representations were explicitly discussed by the teacher or students, as appropriate.
2	The students manipulated or generated two or more representations to represent the same concept, but the connections across the various representations, relationships of the representations to the underlying concept, and applicability or the efficiency of the representations were not explicitly discussed by the teacher or students.
1	The students manipulated or generated one representation of a concept.
0	There were either no representations included in the lesson, or representations were included but were exclusively manipulated and used by the teacher. If the students only watched the teacher manipulate the representation and did not interact with a representation themselves, it should be scored a 0.

3) Students were engaged in mathematical activities.

This item is concerned with the extent of student engagement in activities that are mathematical. Students are considered to be engaged in a mathematical activity when they are investigating, problem solving, reasoning, modeling, calculating, or justifying (each of these could be written or verbal).

Note “most of the students” in an undergraduate mathematics classroom is accepted here to mean more than one-third of the students in the classroom were engaged in mathematical activity, while in a K-12 mathematics classroom it means more than one-half.

It is important to note that one should only focus on what actually happens—not what the teacher assigns watching for students who are off-task.

Score	Description
3	Most of the students spend two-thirds or more of the lesson engaged in mathematical activity at the appropriate level for the class. It does not matter if it is one prolonged activity or several shorter activities. (Note that listening and taking notes does not qualify as a mathematical activity unless the students are filling in the notes and interacting with the lesson mathematically.)
2	Most of the students spend more than one-quarter but less than two-thirds of the lesson engaged in appropriate level mathematical activity. It does not matter if it is one prolonged activity or several shorter activities.
1	Most of the students spend less than one-quarter of the lesson engaged in appropriate level mathematical activity. There is at least one instance of students' mathematical engagement.
0	Most of the students are not engaged in appropriate level mathematical activity. This could be because they are never asked to engage in any activity and spend the lesson listening to the teacher and/or copying notes, or it could be because the activity they are engaged in is not mathematical – such as a coloring activity.

4) Students critically assessed mathematical strategies.

In order for students to flexibly use mathematical strategies, they must develop ways to consider the appropriateness of a strategy for a given problem, task, or situation. This is because not all strategies will work on all problems, and furthermore the efficiency of the strategy for the given context needs to be considered. For students to make such distinctions it is important that they have opportunities to assess mathematical strategies so that they learn to reason not only about content but also about process. This item is concerned with *students* critically assessing strategies, which is more than listening to the teacher critically assessing strategies or asking peers how they solved a task. Examples of critical assessment include students offering a more efficient strategy, asking “why” a strategy was used, comparing/contrasting multiple strategies, discussing the generalizability of a strategy, or discussing the efficiency of different ways of solving a problem (e.g. the selection appropriate tools if needed).

To score high on this item it is the students who must be engaged in the critical assessment, not only the teacher.

Score	Description
3	More than half of the students critically assessed mathematical strategies. This could have happened in a variety of scenarios, including in the context of partner work, small group work, or a student making a comment during direct instruction or individually to the teacher.
2	At least two but less than half of the students critically assessed mathematical strategies. This could have happened in a variety of scenarios, including in the context of partner work, small group work, or a student making a comment during direct instruction or individually to the teacher.
1	An individual student critically assessed mathematical strategies. This could have happened in a variety of scenarios, including in the context of partner work, small group work, or a student making a comment during direct instruction or individually to the teacher. The critical assessment was limited to one student.
0	Students did not critically assess mathematical strategies. This could happen for one of three reasons: 1) No strategies were used during the lesson; 2) Strategies were used but were not discussed critically. For example, the strategy may have been discussed in terms of how it was used on the specific problem, but its use was not discussed more generally; 3) Strategies were discussed critically by the teacher but this amounted to the teacher telling the students about the strategy(ies), and students did not actively participate.

5) Students persevered in problem solving.

One of the *Standards for Mathematical Practice* (National Governors Association Center for Best Practices, Council of Chief State School Officers, 2010) is that students will persevere in problem solving. Student perseverance in problem solving is also addressed in the Mathematical Association of America's Committee on the Undergraduate Program in Mathematics Curriculum Guide (Barker, et al., 2004):

Every course should incorporate activities that will help all students...approach problem solving with a willingness to try multiple approaches, persist in the face of difficulties, assess the correctness of solutions, explore examples, pose questions, and devise and test conjectures.

Perseverance is more than just completion or compliance for an assignment. It should involve students overcoming a road block in the problem solving process.

Score	Description
3	Students exhibited a strong amount of perseverance in problem solving. The majority of students looked for entry points and solution paths, monitored and evaluated progress, and changed course if necessary. When confronted with an obstacle (such as how to begin or what to do next), the majority of students continued to use resources (physical tools as well as mental reasoning) to continue to work on the problem.
2	Students exhibited some perseverance in problem solving. Half of students looked for entry points and solution paths, monitored and evaluated progress, and changed course if necessary. When confronted with an obstacle (such as how to begin or what to do next), half of students continued to use resources (physical tools as well as mental reasoning) to continue to work on the problem.
1	Students exhibited minimal perseverance in problem solving. At least one student but less than half of students looked for entry points and solution paths, monitored and evaluated progress, and changed course if necessary. When confronted with an obstacle (such as how to begin or what to do next), at least one student but less than half of students continued to use resources (physical tools as well as mental reasoning) to continue to work on the problem. There must be a road block to score above a 0.
0	Students did not persevere in problem solving. This could be because there was no student problem solving in the lesson, or because when presented with a problem solving situation no students persevered. That is to say, all students either could not figure out how to get started on a problem, or when they confronted an obstacle in their strategy they stopped working.

6) The lesson involved fundamental concepts of the subject to promote relational/conceptual understanding.

Relational/conceptual understanding is “knowing both what to do and why” (Skemp, 1976). This is in contrast to a procedural understanding as being able to compute certain mathematical activities, but not understanding how the computation works or when one would need to use such a computation and what the answer would mean.

According to the NCTM (2006), certain topics are core to the mathematics learned at each grade level and can form the backbone of the K-8 curriculum. The NCTM extended this concept to the high school level with an emphasis on using these fundamental concepts to make sense of mathematics and deepen students’ relational and conceptual understanding (Martin, et al., 2009). Similar to the NCTM’s guidelines for middle school and high school mathematics lessons, at the undergraduate level the Mathematical Association of America has recommendations in the Committee on the Undergraduate Program in Mathematics Curriculum Guide (Barker, et al., 2004) for departments, programs, and all courses to promote relational/conceptual understanding for both mathematics majors and non-mathematics majors.

Score	Description
3	The lesson includes fundamental concepts or critical areas of the course, as described by the appropriate standards, and the teacher/lesson uses these concepts to build relational/conceptual understanding of the students with a focus on the "why" behind any procedures included.
2	The lesson includes fundamental concepts or critical areas of the course, as described by the appropriate standards, but the teacher/lesson misses several opportunities to use these concepts to build relational/conceptual understanding of the students with a focus on the "why" behind any procedures included.
1	The lesson mentions some fundamental concepts of mathematics, but does not use these concepts to develop the relational/conceptual understanding of the students. For example, in a lesson on the slope of the line, the teacher mentions that it is related to ratios, but does not help the students to understand how it is related and how that can help them to better understand the concept of slope.
0	The lesson consists of several mathematical problems with no guidance to make connections with any of the fundamental mathematical concepts. This usually occurs with a teacher focusing on procedure of solving certain types of problems without the students understanding the “why” behind the procedures.

7) The lesson promoted modeling with mathematics.

Following the “Standards for Mathematical Practice” from the Common Core State Standards (2010) and the recommendations from the MAA’s CUPM Curriculum Guide (Barker, et al., 2004), this item describes lessons that help students to “apply the mathematics they know to solve problems arising in everyday life, society, and the workplace. In early grades, this might be as simple as writing an addition equation to describe a situation. In middle grades, a student might apply proportional reasoning to plan a school event or analyze a problem in the community. By high school, a student might use geometry to solve a design problem or use a function to describe how one quantity of interest depends on another” (National Governors Association Center for Best Practices, Council of Chief State School Officers, 2010).

In an undergraduate classroom, a lesson that promotes modeling might use “radiocarbon dating to illustrate how an initial value problem (IVP) can model a real world situation, and the solution of the IVP then yields obviously useful and interesting results” or “a simple system of differential equations to predict the cyclical population swings in a predator-prey relationship” or even “how modular arithmetic is used in cryptography and the transmission of encoded information” (Barker, et al., 2004).

Score	Description
3	Modeling (using a mathematical model to describe a real-world situation) is an integral component of the lesson with students engaged in the modeling cycle (as described in the Common Core State Standards).
2	Modeling is a major component, but the modeling has been turned into a procedure (i.e. a group of word problems that all follow the same form and the teacher has guided the students to find the key pieces of information and how to plug them into a procedure.); <u>or</u> modeling is not a major component, but the students engage in a modeling activity that fits within the corresponding standard of mathematical practice.
1	The teacher describes some type of mathematical model to describe real-world situations, but the students do not engage in activities related to using mathematical models.
0	The lesson does not include any modeling with mathematics.

8) The lesson provided opportunities to examine mathematical structure. (Symbolic notation, patterns, generalizations, conjectures, etc.)

Following some of the “Standards for Mathematical Practice” (National Governors Association Center for Best Practices, Council of Chief State School Officers, 2010) and the recommendations in the MAA’s CUPM Curriculum Guide (Barker, et al., 2004), lessons should include opportunities for students to contextualize and/or decontextualize in the process of solving quantitative problems, explore and make use of mathematical structure, or to use repeated reasoning to generalize certain categories of problems and their solutions.

Score	Description
3	The students have a sufficient amount of time and opportunity to look for and make use of mathematical structure or patterns.
2	Students are given some time to examine mathematical structure, but are not allowed adequate time or are given too much scaffolding so that they cannot fully understand the generalization.
1	Students are shown generalizations involving mathematical structure, but have little opportunity to discover these generalizations themselves or adequate time to understand the generalization.
0	Students are given no opportunities to explore or understand the mathematical structure of a situation.

9) The lesson included tasks that have multiple paths to a solution or multiple solutions.

As part of having students “make sense of problems and persevere in solving them” (National Governors Association Center for Best Practices, Council of Chief State School Officers, 2010), students must be encouraged to look for multiple methods of solving a problem and to deal with problems that have multiple solutions based upon various assumptions. Additionally, selected tasks with multiple paths to a solution or multiple solutions can increase the cognitive demand of the task for all students through the interaction of the teacher to ask questions of each student at their ability level (Stein & Smith, 1998). This flexibility, “switching (smoothly) between different strategies,” and adaptivity, “selecting the most appropriate strategy” (Verschaffel, Luwel, Torbeyns, & Van Dooren, 2009) enables students to solve problems for which a solution path is not obvious.

Score	Description
3	A lesson which includes several tasks throughout; or a single task that takes up a large portion of the lesson; with multiple solutions and/or multiple paths to a solution and which increases the cognitive level of the task for different students.
2	Multiple solutions and/or multiple paths to a solution are a significant part of the lesson, but are not the primary focus, or are not explicitly encouraged; <u>or</u> more than one task has multiple solutions and/or multiple paths to a solution that are explicitly encouraged.
1	Multiple solutions and/or multiple paths minimally occur, and are not explicitly encouraged; <u>or</u> a single task has multiple solutions and/or multiple paths to a solution that are explicitly encouraged.
0	A lesson which focuses on a single procedure to solve certain types of problems and/or strongly discourages students from trying different techniques.

10) The lesson promoted precision of mathematical language.

This item follows the Standard of Mathematical Practice to “attend to precision”. As such, “Mathematically proficient students try to communicate precisely to others. They try to use clear definitions in discussion with others and in their own reasoning. They state the meaning of the symbols they choose, including using the equal sign consistently and appropriately. They are careful about specifying units of measure, and labeling axes to clarify the correspondence with quantities in a problem” (National Governors Association Center for Best Practices, Council of Chief State School Officers, 2010).

This item also follows the MAA’s CUPM Curriculum Guide recommendation to “develop mathematical thinking and communication skills” which states: “Students should read mathematics with understanding and communicate mathematical ideas with clarity and coherence through writing and speaking” (Barker, et al., 2004).

Whether the communication is verbal or written and originating in the teacher or a student, using precise mathematical language is important. While the teacher cannot control the language used by students, there should be evidence of expectations of the teacher upon the students related to communicating with precise mathematical language. For example, if the lesson is primarily students solving problems, a culture of precision of language should come through in how the students are communicating with one another, both verbal and written.

Score	Description
3	The teacher “attends to precision” in regards to communication during the lesson. The students also “attend to precision” in communication, or the teacher guides students to modify or adapt non-precise communication to improve precision.
2	The teachers “attends to precision” in all communication during the lesson, but the students are not always required to also do so.
1	The teacher makes a few incorrect statements or is sloppy about mathematical language, but generally uses correct mathematical terms.
0	The teacher makes repeated incorrect statements or incorrect names for mathematical objects instead of their accepted mathematical names.

11) The teacher's talk encouraged student thinking.

This item assesses how well the teacher's talk promotes a number of the mathematical practices. Specifically, the practices requiring students to be able to think, reason, argue, and critique during the study of mathematical concepts (National Governors Association Center for Best Practices, Council of Chief State School Officers, 2010). Teachers can greatly impact the level of student thinking and discussion simply by what questions are asked of students. In line with Stein, et al. (2009), the cognitive task level should be maintained at a high level, i.e. procedures with connections and doing mathematics, while questions which are over-scaffolded, rhetorical, or cursory to the level of the students, would score a 1 or a 0.

Specifically about the teacher's talk, this item is referring to the content of the question or statements put forth in the classroom for students to reason and/or discuss. A well planned lesson may contain rich tasks for students to explore or problems to solve, but if the teacher's talk drops or removes student reasoning and problem solving, it has removed or reduced student thinking.

Score	Description
3	The teacher's talk focused on high levels of mathematical thinking. The teacher may ask lower level questions within the lesson, but this is not the focus of the practice. There are three possibilities for high levels of thinking: analysis, synthesis, and evaluation. Analysis: examines/ interprets the pattern, order or relationship of the mathematics; parts of the form of thinking. Synthesis: requires original, creative thinking. Evaluation: makes a judgment of good or bad, right or wrong, according to the standards he/she values.
2	The teacher's talk focused on mid-levels of mathematical thinking. Interpretation: discovers relationships among facts, generalizations, definitions, values and skills. Application: requires identification and selection and use of appropriate generalizations and skills
1	Teacher talk consists of " lower order " knowledge based questions and responses focusing on recall of facts. Memory: recalls or memorizes information. Translation: changes information into a different symbolic form or situation.
0	Any questions/ responses of the teacher related to mathematical ideas were rhetorical in that there was no expectation of a response from the students.

12) There were a high proportion of students talking related to mathematics.

The focus of this descriptor is on the proportion of students talking (frequency). The Standards for Mathematical Practice (National Governors Association Center for Best Practices, Council of Chief State School Officers, 2010) encourages students to be active in making conjectures, exploring the truth of those conjectures, and responding to the conjectures and reasoning of others. In a classroom dominated by only a few students, classroom discourse may appear to be high, but all students must be engaged.

Score	Description
3	More than three quarters of the students were talking related to the mathematics of the lesson at some point during the lesson.
2	More than half, but less than three quarters of the students were talking related to the mathematics of the lesson at some point during the lesson.
1	Less than half of the students were talking related to the mathematics of the lesson.
0	No students talked related to the mathematics of the lesson.

13) There was a climate of respect for what others had to say.

This item adheres to the expectation provided in the third Standard for Mathematical Practice, “Construct viable arguments and critique the reasoning of others.” Given that practice, students are expected to communicate with each other as part of an effective classroom community. Effective communication means that students will listen, question, and critique; this is part of the discourse expected in a mathematics classroom (Sherin, Mendez, & Louis, 2004). This item also encompasses the literature on equity and mathematics in that all students have valuable ideas, strategies, and thinking to share within the mathematics classroom (Boaler, 2006). Equitable spaces include the interactions of students within a mathematical community that increase participation and engagement of all students and work to remove potential barriers (Diversity in Mathematics Education Center for Learning and Teaching, 2007; Gutierrez, 2007; Hiebert & Grouws, 2007; NCTM, 2000; Sherin, Mendez, & Louis, 2004; Yackel & Cobb, 1996). This means creating a climate of respect.

Score	Description
3	Many students are sharing, questioning, and commenting during the lesson, including their struggles. Students are also listening (active), clarifying, and recognizing the ideas of others.
2	The environment is such that some students are sharing, questioning, and commenting during the lesson, including their struggles. Most students listen.
1	Only a few share as called on by the teacher. The climate supports those who understand or who behave appropriately. Or Some students are sharing, questioning, or commenting during the lesson, but most students are actively listening to the communication.
0	No students shared ideas.

14) In general, the teacher provided wait-time.

The appropriate wait time must align with the question/task. In the elementary grades, a teacher may ask students to explain a situation that represents the expression $24 \cdot (1/2)^3$. In middle school, the teacher may ask students to describe why the slope is positive. High school teachers may ask students to explain how linear and exponential functions are similar and different. In each instance, these questions/tasks are not simple yes/no answer and require wait time to provide an answer with meaning and understanding.

Simple Yes/No questions could be asked, but must be accompanied by an explanation. Simple skills or procedural problems should require explanations with the computation and/or procedures. If the class is dominated by rhetorical questions, a score of 0 or 1 is warranted. Even if rhetorical questions are asked, it is possible to score a 2 or 3 if there are questions asked sometimes or frequently that require students to reason, make sense, and articulate thoughtful responses.

Score	Description
3	The teacher frequently provided an ample amount of “think time” for the depth and complexity of a task or question posed by either the teacher or a student.
2	The teacher sometimes provided an ample amount of “think time” for the depth and complexity of a task or question posed by either the teacher or a student.
1	The teacher rarely provided an ample amount of “think time” for the depth and complexity of a task or question posed by either the teacher or a student.
0	The teacher never provided an ample amount of “think time” for the depth and complexity of a task or question posed by either the teacher or a student.

15) Students were involved in the communication of their ideas to others (peer-to-peer).

Both the National Council of Teachers of Mathematics and The Eight Standards for Mathematical Practices, expect teachers to create a mathematical community that includes dialogue around the mathematics content and learning. Students are expected to talk and participate in the discourse of the classroom (Manouchehri & St John, 2006). This item highlights the need for all students to be active participants in the classroom dialogue. Without teacher support and expectations, the classroom discourse can be monopolized or biased against certain populations (Mercer & Wegerif, 1999; Mercer, Wegerif, & Dawes, 1999; Rojas-Drummond & Mercer, 2003; Rojas-Drummond & Zapata, 2004).

This descriptor focuses on the amount of time students spend in communication with their peers at any level, including pairs, groups, informal settings, or whole class settings.

Score	Description
3	Considerable time (more than half) was spent with peer to peer dialog (pairs, groups, whole class) related to the communication of ideas, strategies and solution.
2	Some class time (less than half, but more than just a few minutes) was devoted to peer to peer (pairs, groups, whole class) conversations related to the mathematics.
1	The lesson was primarily teacher directed and little opportunities were available for peer to peer (pairs, groups, whole class) conversations. A few instances developed where this occurred during the lesson but only lasted less than 5 minutes.
0	No peer to peer (pairs, groups, whole class) conversations occurred during the lesson.

16) The teacher uses student questions/comments to enhance conceptual mathematical understanding.

Driscoll (1999; 2007) and Reys, et al. (2009) discuss how teacher questioning can build on student thinking to foster deeper mathematical thinking. In the elementary grades, students can make “over generalized” statements that have a correct nature about them. This is a teachable moment to use. A teacher can ask a question that has the student(s) reexamine their thoughts that would help simplify the over generalizing statement into precise understanding. Reys, et al. (2009) present a simple example, “Student: So every even number is composite. Teacher: Every even number? <Pause with wait time> What about 2?” The teacher’s question stimulates further thought by the student. In secondary grades, Driscoll (1999) indicates that well-timed questions to students should help them shift or expand their thinking, or at least have students thinking about what is important to pay attention to during a lesson. When students are examining expressions, a teacher can ask questions to facilitate mathematical flexibility (Heinze, Star, & Verschaffel, 2009). For example, “What other ways can you write that expression to bring out the hidden meaning? How can you write the expression in terms of the important things you care about?”

Score	Description
3	The teacher frequently uses student questions/ comments to coach students, to facilitate conceptual understanding, and boost the conversation. The teacher sequences the student responses that will be displayed in an intentional order, and/or connects different students’ responses to key mathematical ideas.
2	The teacher sometimes uses student questions/ comments to enhance conceptual understanding.
1	The teacher rarely uses student questions/ comments to enhance conceptual mathematical understanding. The focus is more on procedural knowledge of the task verses conceptual knowledge of the content.
0	The teacher never uses student questions/ comments to enhance conceptual mathematical understanding.

REFERENCES

- Adams, R. J., Wu, M. L., & Wilson, M. (2012). The Rasch rating model and the disordered threshold controversy. *Educational and Psychological Measurement, 72*(4), 547-573.
- Akin, O. & Basturk, R. (2012). Keman egitiminde temel becerilerin Rasch olcme modeli ile degerlendirilmesi [The evaluation of the basic skills in violin training by many facet Rasch model]. *Pamukkale University Journal of Education, 31*(1), 175-187. Retrieved from <https://dergipark.org.tr/pauefd/issue/11112/132860>
- AMATYC (1995). *Crossroads in Mathematics: Standards for Introductory College Mathematics before Calculus*. (D. Cohen, Ed.) Memphis, TN: American Mathematical Association of Two-Year Colleges. Retrieved from <http://www.amatyc.org/?page=GuidelineCrossroads>.
- AMATYC (2006). *Beyond Crossroads: Implementing Mathematics Standards in the First Two Years of College*. (R. Blair, Ed.) Memphis, TN: American Mathematical Association of Two-Year Colleges. Retrieved from <http://beyoncrossroads.matyc.org/>.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Andrich, D. (1978). Relationships between the Thurstone and Rasch approaches to item scaling. *Applied Psychological Measurement, 2*(3), 451-462.
- Andrich, D. (2004). Controversy and the Rasch model: A characteristic of incompatible

paradigms? *Medical Care*, 17-116.

Andrich, D. (2013). An expanded derivation of the threshold structure of the polytomous Rasch model that dispels any “threshold disorder controversy”. *Educational and Psychological Measurement*, 73(1), 78-124.

Andrich, D. (2015). The problem with the step metaphor for polytomous models for ordinal assessments. *Educational Measurement: Issues and Practice*, 34(2), 8-14.

Andrich, D., De Jong, J. H. A. L., & Sheridan, B. E. (1997). Diagnostic opportunities with the Rasch model for ordered response categories. *Applications of Latent Trait and Latent Class Models in the Social Sciences*, 59, 70.

Aryadoust, V. (2015). Fitting a mixture Rasch model to English as a foreign language listening tests: The role of cognitive and background variables in explaining latent differential item functioning. *International Journal of Testing*, 15(3), 216-238.

Baghaei, P., & Carstensen, C. H. (2013). Fitting the mixed Rasch model to a reading comprehension test: Identifying reader types. *Practical Assessment, Research, and Evaluation*, 18(1), 5.

Baird, J. A., Hayes, M., Johnson, R., Johnson, S., & Lamprianou, I. (2013). *Marker effects and examination reliability a comparative exploration from the perspectives of generalizability theory, Rasch modelling and multilevel modelling*. Retrieved from <http://www.ofqual.gov.uk/files/2013-01-21-marker-effectsand-examination-reliability.pdf>

Ballantyne, R., Hughes, K., & Mylonas, A. (2002). Developing procedures for implementing peer assessment in large classes using an action research process. *Assessment & Evaluation in Higher Education*, 27(5), 427-441.

- Bandalos, D. L. (2018). *Measurement theory and applications for the social sciences*. Guilford Publications.
- Barbot, B., Tan, M., Randi, J., Santa-Donato, G., & Grigorenko, E. L. (2012). Essential skills for creative writing: Integrating multiple domain-specific perspectives. *Thinking Skills and Creativity*, 7(3), 209-223.
- Barkaoui, K. (2008). *Effects of scoring method and rater experience on ESL essay rating processes and outcomes* (Doctoral dissertation, University of Toronto).
- Barkaoui, K. (2013). Multifaceted Rasch analysis for test evaluation. *The Companion to Language Assessment*, 3, 1301-1322.
- Barker, W., Bressoud, D., Epp, S., Ganter, S., Haver, B., & Pollatsek, H. (2004). *Undergraduate Programs and Courses in the Mathematical Sciences: CUPM Curriculum Guide, 2004*. Mathematical Association of America. 1529 Eighteenth Street NW, Washington, DC 20036-1358.
- Barrett, S. (2001). The impact of training on rater variability. *International Education Journal*, 2(1), 49-58.
- Basturk, S. (2008). Evaluation of teaching practicum course based on the mentors' opinions. *Educational Sciences and Practice*, 7(14), 93-110. Retrieved from <http://ebuline.com/turkce/arsiv/147.aspx>
- Bejar, I.I. (1983). *Achievement testing: recent advances*. Beverly Hills, CA Sage
- Bell, C. A., Gitomer, D. H., McCaffrey, D. F., Hamre, B. K., Pianta, R. C., & Qi, Y. (2012). An argument approach to observation protocol validity. *Educational Assessment*, 17(2-3), 62-87.
- Birenbaum, M. (1996). *Assessment 2000: Towards a pluralistic approach to assessment*.

- In *Alternatives in Assessment of Achievements, Learning Processes and Prior Knowledge* (pp. 3-29). Springer, Dordrecht.
- Bond T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). Lawrence Erlbaum.
- Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (3rd ed.). New York, NJ: Routledge.
- Bonk, W. J., & Ockey, G. J. (2003). A many-facet Rasch analysis of the second language group oral discussion task. *Language Testing*, 20(1), 89-110.
- Boud, D. (1995). Assessment and learning: contradictory or complementary. *Assessment for Learning in Higher Education*, 35-48.
- Bowers, J., & Smith, W. (2016). Repurposing the MCOP² observation protocol to survey students' views of an active learning course redesign. In *Proceedings of the MTEP Conference, Atlanta: GA*.
- Bramley, T. (2007). Paired comparison methods. In Newton, P., J-A. Baird, H. Goldstein, H. Patrick, and P. Tymms (Ed.). *Techniques for Monitoring the Comparability of Examination Standards*, 264-294. London: QCA.
- Brown, S., Rust, C., & Gibbs, G. (1994). *Strategies for diversifying assessment in higher education*. Oxford: Oxford Centre for Staff Development.
- Cabello, V. M., & Topping, K. J. (2020). Pre-service teachers' conceptions about the quality of explanations for the science classroom in the context of peer assessment. *LUMAT: International Journal on Math, Science and Technology Education*, 8(1), 297-318.
- Casabianca, J. M., Junker, B. W., & Patz, R. J. (2016). Hierarchical rater models. In W. J.

- van der Linden (Ed.), *Handbook of Item Response Theory* (Vol. 1, pp. 449-465).
Boca Raton, FL: Chapman & Hall/CRC.
- Cerezci, B. (2020). Mining the gap: Analysis of early mathematics instructional quality in pre-kindergarten classrooms. *Early Education and Development*, 1-24.
<https://doi.org/10.1080/10409289.2020.1775438>
- Cetin, B., & Ilhan, M. (2017). An analysis of rater severity and leniency in open-ended mathematic questions rated through standard rubrics and rubrics based on the SOLO taxonomy. *Education and Science*, 42(189), 217-247.
<https://doi.org/10.15390/EB.2017.5082>
- Chen, W. H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, 22(3), 265-289.
- Chen, Y., Tong, Y., Xue, Z., Cheng, Y., & Li, X. (2020). Evaluation of the Reliability and Validity of the Behavioral Indicators of Infant Pain Scale in Chinese Neonates. *Pain Management Nursing*, 21(5), 456-461.
- Chen, Y., Yim, R. A., Kogen, R., Stieff, M., & Superfine, A. C. (2020). Rethinking Rater Effects When Using Teacher Observation Protocols. In *Proceedings of the International Conference of the Learning Sciences (ICLS)* (pp. 2022-2029).
- Christensen, K. B., Makransky, G., & Horton, M. (2017). Critical values for Yen's Q3: Identification of local dependence in the Rasch model using residual correlations. *Applied Psychological Measurement*, 41(3), 178-194.
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1), 37-46.

- Cohen, J. (1968). Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4), 213.
- Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability of scores and profiles*. New York: Wiley
- Darling-Hammond, L. (2010). Teacher education and the American future. *Journal of Teacher Education*, 61(1-2), 35-47.
- Darling-Hammond, L., Newton, S. P., & Wei, R. C. (2013). Developing and assessing beginning teacher effectiveness: The potential of performance assessments. *Educational Assessment, Evaluation and Accountability*, 25(3), 179-204.
- Davidson, F. (1991). Statistical support for training in ESL composition rating. In Hamp-Lyons, L. (Ed.), *Assessing second language writing in academic contexts* (pp. 155–164). Norwood, NJ: Ablex.
- DeMars, C. (2010). *Item response theory*. Oxford University Press.
- de Moira, A. P., Massey, C., Baird, J. A., & Morrissy, M. (2002). Marking consistency over time. *Research in Education*, 67(1), 79-87.
- Devellis, R. F. (2011). *Scale development: Theory and applications*. Sage Publications.
- Eckes, T. (2005). Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis. *Language Assessment Quarterly: An International Journal*, 2(3), 197-221.
- Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing*, 25(2), 155-185.

- Eckes, T. (2009). Many-facet Rasch measurement. In S. Takala (Ed.), *Reference supplement to the manual for relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment* (Section H). Strasbourg, France: Council of Europe/Language Policy Division.
- Eckes, T. (2011). Item banking for C-tests: A polytomous Rasch modeling approach. *Psychological Test and Assessment Modeling*, 53(4), 414.
- Eckes T. (2015). *Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments* (2nd ed.). Frankfurt am Main, Germany: Peter Lang.
- Eckes, T. (2019). Implications for rater-mediated language assessment. *Quantitative Data Analysis for Language Assessment Volume I: Fundamental Techniques*.
- Engelhard Jr, G. (1992). The measurement of writing ability with a many-faceted Rasch model. *Applied Measurement in Education*, 5(3), 171-191.
- Engelhard Jr, G. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement*, 31(2), 93-112.
- Engelhard, G. (2002). Monitoring raters in performance assessments. *Large-Scale Assessment Programs for All Students: Validity, Technical Adequacy, and Implementation*, 261-287.
- Engelhard Jr, G. (2013). *Invariant measurement: Using Rasch models in the social, behavioral, and health sciences*. Routledge.
- Engelhard Jr, G., & Wind, S. A. (2013). Rating quality studies using Rasch measurement

- theory. Research Report 2013-3. *College Board*.
- Engelhard, G., & Stone, G.E. (1998). Evaluating the quality of ratings obtained from standard-setting judges. *Educational and Psychological Measurement*, 58(2), 179-196. <https://doi.org/10.1177/0013164498058002003>
- Engelhard Jr, G., Wang, J., & Wind, S. A. (2018). A tale of two models: Psychometric and cognitive perspectives on rater-mediated assessments using accuracy ratings. *Psychological Test and Assessment Modeling*, 60(1), 33-52.
- Erman Aslanoglu, A., Karakaya, I., & Sata, M. (2020). Evaluation of university students' rating behaviors in self and peer rating process via many facet Rasch model. *Eurasian Journal of Educational Research (EJER)*, (89).
- Farrokhi, F., Esfandiari, R., & Dalili, M. V. (2011). Applying the many-facet Rasch model to detect centrality in self-assessment, peer-assessment and teacher assessment. *World Applied Sciences Journal*, 15(11), 76-83.
- Farrokhi, F., Esfandiari, R., & Schaefer, E. (2012). A many-facet Rasch measurement of differential rater severity/leniency in three types of assessment. *JALT Journal*, 34(1), 79-101.
- Feiman-Nemser, S. (2012). *Teachers as Learners*. Harvard Education Press. 8 Story Street First Floor, Cambridge, MA 02138.
- García-Pérez, M. A. (2017). An analysis of (dis) ordered categories, thresholds, and crossings in difference and divide-by-total IRT models for ordered responses. *The Spanish Journal of Psychology*, 20.
- Garrett, L., Guest, K. B., Tameru, A., & Karatas, Z. *Building an active learning nucleus: Examining a case study*. Association of Public Land-Grant Universities.

<https://www.aplu.org/projects-and-initiatives/stem-education/mathematics-teacher-education-partnership/mtep-conferences-meetings/mtep6-materials/25-Building-an-Active-Learning-Nucleus.pdf>

Gleason, J., & Cofer, L. D. (2014). Mathematics classroom observation protocol for practices results in undergraduate mathematics classrooms. In *Proceedings of the 17th Annual Conference on Research on Undergraduate Mathematics Education* (pp. 93-103).

Gleason, J., Livers, S. D., & Zelkowski, J. (2015). *Mathematics classroom observation protocol for practices: Descriptors manual*. University of Alabama. Retrieved from jgleason.people.ua.edu/mcop2.html.

Gleason, J., Livers, S., & Zelkowski, J. (2017). Mathematics classroom observation protocol for practices (MCOP²): A validation study. *Investigations in Mathematics Learning*, 9(3), 111-129.

Gleason, J., Zelkowski, J., Livers, S., Dantzler, J., & Khalilian, J. (2014). Mathematics classroom observation protocol for practices: Validity and reliability. Preprint.

Gomez, K., Kyza, E. A., & Mancevice, N. (2018). *Participatory design and the learning sciences*. Taylor and Francis.

Guler, N. (2008). *Klasik test kurami genellenebilirlik kurami ve Rasch modeli üzerine bir araştırma* (Unpublished doctoral dissertation). Hacettepe University, Ankara.

Haiyang, S. (2010). An application of classical test theory and many-facet Rasch measurement in analyzing the reliability of an English test for non-English major graduates. *Chinese Journal of Applied Linguistics (Foreign Language Teaching & Research Press)*, 33(2).

- Han, C. (2015). Investigating rater severity/leniency in interpreter performance testing: A multifaceted Rasch measurement approach. *Interpreting, 17*(2), 255-283.
- Han, C. (2019). Introduction to Many-Facet Rasch Measurement: Analyzing and evaluating rater-mediated assessments. *Measurement: Interdisciplinary Research and Perspectives, 19*(2), 113-116.
- Hayes, A. F., & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication methods and measures, 1*(1), 77-89.
- Henning, G. (1992). Dimensionality and construct validity of language tests. *Language Testing, 9*(1), 1-11.
- Henning, G. (1996). Accounting for nonsystematic error in performance ratings. *Language Testing, 13*(1), 53-61.
- Hoskens, M., & Wilson, M. (2001). Real-time feedback on rater drift in constructed-response items: An example from the Golden State Examination. *Journal of Educational Measurement, 38*(2), 121-145.
- Hoyt, W. T., & Kerns, M. D. (1999). Magnitude and moderators of bias in observer ratings: A meta-analysis. *Psychological Methods, 4*(4), 403.
- İlhan, M. (2016). Comparison of ability estimates calculated according to classical test theory and multi-faceted Rasch model in measurements made with open-ended questions. *Hacettepe University Journal of Education Faculty, 31*(2), 346-368.
- Johnson, D. W., Johnson, R. T., Holubec, E. J., & Holubec, E. J. (1994). *The new circles of learning: Cooperation in the classroom and school*. ASCD.
- Johnson, E. S., Moylan, L. A., Crawford, A., & Zheng, Y. (2019). Developing a comprehension instruction observation rubric for special education

- teachers. *Reading & Writing Quarterly*, 35(2), 118-136.
- Johnson, R. L., Penny, J. A., & Gordon, B. (2008). *Assessing performance: Designing, scoring, and validating performance tasks*. Guilford Press.
- Jones, E., & Bergin, C. (2019). Evaluating teacher effectiveness using classroom observations: A Rasch analysis of the rater effects of principals. *Educational Assessment*, 24(2), 91-118.
- Karakaya, İ. (2015). Comparison of self, peer and instructor assessments in the portfolio assessment by using many facet RASCH model. *Journal of Education and Human Development*, 4(2), 182-192.
- Kim, Y., Park, I., & Kang, M. (2012). Examining rater effects of the TGMD-2 on children with intellectual disability. *Adapted Physical Activity Quarterly*, 29(4), 346-365. <https://doi.org/10.1123/apaq.29.4.346>
- Knight, S. L., Lloyd, G. M., Arbaugh, F., Gamson, D., McDonald, S. P., & Nolan Jr, J. (2014). Professional development and practices of teacher educators. *Journal of Teacher Education*, 65(4), 268-270.
- Knoch, U., & Chapelle, C. A. (2018). Validation of rating processes within an argument-based framework. *Language Testing*, 35(4), 477-499.
- Knoch, U., Read, J., & von Randow, J. (2007). Re-training writing raters online: How does it compare with face-to-face training? *Assessing Writing*, 12(1), 26-43.
- Kondo-Brown, K. (2002). A FACETS analysis of rater bias in measuring Japanese second language writing performance. *Language Testing*, 19(1), 3-31.
- Kramer, J., Kielhofner, G., Lee, S. W., Ashpole, E., & Castle, L. (2009). Utility of the Model of Human Occupation Screening Tool for detecting client

- change. *Occupational Therapy in Mental Health*, 25(2), 181-191.
- Kubinger, K. D. (2009). Applications of the linear logistic test model in psychometric research. *Educational and Psychological Measurement*, 69(2), 232-244.
- Leckie, G., & Baird, J. A. (2011). Rater effects on essay scoring: A multilevel analysis of severity drift, central tendency, and rater experience. *Journal of Educational Measurement*, 48(4), 399-418.
- Li, Y., Li, S., & Wang, L. (2010). Application of a general polytomous testlet model to the reading section of a large-scale English language assessment. *ETS Research Report Series*, 2010(2), i-34.
- Linacre, J. M. (1989). Many-faceted Rasch measurement. Chicago: MESA Press.
- Linacre, J. M., & Wright, B. D. (1993). *A user's guide to BIGSTEPS: Rasch-model computer program* (pp. 1-122). San Diego, CA: Mesa Press.
- Linacre, J. M. (1995). *Rasch Measurement Transactions, Part 1*. MESA Press, 5835 S. Kimbark Avenue, Chicago, IL 60637.
- Linacre, J. M. (1996). Generalizability theory and many-facet Rasch measurement. *Objective measurement: Theory into practice*, 3, 85-98. Retrieved from <https://files.eric.ed.gov/fulltext/ED364573.pdf>
- Linacre, J. M. (1998). Detecting multidimensionality: which residual data-type works best? *Journal of outcome measurement*, 2, 266-283.
- Linacre, J. M. (1999). Investigating rating scale category utility. *Journal of Outcome Measurement*, 3(2), 103-122.
- Linacre, J. M. (2001). Who wrote Paul's epistles. *Rasch Measurement Transactions*, 15(1), 800-801.

- Linacre, J. M., & Wright, B. D. (2002). Construction of measures from many-facet data. *Journal of Applied Measurement, 4*, 486–512.
- Linacre, J. M. (2003). *Winsteps computer program*, version 3:48, Chicago: www.winsteps.com.
- Linacre, J. M. (2004). Rasch model estimation: Further topics. *Journal of Applied Measurement, 5*(1), 95-110.
- Linacre, J. M. (2006). Rasch analysis of rank-ordered data. *Journal of Applied Measurement, 7*(1), 129-139.
- Linacre, J. M. (2007). *A user's guide to WINSTEPS-MINISTEP: Rasch-model computer programs*. Chicago, IL: winsteps.com.
- Linacre, J. M. (2012). *Winsteps Rasch Tutorial 4*. Retrieved from <http://www.winsteps.com/a/winsteps-tutorial-4.pdf>
- Linacre, J. M. (2017). *Winsteps Rasch measurement computer program*. Beaverton, OR: Winsteps.com.
- Linacre, J. M. (2020). *Facets computer program for many-facet Rasch measurement*, version 3.83.4. Beaverton, Oregon: Winsteps.com
- Livers, S. D., Zelkowski, J., Harbour, K. E., McDaniel, S. C., & Gleason, J. (2020). An examination of the relationships of mathematics self-efficacy and teaching practices among elementary, secondary, and special education educators. *Investigations in Mathematics Learning, 12*(2), 96-109.
- Lord, F. I., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Cambridge, MA: Addison-Wesley.
- Lunz, M. E., & Stahl, J. A. (1990). Judge consistency and severity across grading

- periods. *Evaluation & the Health Professions*, 13(4), 425-444.
- Lunz, M. E., Wright, B. D., & Linacre, J. M. (1990). Measuring the impact of judge severity on examination scores. *Applied Measurement in Education*, 3(4), 331-345.
- Lunz, M. E., Stahl, J. A., & Wright, B. D. (1996). The invariance of judge severity calibrations. In G. Engelhard & M. Wilson (Eds.), *Objective Measurement: Theory into Practice* (Vol. 3, pp. 99-112). Norwood, NJ: Ablex.
- Lynch, B. K., & McNamara, T. F. (1998). Using G-theory and many-facet Rasch measurement in the development of performance assessments of the ESL speaking skills of immigrants. *Language Testing*, 15(2), 158-180.
- Mannarini, S. (2009). A method for the definition of a self-awareness behavior dimension with clinical subjects: A latent trait analysis. *Behavior Research Methods*, 41(4), 1029-1037.
- Marais I. (2013). Local dependence. In Christensen K. B., Kreiner S., Mesbah M. (Eds.), *Rasch Models in Health* (pp. 111-130). London, England: Wiley.
- Marais, I., & Andrich, D. (2008). Formalizing dimension and response violations of local independence in the unidimensional Rasch model. *Journal of Applied Measurement*, 9(3), 200-215.
- Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47(2), 149-174.
- Matsuno, S. (2009). Self-, peer-, and teacher-assessments in Japanese university EFL writing classrooms. *Language testing*, 26(1), 075-100.
- MacMillan, P. D. (2000). Classical, generalizability, and multifaceted Rasch detection of

- interrater variability in large, sparse data sets. *The Journal of Experimental Education*, 68(2), 167-190.
- Marais, H. (2013). *South Africa pushed to the limit: The political economy of change*. Zed Books Ltd..
- McNamara, T. F., & Adams, R. J. (1991). Exploring rater behavior with Rasch techniques. *Language Testing Research Colloquium*, 1-29. Retrieved from <https://files.eric.ed.gov/fulltext/ED345498.pdf>
- McNamara, T. (1996). *Measuring second language proficiency*. London: Longman.
- McNamara, T., & Knoch, U. (2012). The Rasch wars: The emergence of Rasch measurement in language testing. *Language Testing*, 29(4), 555-576.
- Meuwissen, K. W., Choppin, J. M., Cloonan, K. D., & Shang-Butler, H. (2016, April). *Celebrations, confessionals, and creative interpretations: Representing teaching practice in the edTPA as a high-stakes certification exam in New York and Washington States*. Paper presented at the Annual Meeting of the American Educational Association, Washington, DC.
- Mulqueen, C., Baker, D., & Dismukes, R. K. (2000, April). Using multifacet Rasch analysis to examine the effectiveness of rater training. In *15th Annual Conference for the Society for Industrial and Organizational Psychology*.
- Mun, E. Y. (2005). Rater Agreement-Kappa. *Encyclopedia of statistics in behavioral science*.
- Muraki, E., & Bock, R. D. (1993). *PARSCALE: IRT based test scoring and item analysis for graded open-ended exercises and performance tasks*. Scientific Software International.

- Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement, 4*(4), 386-422.
- Myford, C. M., & Wolfe, E. W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part II. *Journal of Applied Measurement, 5*(2), 189-227.
- Myford, C. M., & Wolfe, E. W. (2009). Monitoring rater performance over time: A framework for detecting differential accuracy and differential scale category use. *Journal of Educational Measurement, 46*(4), 371-389.
- National Board for Professional Teaching Standards. (2000). What teachers should know and be able to do. Southfield, MI: Author.
- NCTM. (2000). *Principles and Standards for School Mathematics*. Washington, D.C.: National Council of Teachers of Mathematics.
- NGACBP & CCSSO. (2010). *Common Core State Standards for Mathematics*. Washington, DC: National Governors Association Center for Best Practices & Council of Chief State School Officers.
- North, B. (2003). Scales for rating language performance: Descriptive models, formulation styles, and presentation formats. *TOEFL Research Paper*. Princeton, NJ: Educational Testing Service.
- Nunnally, J. C. (1978), *Psychometric Theory*, 2nd ed., New York: McGraw-Hill.
- Orsmond, P., Merry, S., & Reiling, K. (1996). The importance of marking criteria in the use of peer assessment. *Assessment & Evaluation in Higher Education, 21*(3), 239-250.

- Orsmond, P., Merry, S., & Reiling, K. (1997). A study in self-assessment: tutor and students' perceptions of performance criteria. *Assessment & Evaluation in Higher Education*, 22(4), 357-368.
- Orsmond, P., Merry, S., & Reiling, K. (2000). The use of student derived marking criteria in peer and self-assessment. *Assessment & Evaluation in Higher Education*, 25(1), 23-38.
- Puhl, C. A. (1997). Develop, not judge. Continuous assessment in the ESL classroom. *Forum*, 35(2), 25-31.
- Raczynski, K. R., Cohen, A. S., Engelhard Jr, G., & Lu, Z. (2015). Comparing the effectiveness of self-paced and collaborative frame-of-reference training on rater accuracy in a large-scale writing assessment. *Journal of Educational Measurement*, 52(3), 301-318.
- Saal, F.E., Downey, R.G. and Lahey, M.A. (1980). Rating the ratings: assessing the psychometric quality of rating data, *Psychological Bulletin*, 88(2), 413-428.
- Saito, H. (2008). EFL classroom peer assessment: Training effects on rating and commenting. *Language Testing*, 25(4), 553-581.
- Sambell, K., & McDowell, L. (1998). The construction of the hidden curriculum: messages and meanings in the assessment of student learning. *Assessment & Evaluation in Higher Education*, 23(4), 391-402.
- Sandholtz, J. H., & Shea, L. M. (2012). Predicting performance: A comparison of university supervisors' predictions and teacher candidates' scores on a teaching performance assessment. *Journal of Teacher Education*, 63(1), 39-50.
- Santagata, R., & Sandholtz, J. H. (2019). Preservice teachers' mathematics teaching

- competence: Comparing performance on two measures. *Journal of Teacher Education*, 70(5), 472-484.
- Schaefer, E. (2008). Rater bias patterns in an EFL writing assessment. *Language Testing*, 25(4), 465-493. <https://doi.org/10.1177/0265532208094273>
- Scullen, S. E., Mount, M. K., & Goff, M. (2000). Understanding the latent structure of job performance ratings. *Journal of Applied Psychology*, 85(6), 956.
- Shachar, H., & Sharan, S. (1994). Talking, relating, and achieving: Effects of cooperative learning and whole-class instruction. *Cognition and Instruction*, 12(4), 313-353.
- Sims, M. E., Cox, T. L., Eckstein, G. T., Hartshorn, K. J., Wilcox, M. P., & Hart, J. M. (2020). Rubric Rating with MFRM versus Randomly Distributed Comparative Judgment: A Comparison of Two Approaches to Second-Language Writing Assessment. *Educational Measurement: Issues and Practice*, 39(4), 30-40.
- Sluismans, D., & Prins, F. (2006). A conceptual framework for integrating peer assessment in teacher education. *Studies in Educational Evaluation*, 32(1), 6-22.
- Smith, A. M. (2000). The impact of scale characteristics on the dimensionality of the service quality construct. *Service Industries Journal*, 20(3), 167-190.
- Smith Jr, E. V. (2002). Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals. *Journal of Applied Measurement*, 3(2), 205-231.
- Smith, E. (2005). Effect of item redundancy on Rasch item and person estimates. *Journal of Applied Measurement*, 6(2), 147-163.
- Smith, R. M., & Miao, C. Y. (1994). Assessing unidimensionality for Rasch measurement. *Objective Measurement: Theory into Practice*, 2, 316-327.

- Stiggins, R. J. (1991). Relevant classroom assessment training for teachers. *Educational Measurement: Issues and Practice*, 10(1), 7-12.
- Sudweeks, R.R., Reeve, S., & Bradshaw, W.S. (2005). A comparison of generalizability theory and many-facet Rasch measurement in an analysis of college sophomore writing. *Assessing Writing*, 9(3), 239-261.
<https://doi.org/10.1016/j.asw.2004.11.001>
- Télez, K. (2016). *The teaching instinct: Explorations into what makes us human*. Routledge.
- Topping, K. J. (2009). Peer assessment. *Theory into Practice*, 48(1), 20-27.
- Verloop, N., & Wubbels, T. (2000). Some major developments in teacher education in the Netherlands and their relationship with. *Trends in Dutch teacher education*, 19.
- von Eye, A., & Von Eye, M. (2005). Can one use Cohen's kappa to examine disagreement? *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, 1(4), 129.
- Wainer, H., & Thissen, D. (1996). How is reliability related to the quality of test scores? What is the effect of local dependence on reliability? *Educational Measurement: Issues and Practice*, 22-29.
- Wainwright, C. L., Flick, L., & Morrell, P. (2003). The development of instruments for assessment of instructional practices in standards-based teaching. *Journal of Mathematics and Science: Collaborative Explorations*, 6(1), 21-46.
- Walkington, C., Arora, P., Ihorn, S., Gordon, J., Walker, M., Abraham, L., & Marder, M. (2012). *Development of the UTeach observation protocol: A classroom*

- observation instrument to evaluate mathematics and science teachers from the UTeach preparation program*. Unpublished paper. Southern Methodist University.
- Wang, W. C., Cheng, Y. Y., & Wilson, M. (2005). Local item dependence for items across tests connected by common stimuli. *Educational and Psychological Measurement, 65*(1), 5-27.
- Watley, L. E. (2017). *Structural validity and reliability of two observation protocols in college mathematics*. (Doctoral dissertation, University of Alabama Libraries).
- Wei, R. C., & Pecheone, R. L. (2010). Performance-based assessments as high-stakes events and tools for learning. *Handbook of teacher assessment and teacher quality, 69-132*.
- Weigle, S. (1998). Using FACETS to model rater training effects. *Language Testing, 15*(2), 263-287. <https://doi.org/10.1177/026553229801500205>
- Weigle, S. (1999). Investigating rater/prompt interactions in writing assessment: Quantitative and qualitative approaches. *Assessing Writing, 6*(2), 145-178. [https://doi.org/10.1016/S1075-2935\(00\)00010-6](https://doi.org/10.1016/S1075-2935(00)00010-6)
- Wind, S. A. (2014). Examining rating scales using Rasch and Mokken models for rater-mediated assessments. *Journal of Applied Measurement, 15*(2), 100-132.
- Wind, S. A. (2019). Examining the impacts of rater effects in performance assessments. *Applied Psychological Measurement, 43*(2), 159-171.
- Wind, S. A., & Engelhard Jr, G. (2013). How invariant and accurate are domain ratings in writing assessment? *Assessing Writing, 18*(4), 278-299.
- Wind, S. A., & Engelhard Jr, G. (2017). Exploring rater errors and systematic biases

- using adjacent-categories Mokken models. *Psychological Test and Assessment Modeling*, 59(4), 493-515.
- Wind, S. A., & Peterson, M. E. (2018). A systematic review of methods for evaluating rating quality in language assessment. *Language Testing*, 35(2), 161-192.
- Wolfe, E. W. (2004). Identifying rater effects using latent trait models. *Psychology Science*, 46, 35-51.
- Wolfe, E. W., Chiu, C. W., & Myford, C. M. (2000). Detecting rater effects in simulated data with a multi-faceted Rasch rating scale model. *Objective Measurement: Theory into Practice*, 5, 147-164.
- Wolfe, E. W., Jiao, H., & Song, T. (2015). A Family of Rater Accuracy Models. *Journal of Applied Measurement*, 16(2), 153-160.
- Wolfe, E. W., & McVay, A. (2012). Application of latent trait models to identifying substantively interesting raters. *Educational Measurement: Issues and Practice*, 31(3), 31-37.
- Wolfe, E. W., Myford, C. M., Engelhard, G., Jr. & Manolo, J. R. (2007). *Monitoring reader performance and DRIFT in the AP English Literature and Composition Examination using benchmark essays* (Research Report 2007-2). New York, NY: The College Board.
- Wright, B. D., & Linacre, J. M. (1989). Observations are always ordinal; measurements, however, must be interval. *Archives of Physical Medicine and Rehabilitation*, 70(12), 857-860.
- Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. Chicago: MESA Press.
- Wu, S. M., & Tan, S. (2016). Managing rater effects through the use of FACETS

- analysis: the case of a university placement test. *Higher Education Research & Development*, 35(2), 380-394.
- Yan, X. (2014). An examination of rater performance on a local oral English proficiency test: A mixed-methods approach. *Language Testing*, 31(4), 501-527.
- Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement*, 8(2), 125-145.
- Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, 30(3), 187-213.
- Yurdabakan, İ., ve Oğlun, M. (2011). Öz ve akran değerlendirmenin öğrenme ve bilişüstü bilgi üzerindeki etkisi: sonuçsal geçerlik. 2nd International Conference on New Trends in Education and Their Implications 27-29 April, 2011 Antalya-Turkey.
- Lawrence, S. et al. (2001). Persistence of Web References in Scientific Research. *Computer*. 34, 26-31. doi:10.1109/2.901164.
- Zelkowski, J., & Gleason, J. (2016). Using the MCOP2 as a grade bearing assessment of clinical field observations. In *Proceedings of the fifth annual Mathematics Teacher Education Partnership conference*. Washington, DC: Association of Public and Land-grant Universities.
- Zelkowski, J., Gleason, J., & Livers, S. (2017, July). *Measuring mathematics classroom interactions: An observation protocol reinforcing the development of conceptual understanding*. Paper presented at the 13th International Congress on Mathematical Education, Hamburg, Germany.

VITA

Dr. CHUNLING NIU

EDUCATION

Western Kentucky University Bowling Green, KY
Professional Certificate in Measurement, Evaluation and Research August 2016

Western Kentucky University Bowling Green, KY
Ed.D., Educational Leadership (Educational Assessment & Program Evaluation, Organizational Leadership) May 2015
Dissertation: An Experimental Study on the Impact of Intercultural Sensitivity on Language Motivation

Guangdong Foreign Studies University Guangzhou City, China
A.M., Applied Linguistics (Translation Studies; Simultaneous Interpretation) July 2006
Qualifying paper: Translation of Appellations in Chinese Classical Poetry

Guangdong Foreign Studies University Guangzhou City, China
A.B., English Language and Literature (Translation & Interpretation concentration) July 2000
Senior Honors Thesis: The Greek Tragedy: One Interpretation of *Tess of the d'Urbervilles*

PROFESSIONAL EXPERIENCE

Research Administrative Coordinator Principal, the Training Resources Center, College of Social Work, University of Kentucky Fall 2016-Present

Clinical Program Evaluator, Western Kentucky University Summer 2015-Summer 2016

Clinical Program Evaluator, Western Kentucky University Summer 2015-Summer 2016

Teaching Assistant , Education Leadership Doctoral Program, Western Kentucky University	Summer 2016
Teaching Assistant , Education Leadership Doctoral Program, Western Kentucky University	Spring 2016
Research Assistant , Western Kentucky University	Spring 2014- Spring 2015
Research Assistant , Western Kentucky University	Fall 2013-Spring 2014
Intern , Faculty Center for Excellence in Teaching (FaCET), Western Kentucky University	Spring 2013
Graduate Assistant , Chinese Flagship Program, University College, Western Kentucky University	Fall 2011-Spring 2013
Teaching Assistant , Education Leadership Doctoral Program, Western Kentucky University	Fall 2012
Teaching Assistant , Education Leadership Doctoral Program, Western Kentucky University	Summer 2012
Graduate Assistant , Confucius Institute, Western Kentucky University	Spring 2011
Lecturer , Sichuan International Studies University, China	2006-2010

PUBLICATIONS

Kato, H., Niu, C. , Jin, R. & Yarra, P. (Under review). Early upper endoscopy decreases resource utilization in upper gastrointestinal bleeding: A quasi-experimental study.	2021
Miller, J., & Niu, C. (Under review). Examining self-care practice frequency among social workers: An international comparison.	2021

- Miller, J., **Niu, C.**, & Moody, S. (Under review). Examining self-care practices among public child welfare workers: Implications for employees, employers, and professional member organizations. 2021
- Niu, C.**, Sampson, S., Bradley, K., Kato, H., & Jin, R. (in working progress). Cross-nation comparison of the 4-grade mathematics classroom instruction from the teachers' perspective: A multilevel Rasch modeling analysis of the TIMSS 2015 data 2020
- Niu, C.**, Bradley, K., Sampson, S., Jin, R., Xia, Y., Shen, L., Zhang, J., Wu, R. & Li, N. (in working progress). Simulation study: Evaluating Rater Category Ordering with the JML-Rasch-MFRM Model in Facets 2020
- Niu, C.**, Bradley, K., Wilson, N., & Jin, R. (Under review). Simulation study: The Effects of Missing Data on the Partial Credit Trees DIF Detection Performance 2020
- Niu, C.**, Bradley, K., Jin, R., & Wilson, N. (Under review). A simulation study: Comparing the DIF Detection Capacities of the Rasch Trees Model Tests to Two Common DIF Approaches for Partial Credit Models. 2020
- Niu, C.**, Miller, J., Wu, R., Zhou, X., Shen, L., & Qiu, C. (Under review). Meta-analysis: Evaluating the impact of school social work from 2008 to 2018. 2020
- Miller, J., Cooley, M., **Niu, C.**, Segress, M., Fletcher, J., Bowman, K., & Pachner, T. Assessing the impact of a virtual support group on adoptive parent stress and competence: Results from an urban/rural pilot study. *Child & Family Social Work*, <https://doi.org/10.1111/cfs.12826>. 2020
- Miller, J., **Niu, C.**, & Moody, S. (2020). Child welfare workers and peritraumatic distress: The impact of COVID-19. *Children and Youth Services Review*, *119*, 105508. <https://doi.org/10.1016/j.chilyouth.2020.105508>. 2020
- Niu, C.** & Li, N. (in revision). A simulation study: Comparing multiple testing correction methods in the Rasch Trees DIF analysis. 2020
- Miller, J., Bode, M., Adcock, A., **Niu, C.**, & Freeman, D. (Accepted). I know what I know... unless I don't: Examining faculty knowledge about social work licensing. 2020

- Miller, J., **Niu, C.**, Womack, R., & Shalash, N. Supporting adoptive parents: A study on personal self-care. *Adoption Quarterly*, 22(2), 157-171. 2019
- Miller, J., Cooley, M., **Niu, C.**, Segress, M., Fletcher, J., Bowman, K., & Littrell, L. Virtual support groups among adoptive parents: Ideal for information seeking? *Journal of Technology in Human Services*. DOI: 10.1080/15228835.2019.1637320 2019
- Miller, J., Lee, J., **Niu, C.**, Grise-Owens, E., & Bode, M. Self-Compassion as a predictor of self-care: A study of social work clinicians. *Clinical Social Work Journal*, 47(1), 1-11, DOI:10.1007/s10615-019-00710-6 2019
- Niu, C.**, Miller, J., Bowman, K., Fletcher, J., & Segress, M. (under review). Adoptive parents' virtual support groups: A social network analysis perspective. 2019
- Miller, J., **Niu, C.**, & Moody, S. Investigating the Child Trauma Knowledge of Adoptive Parents: An Exploratory Study. *Adoption Quarterly*, 21(3), 1-18, DOI: 10.1080/10926755.2019.1579134. 2019
- Miller, J., Koh, E., **Niu, C.**, Moody, S., & Bode, M. Examining child trauma knowledge among kin caregivers: Implications for practice, policy, and research. *Children and Youth Services Review*, 100, 112-118. 2019
- Miller, J., Donohue-Dioh, J., **Niu, C.**, Grise-Owens, E., & Poklembova, Z. Examining the self-care practices of social workers in child welfare: A national perspective. *Children and Youth Services Review*, 99, 240-245. 2019
- Miller, J., M., & Cooley, **Niu, C.**, Segress, M., Fletcher, J., Bowman, K., & Littrell, L. Support, information seeking, and homophily in a virtual support group for adoptive parents: Impact on perceived empathy. *Children and Youth Services Review*, 101, 151-156. 2019
- Miller, J., **Niu, C.**, Womack, R., & Shalash, N. Supporting Adoptive Parents: A Study on Personal Self-Care. *Adoption Quarterly*, 22(2), 157-171. 2019
- Miller, J., Koh, E., **Niu, C.**, Bode, M., & Moody, S. Examining child trauma knowledge among kin caregivers: Implications for practice, policy, and research. *Children and Youth Services Review*, 100, 112-118. 2019

- Miller, J., Donohue-Dioh, J., Larkin, S., **Niu, C.**, & Womack, R. Exploring the self-care practices of practicum supervisors: Implications for field education. *The Field Educator*, 8.2, 1-20. 2018
- Miller, J., **Niu, C.** (under review). Examining the self-care practices of gerontological social workers: An exploratory study. *Journal of Gerontological Social Work*. 2018
- Miller, J., Donohue-Dioh, J.**, Larkin, S.**, Gibson, A., & **Niu, C.** (in revision). *Examining the self-care practices of social work administrators: A cross-sectional investigation*. 2018
- Miller, J., Grise-Owens, E., **Niu, C.**, & Shalash, N. (Accepted). Examining self-care practices of bachelor prepared social workers: Implications for undergraduate education. *Journal of Baccalaureate Social Work*. 2018
- Miller, J., Barnhart, S., **Niu, C.**, Donohue-Dioh, J., & Feld, H. (in review). *Self-care practices among nurses: An exploratory examination*. 2018
- Miller, J., Donohue-Dioh, J., **Niu, C.**, & Shalash, N. Self-care among child welfare workers: A research brief. *Child and Youth Services Review*, 84, 137 - 142. 2018
- Miller, J., **Niu, C.**, Sauer, C., Bowman, K., Segress, M., & Benner, K. Foster Parents' Knowledge of Child Trauma: An Exploratory Study. *Journal of Aggression, Maltreatment, and Trauma*, 27(5), 505-522. DOI: 10.1080/10926771.2017.1422839. 2018
- Miller, J., Sauer, C., Bowman, K., Thrasher, S., Benner, K., Segress, M., & **Niu, C.** Conceptualizing Adoptive Parent Support Groups: A Mixed-Method Process. *Adoption Quarterly*, 1, 41-57. 2018
- Dietrich, S., **Niu, C.**, & Zippay, C. The Path to a Model Curriculum in Clinical Teacher Education. *GATEways to Teacher Education*, 28(2), 44-54. 2018
- Houchens, G., **Niu, C.**, Zhang, J., Miller, S., & Norman, A. Do Differences in High School Principal and Assistant Principal Perceptions Relate to Student Achievement? *The NASSP (National Association for Secondary School Principals) Bulletin*. 102(1), 38-57. 2018

- Niu, C.**, Everson, K., Dietrich, S., & Zippay, C. Validity Issues in Assessing Dispositions: The Confirmatory Factor Analysis of A Teacher Dispositions Form. *Journal of the Southeastern Regional Association of Teacher Educators*, 26(2), 41-49. 2017
- Miller, J., Collins-Camargo, C., **Niu, C.**, & Jones, B. Exploring Member Perspectives on Participation on Child Welfare Citizen Review Panels: A National Study. *Child Abuse and Neglect: The International Journal*, 72, 352 – 359. 2017
- Miller, J., Pope, N., Lee, J., Grise-Owens, E., & **Niu, C.** (Under Review). Exploring the Self-Care Practice of Clinical Social Workers: Implications for Practice, Education, and Research. *Clinical Social Work Journal*. 2017
- Houchens, G., Zhang, J., Davis, K., **Niu, C.**, Chon, K., & Miller, S. The Impact of Positive Behavior Interventions and Supports on Kentucky Teachers' Perceptions of Teaching Conditions and Student Achievement. *Journal of Positive Behavior Interventions*, 19 (3), 168-179. 2017
- Miller, J., Benner, K., Thrasher, S, Pope, N., Dumas, T., Damron, J., Segress, M., & **Niu, C.** Planning A Mentorship Initiative for Foster Parents: Does Gender Matter? *Evaluation and Program Planning*, 64, 78-84. 2017
- Miller, J., Benner, K., Pope, N., Dumas, T., Damron, J., Segress, M., Sloan, M., Thrasher, S, & **Niu, C.** Conceptualizing effective foster parent mentor programs: A participatory planning process. *Child and Family Services Review*, 73, 411-418. 2017
- Zhang, J., **Niu, C.**, Shahbaz, M., Anderson, R., & Nguyen-Jahiel, K. What Makes a More Proficient Discussion Group in English language Learners' Classrooms? Influence of Teacher Talk and Student Backgrounds. *Research in the Teaching of English*, 51(2), 183-208. 2016
- Zhang, J., Cabrera, J., **Niu, C.**, Zippay, C., & Dietrich, S. (Under Review). Pre-service Teachers' Perceptions of Clinically Based and Non-Clinically Based Teacher Preparation Programs. *Teacher Development* 2015
- Zhang, J., Li, H., **Niu, C.**, Chen, Y., Dong, Q., & Xu, J. (Under Review). The Role of Orthography in Oral Vocabulary Learning in Chinese as a Second Language 2014

- Niu, C.**, Chon, K., Zhang, J., Miller, S., Houchens, G., & Norman, A. (Under Review). The TELL Kentucky Survey of Teacher Working Conditions: Do Differences in Teacher and Principal Perceptions Relate to Student Outcomes? 2013
- Niu, C.**, Zhang, J., Chon, K., Norman, A., & Miller, S. (Under Review). Teacher perceptions of teaching conditions and student achievement in regional high schools in Kentucky. 2013
- Tao, X., **Niu, C.** (Ed.) (in press). *English Debating*. Beijing: Foreign Language Teaching and Research Press. 2013
- Tao, X., **Niu, C.**, & Zhang, P. (Ed.) (2011). *English Public Speaking*. Beijing: Peking University Press. 2011
- Beck, M., translated by **Niu, C.**, & Zhang, P. (2010). *Finding Your Own North Star*. Chongqing: Chongqing University Press. 2010
- Li, F., Feng, X., Zhang, P., & **Niu, C.** (Ed.) (2009). *A New Course Book of Practical Interpreting: Theory, Skills and Practice*. Chengdu: Sichuan People's Press. 2009
- Niu, C.** (2007). Exploring the “effortless perfection”: Discerning the “intention” in poetry translation from the hermeneutics perspective. *Journal of Chongqing University of Posts and Telecommunications*, 19(z1), 76-80. (In Chinese) 2007

AWARDS AND SCHOLARSHIPS

- The United Way of Southern Kentucky (UWSK) Shooting Star Award for persons who exemplified courage and tenacity in overcoming obstacles 2015
- The Outstanding Educational Leadership Doctoral Student in the Organizational Leadership Specialization 2014
- Sichuan International Studies University 2009 Award for Teaching Excellence 2009
- Sichuan International Studies University 2007 Award for Outstanding Young Scholars 2007
- Han Suyin's National Award for Young Translators 2004

The First-Class Scholarship for Academic Excellence, Guangdong Foreign 1998-2000
Studies University, China

Scholarship for Outstanding College Freshmen, Anhui Provincial Bureau of 1996-2000
Education, China