2010-12-17

The Effect of Raters and Rating Conditions on the Reliability of the Missionary Teaching Assessment

Abigail Christine Ure
Brigham Young University - Provo

The Effect of Raters and Rating Conditions on the Reliability of the Missionary Teaching Assessment

Abigail C. Ure

A dissertation submitted to the faculty of
Brigham Young University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Richard R. Sudweeks, Chair
Randall Davies
Tim Morrison
Peter Rich
Stephen Yanchar

Department of Instructional Psychology and Technology

Brigham Young University

April 2011

ABSTRACT

The Effect of Raters and Rating Conditions on the Reliability of the Missionary Teaching Assessment

Abigail Christine Ure

Department of Instructional Psychology and Technology

Doctor of Philosophy

This study investigated how 2 different rating conditions, the controlled rating condition (CRC) and the uncontrolled rating condition (URC), effected rater behavior and the reliability of a performance assessment (PA) known as the Missionary Teaching Assessment (MTA). The CRC gives raters the capability to manipulate (pause, rewind, fast-forward) video recordings of an examinee's performance as they rate while the URC does not give them this capability (i.e., the rater must watch the recording straight through without making any manipulations). Few studies have compared the effect of these two rating conditions on ratings. Ryan et al. (1995) analyzed the impact of the CRC and URC on the accuracy of ratings, but few, if any, have analyzed its impact on reliability.

The Missionary Teaching Assessment is a performance assessment used to assess the teaching abilities of missionaries for the Church of Jesus Christ of Latter-day Saints at the Missionary Training Center. In this study, 32 missionaries taught a 10-minute lesson that was recorded and later rated by trained raters based on a rubric containing 5 criteria. Each teaching sample was rated by 4 of 6 raters. Two of the 4 ratings were rated using the CRC and 2 using the URC.

Camtasia Studio (2010), a screen capture software, was used to record when raters used any type of manipulation. The recordings were used to analyze if raters manipulated the recordings and if so, when and how frequently. Raters also performed think-alouds following a random sample of the ratings that were performed using the CRC. These data revealed that when raters had access to the CRC they took advantage of it the majority of the time, but they differed in how frequently they manipulated the recordings. The CRC did not add an exorbitant amount of time to the rating process.

The reliability of the ratings was analyzed using both generalizability theory (G theory) and many-facets Rasch measurement (MFRM). Results indicated that, in general, the reliability of the ratings obtained from the 2 rating conditions were not statistically significantly different from each other. The implications of these findings are addressed.

Keywords: generalizability theory, many-facet Rasch measurement, performance assessment, microteaching, reliability, rater behavior, rater cognition, video recording, assessment in teacher education

Table of Contents

List of Tables ix

List of Figures xi

List of Equations xiii

Chapter 1: Introduction 1

Statement of Purpose 5

Rationale 6

Research Questions 6

Background 7

Chapter 2: Literature Review 10

Performance Assessment 10

Performance assessment in teacher education 12

Use of microteaching for assessing prospective teachers 13

Rater Cognition 17

Use of Video in Rater Observations 20

Estimating Reliability 23

Generalizability theory 24

Facets 25

Relative versus absolute decisions 28

G-study summary statistics 28

D-study 29

Many-facet Rasch measurement 29

G theory versus MFRM ... 31

Chapter 3: Method ... 34

Study Participants ... 34

Missionaries ... 34

Raters ... 34

Investigators ... 35

Design ... 35

Instrumentation ... 37

Assessment procedure ... 37

Rating scales ... 38

Procedure ... 41

Analyses Used to Address the Research Questions ... 44

Research Question 1: Rater behavior ... 44

Research Question 1a: Raters' usage of controlled rating condition ... 44

Research Question 1b: Raters' reasons for reviewing recordings ... 44

Research Question 1c: Time spent reviewing recordings ... 44

Research Question 1d: Changes in reviewing behavior across criteria ... 45

Research Question 2: Variability attributable to each source of variance ... 45

Research Question 2a: Reliability across criteria ... 49

Research Question 2b: Impact of number of raters on reliability ... 52

Research Question 2c: Reliability across rating conditions ... 52

Research Question 2d: Impact of the use of the CRC on reliability ... 53

Research Question 3: Performance of rating scale categories ... 54

Chapter 4: Results 55

Research Question 1a: Raters’ Usage of the Controlled Rating Condition 55

Use of various types of manipulation 57

Pausing 57

Rewinding 57

Fast-forwarding 59

A different way of aggregating the video manipulation data 59

Rewind′ 61

Fast-forward′ 64

Summary 64

Research Question 1b: Raters’ Reasons for Reviewing Recordings 64

Research Question 1c: Time Spent Reviewing Recordings 67

Research Question 1d: Changes in Reviewing Behavior Across Criteria 69

Research Question 2: Variability Attributable to Each Source of Variance 71

Generalizability study findings 71

Variance component for the missionary facet 72

Variance component for the rater facet 72

Variance component for the residual 74

Many-facet Rasch measurement findings 74

Missionary facet 76

Rater facet 78

Research Question 2a: Reliability Across Criteria 80

Generalizability study findings 80

Many-facet Rasch measurement findings 80

Missionaries 81

Raters 83

Research Question 2b: Impact of Varying Number of Raters on Reliability 83

Research Question 2c: Reliability Across Rating Conditions 84

Generalizability study findings. 84

Many-facet Rasch measurement findings 85

Research Question 2d: Impact of the Use of the CRC on Reliability 87

Research Question 3: Performance of Rating Scale Categories 92

Chapter 5: Discussion 102

Reflections on Findings for Each Research Question 102

Research Question 1a: Raters' usage of controlled rating condition 102

Research Question 1b: Raters' reasons for reviewing recordings 104

Research Question 1c: Time spent reviewing recordings 105

Research Question 1d: Changes in reviewing behavior across criteria 106

Research Question 2: Variability attributable to each source of variance 107

Research Question 2a: Reliability across criteria 108

Research Question 2b: Impact of number of raters on reliability 109

Research Question 2c: Reliability across rating conditions 110

Research Question 2d: Impact of the use of the CRC on reliability 111

Research Question 3: Performance of rating scale categories 112

Comparison of G Theory and MFRM 112

Implications for Microteaching 113

Limitations 114

Recommendations for Future Research 114

References 117

Appendix A: Missionary Teaching Assessment Rating Scale 124

Appendix B: Informed Consent Form for Raters 127

Appendix C: Protocol for Rater Think-Alouds 130

Appendix D: Exit Interview Questions 132

List of Tables

1. Study Design 36
2. Raters' Usage of Video Manipulations 56
3. Raters' Usage of Video Manipulation 62
4. Reasons Raters Reported for Reviewing Recordings 66
5. Additional Rating Time Added by Raters due to Video Manipulation 68
6. Criteria Associated With Each Rewind and Reason 70
7. Amount of Variability in the Ratings Attributed to Each Source 73
8. MFRM Analysis of Missionaries 77
9. MFRM Analysis of Raters 79
10. Phi Coefficients by Criteria and Rating Condition for Two Raters 80
11. MFRM Analysis of Criteria 81
12. MFRM Missionary Group-Level Statistics for Five Criteria 82
13. MFRM Summary Rater Statistics for Five Criteria 83
14. Phi Coefficients for Two Rating Conditions Across Five Criteria 85
15. MFRM Analysis of Rating Conditions 86
16. MFRM Missionary Reliability Estimates for Rating Conditions by Criteria 87
17. MFRM Missionary Group-Level Statistics by Rater 88
18. MFRM Missionary Group-Level Statistics by Rating Condition and Rater 90
19. Statistics by Rater from Feldt Test Comparing Rating Conditions 91
20. MFRM Rating Scale Category Statistics for Shows Warmth and Concern 93
21. MFRM Rating Scale Category Statistics for Listens 94

22. MFRM Rating Scale Category Statistics for Asks Questions 96

23. MFRM Rating Scale Category Statistics for Adjusts to Needs 98

24. MFRM Rating Scale Category Statistics for Invites Others to Make Commitments.........100

List of Figures

1. Description of teaching situation presented to missionaries. 39
2. Description of teaching situation presented to investigators. 39
3. Division of four subsets for each criterion and rating condition combination. 48
4. Percentage of recordings where pause, rewind, or fast-forward function used one or more times. 58
5. Frequency of rewind and rewind′. 63
6. Average use of rewind and rewind′ per recording. 63
7. Frequency of fast-forward and fast-forward′. 65
8. Average use of fast-forward and fast-forward′ per recording. 65
9. Facets map displaying calibrations of missionaries, raters, rating conditions, and criteria. 75
10. Projected reliability of absolute decisions obtained by varying the number of raters rating each examinee 84
11. Category probability curves for Shows Warmth and Concern 93
12. Category probability curves for Listens. 95
13. Category probability curves for Asks Questions with original scale. 97
14. Category probability curves for Asks Questions with collapsed scale. 97
15. Category probability curves for Adjusts to Needs. 99
16. Category probability curves for Invites Others to Make Commitments. 101
17. Category probability curves for Invites Others to Make Commitments for collapsed scale. 101

Appendix A1: Missionary Teaching Assessment rating scale for five criteria included in the study..125

Appendix A2: Missionary Teaching Assessment rating scale for four criteria not included in the study. 126

List of Equations

1. Weighted Mean for Subdividing Method 46

2. Feldt Test 50

Chapter 1: Introduction

Unlike multiple-choice tests, performance assessments (PA) require examinees to construct responses to complex tasks that are similar to realistic problems (Braden, 2005). PA tasks include presentations, research projects, role-plays, experiments, portfolios, working through case studies, etc. (Zenisky, 2007). PAs can also be used to assess teaching ability. They allow supervisors to observe the teaching process in a naturalistic setting and provide more detailed and precise evidence than other data sources (Waxman, 2003). Although PAs have a number of benefits due to the richness of the data they provide, they also have some problematic methodological and feasibility issues that should be considered before using them.

Performance assessment is not a true assessment unless the quality of the examinee's performance has been evaluated. Typically, an examinee's performance on a PA must be observed and judged by a rater. Consequently, one of the primary methodological issues associated with PAs is the reliability of these rater-mediated judgments. In general, reliability refers to the consistency of the scores obtained from a measurement procedure. In the context of PA, it refers to the consistency of the ratings. Two or more ratings of each examinee's performance are necessary in order to estimate consistency, but at least two kinds of inconsistency (i.e., a lack of reliability) are possible. If the multiple ratings necessary to estimate consistency are obtained by having two or more raters rate the same performance on a single rating occasion, then the degree of consistency is defined in terms of interrater reliability. If the multiple ratings were obtained by having a single rater rate each examinee's performance on two or more rating occasions then the degree of consistency is defined in terms of intrarater reliability. A third alternative involves having each examinee's performance rated by multiple raters on multiple rating occasions. This last alternative is more informative because it permits estimates of both interrater and intrarater reliability.

Studies of the reliability of ratings typically include multiple sources of inconsistency such as differences in raters, test occasions, and tasks. Popham (1990) pointed out three common sources of error that come from the rating process: (a) rating-instrument flaws, (b) procedural flaws, and (c) rater personal-bias errors. Rating instrument flaws are introduced when rating scales are vaguely defined which likely leads to inconsistent interpretations by raters. The rating scale categories may be poorly defined or have overlapping descriptors. Procedural flaws occur when there are problems in the rating operation. For instance, a rater may be overwhelmed with too many traits to rate at the same time or a rater may be asked to rate for an extended period of time leading to fatigue.

Many PA studies have focused on aspects of the rating process such as development, training, and rater personal-bias errors. According to Joe (2009), "relatively little emphasis has been placed on the degree to which aspects of performance assessments (e.g., scoring rubrics and procedures) adversely influence the rater, from a cognitive (decision-making) or physiological perspective (e.g. fatigue)" (p. 18).

Although rating errors can never be fully eliminated, they can and should be reduced as much as possible. Many studies have sought to identify ways to decrease the inevitable error that comes with human raters. This issue has been approached from a variety of vantage points. One approach has been to understand the cognitive processes a rater uses when making a judgment (DeCarlo, 2005; Joe, 2009; Lumley, 2002; Orr, 2002;

Suto & Greatorex, 2008; Wolfe & Feltovich, 1994). The hope is that by understanding these cognitive processes, the error that comes from the decision-making process raters go through can be minimized through means such as rater training. A key element in the decision-making process is the observation phase. Does one rater observe the same stimuli that another rater observes? How do differences in observation or perception affect the ratings given by a rater?

Rating procedures such as how a rater observes a ratee have changed with advances in technology. Before audio/visual recordings were commonplace, all observations had to be made in person as the event occurred. Audio/visual recordings have introduced much more flexibility into the rating process. This technology has increased the ease with which raters can rate PAs. A performance can be recorded in a remote location and rated by a trained rater at a later time and in a different location. If the task the subject is performing is complex, a rater can pause or review the recording to better analyze it. A rater can also review the recording if they lose focus. All of these advantages are likely to have an effect on what a rater observes and ultimately on the ratings.

Many performance observations are still observed and judged live. One study (Ryan et al., 1995) compared the accuracy of ratings from raters who observed a performance "directly" (in person as it occurred) and those who observed a performance "indirectly" (via a video recording). Ryan et al. defined accuracy as how close raters' ratings were to ratings awarded by expert raters. He found no significant difference between the accuracy of the two groups. No study has compared the reliability of the ratings from these two rating procedures.

One way to minimize the influence of differences between raters is to require each rater to rate the performance of each examinee and then compute the mean ratings for each examinee averaged across raters. Similarly, one could compute the mean rating for each examinee averaged across rating occasions or across facets. The reliability of mean ratings generally increases as more occasions, tasks, and raters are added to the assessment. It also increases through proper rater training and appropriate rating procedures. Unfortunately, an increase in reliability often comes with a tradeoff in feasibility. Each of the additional elements that can contribute to increased reliability also bring additional costs such as time taken away from students to administer the assessment, the cost of paying additional raters, the cost of in-depth rater training sessions, and the cost of technology. These added costs may make it unfeasible for many organizations to conduct these more extensive assessments. Test developers must find a good balance between an acceptable level of reliability and cost.

The Missionary Training Center (MTC) in Provo, Utah is an institution that regularly administers PAs to young men and women who are training to serve as missionaries. The MTC is sponsored by The Church of Jesus Christ of Latter-day Saints (LDS Church) and trains thousands of missionaries every year how to effectively teach basic doctrinal principles. A PA, known as the Missionary Teaching Assessment (MTA), was developed by the MTC Research and Evaluation department. The MTA assesses nine different criteria or teaching skills but only five were included in the study in order to simplify the rating process and not overwhelm the raters with too many criteria.   The criteria included in this study were (a) Shows Warmth and Concern, (b) Listens, (c) Adjusts to Needs, (d) Asks Questions, and (e) Invites Others to Make Commitments. The full MTA rubric is displayed in Appendix A. The purpose of the MTA is to gather systematic observations to assess and track missionaries' current teaching ability. Describing the current status of instructional practices and identifying instructional problems is one of the key purposes of teacher observations (Waxman, 2003). In order for the Research and Evaluation department to effectively measure missionary teaching performance, the MTA must be both reliable and feasible.

Statement of Purpose

This study had two purposes. The first purpose was to explore how MTA raters use digital recordings when they are given the capacity to control the pace at which they view them (e.g., pause and rewind) and how this capability affected the reliability of the ratings. Currently, teachers assess missionaries while they teach. In the Teaching Resource Center they watch a missionary's performance from a TV monitor in another room. Because the performance is live, they do not have the ability to pause or rewind. In this respect, the observation method they are employing is similar to observing in-person. The researcher was interested in exploring how raters would view the performance when rating if it was not live and they had the capability to control the pace. Throughout this study, the researcher refers to this capability or the lack of it as the rating condition. The rating condition that gives the rater the ability to pause or rewind is referred to as the controlled rating condition (CRC) and the rating condition that does not give the rater this ability is referred to as the uncontrolled rating condition (URC).

The second purpose of this study was to make recommendations for improving the reliability of MTA ratings while keeping costs at feasible levels. As previously discussed, reliability is one of the central problems of PAs. Assessing reliability and making necessary changes to increase it are essential to ensuring the psychometric soundness of any instrument. Increasing the reliability of the MTA will give the MTC the ability to draw generalizable conclusions from it in the future.

Rationale

There were both practical and theoretical purposes for conducting this study. From a practical perspective, the data gathered on rating conditions should be beneficial to the MTC as well as other institutions implementing PAs. Many studies on the reliability of PAs have focused on reliability across raters and tasks. Yet, no studies have analyzed the effects of rating condition on reliability. The researcher was interested in knowing how the CRC affected reliability. If using the CRC increased the reliability of the ratings, then it was an option that should be considered when rating PAs. If the CRC did not contribute or contributed very little to reliability, then the added time and cost may not be justified.

Another practical outcome of this study was the improvement of the MTA. Analyzing the reliability of the MTA and making suggestions concerning how to increase the reliability in a cost effective way will benefit the MTC's ability to systematically measure missionary teaching performance.

Results from this study will also contribute to the literature on rater cognition. A better understanding of if and how raters utilize the capability to control the pace of digital recordings can provide data on rater cognition during the observation phase. Also, how raters' use of the capability changes from one criteria to another gives insight into whether or not certain criteria are more complex to assess thus leading to cognitive overload.

Research Questions

This study focused on the following specific questions:

1. When raters were able to control the pace (e.g., pause, rewind, and fast-forward) in which they viewed digital recordings of missionary trainees' teaching performance, to what extent and for what reasons did they use this capability?
    a. How often did the raters manipulate the recordings?
    b. Why did raters manipulate the recordings?
    c. How much time did raters spend engaged in reviewing segments?
    d. How did the raters' reviewing behavior vary from one rating criterion to another?
2. What percent of the variability in missionaries' ratings was due to estimated differences in the missionaries' teaching ability and what percent was due to inconsistencies between (a) the raters, (b) the rating conditions, and (c) the various possible interactions between these sources of variability?
    a. How did the reliability of the ratings vary from one rating criterion to another?
    b. How did the reliability vary as a function of the number of raters?
    c. How was the reliability influenced by the rating condition used?
    d. How did a rater's use of the controlled rating condition affect the reliability of their ratings?
3. How well did the categories in each of the rating scales function and which categories, if any, need to be combined or revised?

Background

Male missionaries are allowed to commence their two-year mission work when they turn 19-years old. The Church strongly recommends that every worthy male member of the Church serve a two-year mission. On the other hand, females may serve an 18-month mission when they are 21-years old if they would like but there is less of an expectation for them to do so. Because of this, the MTC Research and Evaluation department estimates that only 15% of missionaries are female. After potential missionaries submit an application, they are assigned to an area of service in the world by LDS Church leaders. For missionaries who are assigned to missions where a language other than their native language is spoken, they are expected to learn the language of the area. Missionaries are sent to 1 of the 17 MTCs located around the world. The largest MTC is located in Provo, UT. Approximately 20,000 missionaries are trained at the Provo MTC every year. Missionaries are taught how to teach others the doctrinal principles of the LDS Church as well as the foreign language they will be speaking (if applicable).

In order to aid the missionaries in learning how to teach, the MTC has used some form of a teaching performance assessment for the last 25 years in a number of formal and informal situations. Most of the PAs at the MTC are used for formative or instructive purposes. Missionaries are currently observed and receive formative feedback on an almost daily basis. These observations take place in the classroom where missionaries role-play teaching experiences with other missionaries or with their teachers.

Missionaries are also observed and given feedback in the Teaching Resource Center (TRC), the Referral Center (RC), and the Teaching Evaluation Center (TEC). In the TRC, missionaries teach volunteers who play the role of an investigator of the LDS Church. Teachers observe their performance via a TV monitor that streams the live teaching performance and provide written and oral feedback. In the RC, missionaries talk to individuals who call in for free products such as a copy of the Bible, the Book of Mormon, or a Church-produced video about Christ. Missionaries talk with these people, share their testimonies, and persuade them to have local missionaries deliver the selected product and share a message with them. Teachers can listen in on the RC phone calls and give the missionaries immediate feedback on their performance. In the TEC, missionaries receive a more formal assessment of their teaching. Missionaries teach employees who play the part of an investigator. These employees have received more advanced training in missionary teaching evaluation. Immediately following the missionaries' performance, these employees provide them with immediate feedback on their teaching and give them another opportunity to teach and implement the feedback.

PAs and ratings are a central part of evaluating missionaries' teaching while in the MTC, but the assessments tend to be formative and less formal. The MTC has not taken a systematic approach to evaluating the teaching performance of missionaries across the MTC to assess how it is doing as a whole. This is the purpose of the Missionary Teaching Assessment (MTA). The MTA is a procedure that includes having missionaries teach a person acting as investigator with some task specification. The performance is then rated by a rater based on a rubric. This instrument is the focus of this study. Because it is summative in nature and will be used to make decisions about the teaching ability of missionaries and the quality of the curriculum, it is vital that the MTA be both reliable and feasible.

Chapter 2: Literature Review

In this review, the researcher defines and discusses the history of performance assessments including its use in the field of education. She then reviews two topics that are pertinent to this study concerning the rating of PAs: (a) rater cognition and (b) the use of video in rater observations. Finally, the researcher reviews generalizability theory and many-facet Rasch measurement, the two statistical models that were used to analyze the reliability and generalizability of the MTA.

Performance Assessment

Performance assessment (PA) is a broad term that has a variety of meanings (Palm, 2008). PAs differ from the typical multiple-choice assessment in that a student must construct their response as opposed to just selecting it from a group of options. Response construction is a necessary but not sufficient quality of PAs. Arter (1999) states that PAs do not "include all constructed-response-type items (especially short answer and fill in the blank), but, admittedly, the line between constructed response and performance assessment is thin." Stiggins et al. (2003) offer the following definition: "The term performance assessment (PA) is typically used to refer to a class of assessments that is based on observation and judgment. That is, in PAs an assessor usually observes a performance or the product of a performance and judges its quality" (p. 134). As this definition stipulates, a PA can evaluate either a performance such as a musical performance or presentation or an end product such as an essay or a culinary creation.

A movement toward the use of PAs in the educational system began in the mid-1980s and has continued to this day (Stiggins et al., 2003). This growth in popularity came as a result of dissatisfaction with selected-response assessments that focused on memorization. This dissatisfaction was the result of a shift in educational paradigms.

> For centuries, the prevailing assumption about learning has been that the teacher tells, shows, or demonstrates facts, knowledge, rules of action, and principles, and then students practice them. . . . By the mid-1980s another model of the mind and pedagogy emerged, locking horns with the heretofore governing model. This new paradigm is rooted in the belief that there is "construction of knowing in a socio-cultural context" that embodies "investigatorial styles of learning." It is this model of learning that has driven the survey of the "new" or "authentic" assessment movement. (Madaus & O'Dwyer, 1999, p. 689)

Also during the 1980s, many stakeholders in education began asking what skills our future workforce needed. With knowledge doubling every 3 years, students needed to know how to do more than just memorize (Stiggins et al.). All of these changes in thinking led to the major shift toward PAs in schools.

Like all assessments PAs have their strengths and weaknesses that need to be taken into account before using them. Many advocates believe that they have increased validity because they are able to elicit higher-order thinking skills and they preserve the complexities that are a part of real life situations (Ryan, 2006). PAs allow educators to directly observe and make judgments about a competency or proficiency and assess a broad range of learning outcomes (Stiggins et al., 2003).

The increase in validity comes with tradeoffs including feasibility and technical issues. PAs are often unfeasible for schools to employ because they are more expensive to develop, administer, and score. The primary technical issue is the reliability of tasks and scoring. Is one task comparable to any other task a student may receive? Is one rater as severe in their rating as another? Do raters interpret the scoring criteria differently from one another? Other criticisms of PAs include poor quality tasks, incorrect or poorly defined performance criteria, and an inappropriate sample of tasks (Stiggins, 1994).

Performance assessment in teacher education. Teacher testing has been a part of the educational system since the early part of the 20th century (Cruickshank & Metcalf, 1993). Some of the methods include tests of subject matter knowledge, peer reviews, classroom observation, student evaluations, students' achievement test scores, teacher performance tests, and teacher self-evaluations (Haertel, 1988). Traditionally teacher candidates have been assessed using paper-and-pencil standardized tests. Alternative forms of teacher assessment such as PAs and portfolios began appearing more and more in the 1980s. Large-scale alternative assessments for teachers include the Teacher Assessment Project (TAP) and the National Board for Professional Teaching Standards (NBPTS).

Alternative assessments began gaining favor in the 1980s because of the many shortcomings of the objective paper-and-pencil forms of assessment such as the Pre-Professional Skills Test (PPST) and the National Teacher Examination (NTE). These assessments measure teacher candidates' general knowledge, subject matter content knowledge, and pedagogical knowledge. One of the shortcomings is the fact that paper-and-pencil tests fail to be predictive of future teacher performance (Haertel, 1988). What teacher candidates know cognitively about their subjects or about teaching does not necessarily transfer to how they perform in a classroom. Other shortcomings are that they often focus on the recall of subject matter content, generic and not subject-matter specific pedagogy is assessed, and critical teaching skills are not measured (Haertel, 1991).

PAs such as microteaching have shown to have better than average predictive validity (Cruickshank & Metcalf, 1993). PAs allow supervisors to directly assess teaching performance abilities (Pecheone & Chung, 2006). Not only are these PAs valuable in measuring teaching ability, but they also provide valuable opportunities for student teacher learning and growth and they promote systematic change in schools (Delandshere & Petrosky, 1998).

Use of microteaching for assessing prospective teachers. Microteaching is a type of teaching PA that allows preservice or inservice teachers to practice particular teaching skills in a more controlled setting. In microteaching, a teacher teaches a brief 5 to 20 minute single concept lesson to a small group of pupils, generally three to five, who are usually volunteers or peers. The teacher focuses on one teaching skill such as introducing a lesson, teaching with clarity, responding to silence and nonverbal cues, and using visual aids. This abbreviated lesson allows the teacher candidate to practice a particular teaching skill in a low risk environment (Shore, 1972). The teaching experience is followed by a critique from a supervisor, teacher, or even a peer. Feedback is generally followed by an opportunity for the teacher to reteach the lesson with applicable improvements.

The microteaching method was developed in 1963 at Stanford University. It grew in popularity and quickly spread to more than half of the teacher education programs in the U.S. Eventually, programs became overwhelmed by its complexity and many ceased to use the method. According to Allen and Wang (2002), microteaching began to reemerge in the late 1980s and 1990s as many programs began adopting a more scaled down model. Often, microteaching occurs without a supervisor. Feedback is provided by peers, thus simplifying the process and lowering costs.

Microteaching does not take the place of student teaching or internships. It is generally used to give preservice students an opportunity to have some teaching experience and develop specific teaching skills before they enter a real classroom (Trent-Wilson, 1990). This method also helps to bridge the gap between theory and practical application (Brent, 1996).

Although microteaching experiences can last anywhere from 5 to 20 minutes, Allen and Ryan (1969) found that a 4 minute teaching experience was as effective as a 7 minute teaching experience. Research at Stanford confirms the usefulness of shorter microteaching lessons. Five minutes is often sufficiently long to practice many teaching skills. When lessons are much longer, training sessions become increasingly complex and tend to lose focus (Allen & Wang, 2002).

When resources permit, microteaching experiences are video recorded and later reviewed by the student teacher and their supervisor. The student teacher can review his/her teaching performance with the supervisor and/or peers providing both positive and negative feedback. The student teacher is given another opportunity to teach and improve the applicable teaching skills (Allen & Wang, 2002). Video recording teaching performances and reviewing them has proven to be more effective in improving teaching than not recording them (Kpanja, 2001).

Microteaching is used to provide formative feedback and has been shown to be effective in improving teaching skills. Kallenbach and Gall (1969) conducted a study comparing teacher education students trained with a microteaching approach and students who had a conventional classroom observation and student teaching approach. They found no significant difference between the teacher effectiveness ratings of the two groups. The significant finding was that microteaching was able to deliver comparable results in only one fifth of the time and with fewer administrative problems.

Allen and Fortune (as cited in McKnight, 1971) conducted a similar study where they compared two groups of preservice teachers. One group participated in 10 hours of microteaching while the other group spent 25 hours observing a classroom and functioning as a teacher's aide. The students who participated in microteaching received slightly higher teacher effectiveness ratings than their peers in the observation program. The majority of the students in the microteaching program (89%) believed the experience had been valuable for them. Additionally, Allen and Fortune reported the finding that microteaching situations were valid predictors of subsequent classroom performance.

Not only did the benefits of microteaching manifest themselves in teacher effectiveness ratings, but students also indicated in surveys that they believed the method was valuable. Benton-Kupper (2001) conducted a survey of students in a general secondary methods course following their microteaching experience to assess their perspective of the method. Students indicated that they had very positive feelings about microteaching and that it increased their confidence as a teacher. Microteaching instilled within them the value of reflecting on their teaching and they believed that the use of videotapes was conducive to feedback and reflection. Additionally, students appreciated being able to observe their peers teaching because it gave them new ideas and strategies for teaching.

Another study (Bolton, 1996) examined how using microteaching to assess students impacted student teachers' self-efficacy. Bolton compared two groups. One group was assessed using a traditional objective nonperformance-based exam while the other was assessed through microteaching. Bolton found that the students in the microteaching group had greater self-efficacy at a statistically significant level in the following four areas: (a) writing objectives, (b) developing task analyses, (c) developing lesson plans, and (d) teaching a lesson.

In microteaching students focus on only one skill at a time. The lesson they teach is simplified, and they teach it in a very small and controlled environment. One of the greatest drawbacks of microteaching is that studies have not indicated that the skills they learn in such a simplified environment are transferred to the complex atmosphere of an actual classroom. Peterson (1973) compared how well two groups of student teachers implemented 13 specific questioning skills into their classroom discussions during a field experience. One group practiced these questioning skills prior to their field experience using microteaching while the other group did not have this opportunity. A comparison between the two groups showed no significant differences between their regular classroom discussions. Peterson recommends that more needs to be done to aid in the transfer of the skills taught in microteaching environments.

Other studies have found similar results. Rose and Church (1998) performed a literature review of studies on various methods of training preservice and inservice teachers that used direct observation to assess the impact of the training on their teaching behaviors and skills. Rose and Church reported that the microteaching studies showed weak and inconsistent training effects especially compared to other training procedures.

They reported that the microteaching procedures provide practice that is too far removed from the actual classroom.

Other factors have impeded a more widespread use of microteaching. Microteaching and other PAs are more costly to develop and implement than standardized tests. A preoccupation exists with field-based experiences that take the place of on-campus laboratory experiences like microteaching. Finally, there is a lack of agreement concerning what constitutes desirable professional practice (Cruickshank & Metcalf, 1993).

In most instances microteaching is used as a means of helping teacher candidates practice and improve their teaching skills. They are used as a formative assessment and less as a summative assessment. Researchers have been interested in assessing their ability to function as a teaching tool or their validity as an assessment tool. Few studies have considered the reliability of their ratings or the feedback that comes from supervisors or peers.

The MTA has many similarities to microteaching. It allows "preservice" missionaries to practice particular teaching skills in a controlled setting. The missionaries teach a brief, single concept lesson. The lessons are video recorded and later reviewed. One way that it differs from microteaching is that it is summative in nature and the missionary does not receive either feedback concerning his/her performance or an opportunity to reteach the lesson. Unlike most of the microteaching studies, this study assessed the reliability of the MTA.

Rater Cognition

Prior to 1980 performance evaluation literature focused on improving the instruments used in the evaluations. In the early 1980s, researchers such as Feldman

(1981) began looking more closely at the way raters gathered information and formulated judgments. In order to better understand and improve the rating process, researchers over the past 3 decades have sought to create a comprehensive model of the cognitive processes raters experience during the rating process (Arvey & Murphy, 1998). Most of the performance evaluation research has come from the human resources sector and is centered on the rating of subordinates by their supervisors.

Although each model of rater cognition has its own unique attributes, they all contain the same general phases: (a) observation of behavior, (b) encoding, (c) storage, (d) retrieval, and (e) integration of information. The first phase, the observation of ratee behavior, is the most pertinent to this study and will therefore be the main focus of this literature review.

DeNisi (1996) noted that this first step in the appraisal process was critical because the accuracy of an evaluation is dependent on the information available to the rater:

> The decision making process in performance appraisal begins with raters acquiring information, and the outcome of this process (i.e., the performance information available to the rater) will determine the evaluation made. Since we assume that raters cannot observe all aspects of the performance of each ratee because of conflicting demands on their attention, or simply because of physical constraints, raters will make decisions based only upon samples of the ratee's performance. Even if two raters are observing the same ratee then, if they engage in different information search or acquisition activities, they will have different information available to them, and so will likely arrive at different evaluations. (p. 31)

A study by Sanchez and De La Torre (1996) confirmed the fact that the accuracy of dimensional ratings is a function of the accuracy of memories.

Kolk, Born, van der Flier, and Olman (2002) conducted a study on cognitive load during the observation phase. Kolk et al. stated that taking notes during the observation phase facilitated verbal encoding of behavior but the dual task of observing and note-taking could lead to cognitive overload. As a result, the rater may make observational and rating errors. They may miss key behaviors while writing down an observation or they could incorrectly classify behaviors. Kolk and his colleagues also hypothesized that experienced raters perform better under the cognitive demands during the observation phase than less experienced raters. Past studies support that experience and practice with a task leads to a decrease in the cognitive resources needed.

Kolk et al. (2002) used a group of experienced raters and a group of inexperienced raters in their study. Half of the raters from each group were instructed to take notes while observing a performance while the other half was instructed to withhold taking any notes until after the observation. They found that more experienced raters did produce significantly more differentially accurate ratings than inexperienced raters. Differential accuracy in this context pertains to how favorably/unfavorably inexperienced raters rated each candidate compared to how favorably/unfavorably expert raters rated each candidate on a dimensional level as opposed to an overall level. The group of raters who postponed taking notes until the end had a slightly higher interrater reliability (.93) than the group who took notes while observing (.85) although this difference was not statistically significant. Kolk et al. acknowledged that the lack of significance may have been due to the small sample size of videotaped candidates that were assessed (*n* = 3).

When rating the MTA, raters' use of the two different rating conditions could have potentially caused them to gather different samples of the ratees' performances. It would seem logical that a rater would be able to observe a larger and a more complete sample of a missionary's behavior if they rewound and reviewed something that they missed the first time through. On the other hand, a rater using the CRC may opt to focus on a narrower subset of behaviors knowing that they would have the opportunity to go back later and review the video for other behaviors. Ryan et al. (1995) stated, "Videotaping may lead to less attention overall, as the need to be vigilant in observation is less when one knows there is the capacity to replay and catch what is missed" (p. 665). If one rater used the controlled rating condition and another one did not or if one rater manipulated the ratings much more frequently than another, then the two raters would be "engaging in different information search or acquisition activities" (DeNisi, 1996). If DeNisi's hypothesis held true for this study, then the raters should have "likely [arrived] at different evaluations" (DeNisi). If differing search activities lead to different evaluations or results both between raters and within raters, then both the interrater and intrarater reliabilities could potentially be impacted.

Use of Video in Rater Observations

The introduction of video and now digital recordings has greatly enhanced the field of PAs. Raters are now able to evaluate ratees remotely. Ryan et al. (1995) termed this type of remote observation as indirect observation while in-person observation was termed direct observation. Indirect observation has the advantages of reducing assessor fatigue, increasing the number of raters that can observe, enhancing the credibility of an assessment, lowering costs, and providing the ability to review videotapes when there are disagreements among raters.

Indirect observation can occur in two different ways. The rater can watch the video and pause or rewind it when necessary or he can just watch it without any type of manipulation. Ryan et al. (1995) refers to the former as controlled observation and the latter as just indirect observation. In this study, the researcher refers to controlled observation as the controlled rating condition (CRC) and indirect observation as the uncontrolled rating condition (URC).

Watching a performance directly or indirectly without any control over the recording could potentially put pressure on raters since they are not able to pause or review what is going on. According to DeNisi (1996), raters facing time pressures consider fewer pieces of information and are more likely to search for negative information. They are also more likely to rely upon the results of past evaluations. Raters give the greatest weight to the pieces of information most easily retrievable from memory.

Little research has been conducted comparing these different methods of observation. Ryan et al. (1995) reviewed two studies that compare direct and indirect observation (without control). One study showed affective differences between the two groups (how they felt about the evaluation process) but no significant differences between the ratings. The second study showed that raters observing indirectly awarded significantly higher ratings on 5 of 12 dimensions and on the overall rating. However, Ryan et al. stated that both of these studies were inconclusive due to their small sample sizes.

Ryan et al. (1995) conducted a study to determine if indirect observation affected the accuracy of ratings when compared to direct observation. They observed two different types of accuracy: (a) behavioral-accuracy (e.g., recognition of specific behaviors) and (b) classification-accuracy (e.g., ratings compared with a true score). Ryan et al. found that there was not a significant difference between direct and indirect observation.

In the same study, Ryan et al. (1995) also compared the controlled and indirect observational methods. Raters in the controlled observational group paused (0-45 times) more than they rewound (0-13 times). Raters also indicated they found pausing more helpful than rewinding. The purpose for pausing was to give them more time to record observations and the purpose of rewinding was to observe something they may have missed or to ensure that they had not missed anything. Ultimately, Ryan et al. found that controlled observation did have some effects on accuracy, but the effects were neither large nor consistent. Therefore, they concluded that controlled observation did not increase accuracy.

This same report states that the advantages of controlled observation may not have been manifested in this study because of the short time span of the rating sessions. Raters were able to maintain a high level of alertness. The advantages may become more apparent when a rater is fatigued. Another limitation was that the observed situation was a group discussion that contained a lot of noise (behavior that was not pertinent to the rating objective). Ryan et al. suggest that the modes of observation should be compared in an exercise involving only one subject.

This dissertation incorporated some of the attributes that Ryan et al. (1995) suggested should be included in future research. The MTA contained much less noise since it incorporated only one missionary and one investigator. Although the teaching samples were not lengthy (10 min), each rater was required to rate many more samples (16 to 32) thus potentially leading to fatigue. Raters in the Ryan et al. study only rated one group discussion of an unknown length. One of the goals of this dissertation was to conduct another study to explore if and how the CRC and URC affect ratings. Ryan et al. analyzed their affects on accuracy and this study analyzed their affects on reliability.

Estimating Reliability

Assessments are used to make inferences about a person's true ability. The Missionary Teaching Assessment (MTA) is used to make inferences about how a missionary would perform in a broad range of similar situations. Ratings obtained from the MTA vary from missionary to missionary not only because each missionary possesses different levels of teaching ability but also because assessments always contain a degree of measurement error.

Measurement error can come from many sources including the following:

1. Raters—Raters come with biases and vary in their level of rating severity.
2. Teaching occasions—A teacher's teaching performance may vary from day to day due to the teacher's understanding of the content of the lesson presented, the kinds of questions and concerns raised by the learner, and anxiety, illness, lack of sleep, etc. on the part of the teacher.
3. Rating occasions—A teacher's ratings may differ from one rating occasion to another due to their ability to focus, mood, etc.
4. Teaching tasks—One teaching task may be more difficult than another.

There may also be interactions among all of these sources of error. For instance, a teaching task may be less difficult for one teacher than another because the subject matter elicited by the task is fresh in his mind since he just happened to study it earlier that day. To create an assessment that makes reliable inferences about a person's ability, these sources of error must be mitigated.

Multiple frameworks have been created to quantify measurement error. Two include generalizability theory and many-facet Rasch measurement. The strengths and weaknesses of the two have been debated. This study will use both frameworks to estimate the measurement error associated with MTA ratings.

Generalizability theory. Generalizability theory (G theory) is a framework for analyzing how well observed scores allow users to make generalizations about a person's behavior in a defined universe of situations (Shavelson & Webb, 1991). Instead of partitioning an observed score into just two parts, the true score and the error as found in classical test theory, a G-study partitions the error variance into multiple components representing several different sources of error. Knowledge of the relative size of the different variance components permits researchers to make informed decisions about how to improve a measurement procedure. Another advantage of using G theory is that it can estimate the reliability of the mean rating for each examinee while simultaneously accounting for both interrater and intrarater inconsistencies as well as inconsistencies due to various possible interactions. Classical reliability procedures do not allow the researcher to simultaneously estimate the amount of measurement error from multiple sources.

In this section, the researcher will discuss the following aspects of G theory: (a) facets, (b) relative versus absolute decisions, (c) G-study summary statistics, and (d) D-studies.

Facets. As previously stated, G-studies partition error into multiple sources. Each major source of error (e.g., raters and occasions) is called a facet and each level of a facet (e.g., number of raters and number of items) is called a condition. Before a G-study can be conducted, the facets and conditions must be defined (Shavelson & Webb, 1991).

Sources of variability that contribute to measurement error include not only the individual facets, but also all possible interactions among the facets and between the various facets and the object of measurement. In a one-facet design, there are three sources of variability: (a) the object of measurement, (b) the facet, and (c) the interaction between the object of measurement, the facet, and any additional random or unidentified variance. For example, if persons (p) were the object of measurement and items (i) were the facet, then the three sources of variability would be p, i, and p × i, e.

As additional facets are introduced, the number of sources of variability grows in a nonlinear fashion due to the increasing number of possible interactions. A two-facet design has 7 sources of variability. If the facets in a two-facet measurement were items (i) and occasions (o), then the sources of variability would include the following:

1. persons (p)
2. items (i)
3. occasions (o)
4. p x i interaction
5. p x o interaction
6. i x o interaction
7. p x i x o, e interaction, unidentified or random variability

A three-facet fully crossed design produces estimates of 15 sources of variability; a four-facet fully crossed design yields estimates of 31 sources of variability, etc.

In G theory facets are classified as either fixed or random. A *random facet* is one where there are an infinite number of conditions associated with it. If the levels selected for a particular facet in a study are treated as a random sample that could be exchanged with any other sample from the same universe, then the facet is classified as random. For example, if multiplication items were the facet and a sample of multiplication items is included in the study, then the facet is random. The selected multiplication items could be exchanged with another random set of multiplication items.

Fixed facets have a limited number of conditions that are not considered exchangeable. For example, if the items in a test are selected to assess only two different subject-matter areas (e.g., reading and mathematics), then this facet would be classified as fixed since there are only two conditions and the user is not interested in generalizing to other subjects.

The distinction between random and fixed facets is important because it affects how the error is calculated. Error variance for a fixed facet is calculated by averaging over the conditions of the facet. In instances where it is not logical to average across conditions (e.g., averaging scores of teacher behavior taken from math and reading instruction) then a separate G-study should be conducted for each condition of the fixed facet (Shavelson & Webb, 1991).

Facets can also be defined as being fully crossed or nested. Facets are fully crossed when every level of one facet appears in conjunction with every level of another facet. For instance, the facets raters and occasions are fully crossed if each rater rates each person on every occasion. If Raters 1 and 2 rate the first occasion and Raters 3 and 4 rate the second occasion, then raters would be nested within occasions (r:o). "One facet is said to be nested within another facet when two or more conditions of the nested facet (raters) appear with one and only one condition of another facet (occasions)" (Shavelson & Webb, 1991, p. 11). A study that has both fully crossed and nested facets is called a mixed design.

When the procedure for collecting ratings involves a nested design, the variance associated with the nested facet cannot be calculated independently of the other facets. For example, in a two-facet fully crossed design where the facets are raters (r) and occasions (o), the seven sources of error are persons (p), raters (r), occasions (o), person-by-rater interaction (p x r), person-by-occasion interaction (p x o), rater-by-occasion interaction (r x o), and person-by-rater by occasion interaction and any additional random or unidentified variance (p x r x o, e). However if raters are nested within occasions, the variance associated with raters cannot be separated from the variance for occasions. Only five as opposed to seven sources of variance can be calculated and they are persons (p), occasions (o), person-by-occasion interaction (p x o), raters combined with the occasion-by-rater interaction (r, r x o), and the person-by-rater interaction combined with the person-by-rater-by-occasion interaction and any additional unmeasured variance (p x r, p x r x o, e). Since the same raters did not rate all occasions, it is impossible to know the variance that is uniquely attributable to raters. Thus, the rater effect is confounded by the occasion-by-rater effect. Again, since the variance from raters cannot be parsed from the variance of the occasion-by-rater interaction, it is impossible to calculate the variance for the person-by-rater interaction. The person-by-rater interaction is confounded by the three-way interaction and unmeasured error.

Relative versus absolute decisions. Another aspect to consider when designing a G-study is whether relative or absolute decisions will be drawn from the output. A relative decision is drawn when a person's standing among other individuals is the focus (norm-referenced). For instance, an orchestra teacher who wanted to rank order his students through a performance assessment would care only about how many mistakes a student made compared to another student in the class. Absolute decisions are drawn when attention is given to how well someone performs relative to an absolute level of performance and not relative to his or her peers (criterion-referenced). This distinction is important because it affects how the overall error variances are calculated.

G-study summary statistics. After the design of a G-study has been defined (e.g., number of facets, absolute vs. relative decisions, and nested vs. fully crossed design) the variance among ratings can be partitioned into its components and four possible summary statistics can be calculated. The statistical model used to partition the variance is the analysis of variance (ANOVA). The magnitude of each variance component tells us how much each facet contributes to the overall measurement error.

Two of the four summary statistics are the relative error variance and the absolute error variance. The relative error variance is used for relative decisions and the absolute error variance is used for absolute decisions. The error variances are a sum of two or more variance components estimated in the G-study. In relative decisions, only the variance components that have an interaction with the object of measurement are used in defining the error. If the MTA was used to make decisions about how well a missionary performed relative to other missionaries in his group, a relative decision, then we would use the error variances that interact with the object of measurement, which in this case is missionaries (m). If raters (r) and occasions (o) were our facets, the relative error variance would be the sum of the error variances from the missionary by rater interaction, missionary by occasion interaction, and the missionary by rater by occasion interaction that also includes any unmeasured or unsystematic variance. If we wanted to compare a missionary to a benchmark score, an absolute decision, then the absolute error variance would be the sum of all variance components except the object of measurement (Shavelson & Webb, 1991).

The error variances along with the variance component for the object of measurement are used to compute the two reliability coefficients. The reliability coefficient for relative decisions known as the *g*-coefficient includes the relative error variance in its denominator. The reliability coefficient for absolute decisions known as the phi (Φ) coefficient includes the absolute error variance in its denominator.

D-study. While G-studies estimate the magnitude of the various sources of error, a D-study uses information from the G-study to design a measurement procedure that will minimize the sources of error. The D-study projects how each of the two error variances and each of the two reliability coefficients described above vary as a function of changing the number of raters, rating occasions, and teaching occasions. Additionally, a D-study projects how the four summary statistics vary in size as a function of using a different design (e.g., nested design) to collect the ratings.

Many-facet Rasch measurement. Many-facet Rasch measurement (MFRM) is an extension of the simple, one-parameter Rasch Model. Instead of assessing just one facet, tasks, MFRM can assess multiple facets simultaneously including sources of systematic error from raters, occasions, and tasks (Sudweeks, Reeve, & Bradshaw, 2004).

Like G theory, MFRM can partition the error variance into multiple sources, but it can break it down even further. MFRM gives group-level statistics for each facet analogous to the main effects calculated in G theory, but it also provides individual-level statistics. It allows researchers to assess each individual person, rater, occasion, and item. If one rater rates more severely than other raters, then MFRM would allow a researcher to detect that and implement an intervention.

MFRM also gives researchers more detailed information about each facet. Myford and Wolfe (2003) report that MFRM gives researchers the ability to analyze the following five rater errors: (a) leniency/severity, (b) central tendency, (c) restriction of range, (d) halo, and (e) differential leniency.

MFRM also provides fit statistics that show how well each facet at a group or individual-level performs relative to the expected value predicted by the MFRM model. The fit statistics are reported as mean squares that are calculated by dividing a chi-square statistic by its degrees of freedom. Fit statistics include infit and outfit statistics. Outfit statistics are highly influenced by outliers. On the other hand, infit statistics are weighted and are more sensitive to unexpected patterns of small residuals or nonoutliers. The fit statistics have an expected value of 1 and a range from zero to infinity. If a fit statistic is less than 1, then the data are probably redundant, dependent, or constricted. However, if a fit statistic is greater than 1, the data are inconsistent, contain unexpected variability, or are subject to extremism (Smith & Kulikowich, 2004). If a facet is performing as expected, then both the infit and outfit statistics will fall between 0.5 and 1.5 (Linacre, 2002).

Just as G theory provides an overall reliability statistic known as the *g*-coefficient, MFRM provides two different reliability statistics known as the *reliability of separation* *index* and the *separation ratio*. The reliability of separation index is analogous to estimates of internal consistency such as Cronbach's alpha coefficient and ranges from 0 to 1.0. This statistic shows how much variance exists among conditions or elements within a facet along the continuum. Since it is desirable for the variance to come from actual differences among people and not from other facets associated with a measurement, it is desirable for the reliability of separation index for persons to be as close to 1.0 as possible and all other facets to be as close to zero as possible.

The separation ratio ranges from 1.0 to infinity. Like the reliability of separation index, high values are desirable for the person facet and low values are desirable for all other facets.

G theory versus MFRM. In the literature, researchers have compared G theory and MFRM in order to determine their strengths and weaknesses. Lynch and McNamara (1998) compared these two model using data from an ESL speaking skills PA. They compared them using the analogy of a microscope. MFRM has a high level of magnification and allows a researcher to examine every imperfection. G theory has a lower level of magnification and allows a researcher to see the net effect of the blemishes. MFRM revealed to Lynch and McNamara numerous person-by-rater and person-by-item interactions that were biased, whereas the G-study revealed to that these biases were washed out at the aggregated level. With the G-study's group-level statistics and the D-study, G theory is useful in making decisions concerning test design while MFRM provides information to make adjustments among particular raters and items.

Sudweeks et al. (2004) concluded that G theory and MFRM both have their strengths. The focus of the two methods differs thus making their appropriateness dependent on the research context. The information from both analyses can be used to complement each other.

Smith and Kulikowich (2004) compared the two measurement models using scores from a complex problem-solving skills assessment. They found that the relative magnitudes of the variation among the facets were comparable, but they differed in how they handled the sources of variation. A major difference between the two models is that G theory assumes that the measurement scales are interval when many measurement scales are in fact ordinal. "This makes valid comparisons between individuals or items difficult as equal raw score differences between pairs of points do not necessarily imply equal amount of construct under investigation" (Smith & Kulikowich, p. 621). Because MFRM is based on a standardized unit of measurement, logits or logarithm of odds, a researcher is able to compare various facets to each other. For instance, the difficulty level of an item, the ability level of a person, and the severity of a rater can all be compared to one another.

Another advantage of the MFRM model is that parameters can be estimated separately from one another. The ability level of a person is not affected by the distribution properties of items or raters. If a person receives a particularly difficult item and is rated by a severe rater, their ability score will be adjusted accordingly so that they can be accurately compared to other people in the sample. MFRM is not dependent on an evenly distributed sample. The facet estimates should remain constant across various samples from the same population.

G theory does not possess this attribute and is more dependent upon the distribution of its sample. Statistics from a G-study are affected by the severity of raters and the difficulty of items. G theory requires homogeneity among raters and items. Each rater and item should be interchangeable with any other rater and item in the universe of possibilities. MFRM supports heterogeneity within the facets.

MFRM produces statistics at both the overall or group-level and the individual-level giving information on individual items, persons, and raters. G theory only produces group-level statistics.

Overall, as discussed above, MFRM contains many advantages over G theory. Because of these differences between the two models, Smith and Kulikowich (2004) recommend that researchers select a model that is appropriate to their purpose.

In his master's thesis, Alharby (2006) compares two different approaches to scoring, holistic and analytic, as well as two methods of assessing the reliability of a measure, G theory and MFRM. Alharby explored the interaction between the two scoring approaches and the two methods of measuring reliability. He conducted a G-study and found that the analytic scoring method had a higher *g*-coefficient than holistic scores. When he conducted an MFRM analysis, he found that the holistic scoring method had a better fit than the analytical method.

Studies that have compared G theory and MFRM to one another seem to agree that both models have their advantages and disadvantages. The decision of which model to use should be dependent on the purpose of the study. Because both models have their strengths and weaknesses, the researcher chose to analyze the MTA data using both models in order to create a more comprehensive picture of how ratings from the MTA were performing.

Chapter 3: Method

Study Participants

Missionaries. The missionaries selected to participate in this study were native English-speaking missionaries preparing to serve in English-speaking missions and thus not learning a foreign language. Missionaries not learning a foreign language stay in the MTC for 3 weeks. The sample included 32 missionaries in their final week at the MTC. This particular group of missionaries was selected to participate for the purpose of simplifying the process. Since the focus of this study was the instrument itself and not to make generalizations about the missionary population as a whole, it was less important to have a group that was representative of the general missionary population.

Raters. Six raters were selected to participate in this study. They were current employees of the MTC who had previously served as missionaries. Four of the raters were current MTC teachers and had never had any experience with the instrument before this study. They were familiar with the criteria since they teach these teaching skills on a regular basis, but they had never seen the MTA rubric before. The fifth rater was an employee of the MTC Research and Evaluation department. He/she had seen the instrument before but had never used it to rate missionaries. The sixth rater was a former employee in the MTC Research and Evaluation department who had recently accepted a job in another department at the MTC. He/she had been integrally involved in the development of the instrument and had used it to rate missionaries during pilot studies. These six raters were a convenience sample selected based on their availability or experience working with the MTC Research and Evaluation department. The raters were trained on using the rating scale and were paid for the time they spent in training, rating, and being interviewed.

Investigators. Two investigators were selected to participate in this study. One investigator was taught by half of the missionaries and the other investigator was taught by the other half. Only two were used to decrease the variance that is introduced from different investigators. The investigators were Teaching Resource Center (TRC) employees. Because missionaries have a lot of flexibility as to what they teach and how they teach it, the investigators had to give some unscripted responses. They were instructed beforehand by an MTC Research and Evaluation employee to keep their responses as uniform as possible. They were to present their concern when it was elicited by the missionary and they were not to create any other concerns.

Design

Because of limitations on their availability, four MTC raters rated the performance of only 16 of the 32 missionaries. The remaining two raters rated the performance of all 32 missionaries. It was important to have the two raters rate all 32 missionaries in order to create connectivity among the data which is an important element in the MFRM analysis although it did not resolve all of the connectivity issues. Having different raters rate different numbers of missionaries led to an unbalanced design which caused some problems in the G-study. The researcher will discuss these problems in greater detail later in this chapter. Overall, four different raters rated each missionary. Table 1 illustrates this design. Each X in the table represents an observation. Although it would have been ideal to have a fully crossed design where every rater rated every missionary, such a design was not possible due to limited resources.

Table 1

Study Design

	Controlled rating condition						Uncontrolled rating condition					
M	R1	R2	R3	R4	R5	R6	R1	R2	R3	R4	R5	R6
1	X		X							X		X
2	X		X							X		X
3	X		X							X		X
4	X		X							X		X
5	X		X							X		X
6	X		X							X		X
7	X		X							X		X
8	X		X							X		X
9		X	X								X	X
10		X	X								X	X
11		X	X								X	X
12		X	X								X	X
13		X	X								X	X
14		X	X								X	X
15		X	X								X	X
16		X	X								X	X
17				X		X	X		X			
18				X		X	X		X			
19				X		X	X		X			
20				X		X	X		X			
21				X		X	X		X			
22				X		X	X		X			
23				X		X	X		X			
24				X		X	X		X			
25					X	X		X	X			
26					X	X		X	X			
27					X	X		X	X			
28					X	X		X	X			
29					X	X		X	X			
30					X	X		X	X			
31					X	X		X	X			
32					X	X		X	X			

X = rating obtained; M = missionary; R = rater.

Each of the 32 missionaries taught a 10-minute lesson. Pilot studies indicated that 10 minutes was sufficiently long for a rater to observe and make an informed judgment on the relevant criteria. All of the missionaries received the same teaching situation and taught one of the two investigators.

Each rater observed half of the teaching performances using the uncontrolled rating condition (URC) meaning they watched the recordings through only once without being able to pause or rewind. Raters used the controlled rating condition (CRC) for the other half of the ratees meaning they had the ability to pause or review the recordings at will.

Instrumentation

As stated in the Introduction, the MTA is a procedure that includes having missionaries teach a person acting as investigator with some task specification. The performance is then rated by a rater based on a rubric. It was created by the MTC Research and Evaluation department to assess a missionaries' ability to teach the gospel. Although it may serve other functions at a future time, the current purpose of this assessment is to measure and track missionary performance for administrative uses, not to provide feedback or aid the missionaries in improving their performance.

Assessment procedure. For this study, a missionary was put in an actual teaching situation with a TRC employee role playing the part of an investigator. They were in a room that was made to look like someone's home, the environment a missionary would typically teach in. Missionaries usually teach in pairs, but for the purpose of this assessment, only one missionary taught at a time. They did not have a companion with them. This was done so that the missionary who taught second would not have an unfair advantage because he knew what the investigator's concern was before he began teaching.

Before the missionary entered the room, he read a situation telling him what he would be teaching the investigator and the context. The investigator also read the situation and was prompted as to how he should respond to the missionary. Each missionary received the same teaching situation and had 10 minutes to complete the task. Although the researcher wanted to include more than one teaching situation and more than one teaching occasion in the study to understand the effect they had on the ratings, she was not able to do so because of limited resources at the MTC. The teaching situation required the missionary to go beyond just the presentation of material. It focused on a problem or need of an investigator that the missionary had to identify and handle. The rating criteria focused on the interaction between the missionary and investigator and the missionary's ability to adjust his teaching to fit the investigator's needs. Figures 1 and 2 contain the situation each missionary and investigator received in this study.

Rating scales. The criteria used to rate the missionaries' teaching is founded on the content of the missionary training manual entitled *Preach My Gospel* (2004, see chapters 3, 10, and 11) and the *Effective Teaching* document written by the Missionary Department. Although this instrument assesses many different teaching skills, this study assessed only the following five criteria:

1. Shows Warmth and Concern
2. Listens
3. Asks Questions
4. Adjusts to Needs
5. Invites Others to Make Commitments

Instructions to Missionary: You are in the middle of teaching the Plan of Salvation lesson to an investigator. You have previously taught Lesson 1. You should try to accomplish the following in this teaching visit:

- Begin teaching "Kingdoms of Glory"
- Identify questions or concerns the investigator may have and adapt your lesson accordingly.
- Invite the investigator to make a commitment.

Figure 1. Description of teaching situation presented to missionaries.

Instructions to Investigator: You are meeting with the missionaries for the second time. They are teaching you about Plan of Salvation, specifically about the Kingdoms of Glory. Do the following in your role as an investigator:

- Ask whether they believe that only people from the LDS Church will be able to make it into the Celestial Kingdom?

Note: The missionary should be proactive in discovering your feelings/concerns, so please give them the opportunity to ask questions before volunteering this information. If too much time has passed and it looks like they aren't going to ask the right questions, feel free to interject your question/concern.

Figure 2. Description of teaching situation presented to investigators.

The other four criteria that were purposely not included in this study are described in detail in Appendix A and are summarized below:

1. Begins the Lesson
2. Teaches for Understanding
3. Uses Scriptures
4. Testifies

These criteria were excluded to reduce the number of attributes raters had to focus on simultaneously. The Begins the Lesson criterion was not included because the teaching sample had to be kept short so the directions given to missionaries instructed them to start in the middle of the lesson.

Each criterion is rated on a 7-point scale anchored by four descriptors. A descriptor is provided for the first, third, fifth, and seventh levels. The scale was originally a 5-point scale but was changed to a 7-point scale in hopes of helping the raters to more clearly differentiate among missionaries' teaching ability, thereby increasing the variability in the ratings.

The rating scales were developed by the MTC Research and Evaluation department through an iterative process. A group of MTC employees collaborated in creating these scales. They were tested many times using examples of actual missionaries teaching. The scales were revised over a period of time based on a series of iterative tryouts and revisions. After the rating scales reached what was believed to be a stage of acceptability, the MTC administrative president critiqued them. Further revisions were made from his feedback. The final rating scales that were used in this study can be found in Appendix A.

Procedure

The selected missionaries' teaching experiences were recorded by a digital camera in the Teaching Resource Center (TRC). Two rooms were set up, each with one of the selected investigators and were equipped with a discrete digital camera. Teaching in the TRC is a regularly scheduled event that missionaries participate in during their time at the MTC so the missionaries were aware of the format and the fact that they were being recorded.

The missionaries were not explicitly told what criteria they would be rated on but the criteria could be inferred from the teaching situation they were given. The criteria were based on teaching skills they had been taught in their classes. Missionaries are expected to integrate the selected teaching skills in every lesson they teach. The situation and instructions they were given should have elicited the target behaviors. The instruction to "identify questions or concerns the investigator may have and adapt your lesson accordingly" should have prompted the missionary to ask questions (Criterion 3), listen (Criterion 2), and adjust to needs (Criterion 4). The instruction to "invite the investigator to make a commitment" parallels Criterion 5, Invites Others to Make Commitments. Missionaries are taught to include Criterion 1, Shows Warmth and Concern, in every lesson they teach.

A missionary was taken to a room where one of the trained investigators was waiting for him. Before he entered the room, the missionary was given a written description of the teaching task and situation. He was given an opportunity to read through the situation and organize his thoughts which typically took 1 to 2 minutes. An MTC employee facilitated the process (e.g., directed him to the correct room and provided him with the situation) and asked if he had any questions. If the missionary asked questions, the facilitator tried to clarify the instructions without expanding on them. After all questions were answered, the missionary entered the assigned room, met the investigator, and began teaching. The missionary was given 10 minutes to teach the lesson and interact with the investigator. At the end of the 10 minutes, the facilitator knocked on the door giving the missionary the indication that his time is up. Missionaries were not required to use the entire 10 minutes.

Once all 32 selected missionaries recorded their teaching experiences, the rating process began. Prior to rating, the six raters participated in a two-hour training session provided by the researcher. The training included an introduction to the study, an overview of the study design, and an introduction to the MTA including an explanation of the five criteria and their scales. The raters were taught about common rater errors such as halo and central tendency so that they could avoid such behaviors. The researcher instructed the raters on how to use the video player as well as the screen capture software, Camtasia Studio (2010). The raters practiced rating approximately eight teaching samples that were not a part of the study. The missionaries in these teaching samples did not teach the same principles as the missionaries in the study. Some of the teaching samples lasted much longer than 10 minutes so only a portion of the teaching sample was viewed. The raters watched the samples together and then rated them individually. After rating, each rater shared the ratings they assigned to the missionary and explained why they gave those ratings. Any discrepancies among the ratings were discussed and the raters sought to come to a consensus on the appropriate rating. The raters did not receive any instruction concerning how they should manipulate the recordings. Any manipulations were left up to their discretion.

The order in which ratees are rated can often have an effect on their ratings. If a missionary is always rated last, then raters may have a tendency to be more lenient or sloppy because they are fatigued and want to finish rating. If a missionary is always rated first, a rater may be too lenient or severe because they have not seen the performances of other missionaries and therefore do not have anyone else to compare him to. To control for factors like these, the researcher randomized the order in which each rater rated their assigned missionaries. Also, raters switched off every other time between rating a missionary using the CRC and the URC.

The raters recorded their screen while they were rating using the CRC with screen capture software known as Camtasia Studio (2010). This software made a digital recording of everything the rater saw on the screen while they were rating as well any movement of the cursor and any manipulation of the recording. The researcher randomly selected missionaries out of each rater's rating pool that were rated using the CRC. Three missionaries were selected from the rating pools of the raters that rated 16 missionaries while six missionaries were selected for the two raters who rated 32 missionaries. After a rater completed rating one of these selected missionaries, the rater watched the entire teaching sample over again and performed a think-aloud where they verbalized the thought processes they had as they were rating. Raters were instructed to be candid in their responses and to also indicate each time they manipulated the recording and why they did so. The researcher made an audio/visual recording of each think-aloud. The protocol for the think-alouds is contained in Appendix C. Each rater was also given an exit interview upon completion of all of their ratings. The questions included in the exit interview are contained in Appendix D.

Analyses Used to Address the Research Questions

Research Question 1: Rater behavior. Research Question 1 asked, “When raters were able to control the pace in which they viewed digital recordings of missionary trainees’ performances (e.g., pause, rewind, etc.), to what extent did they use this capability to review previously viewed segments?” To answer this question, Research Question 1 contained four subquestions. The analysis plan for each subquestion is contained in the following sections.

Research Question 1a: Raters’ usage of controlled rating condition. To answer Research Question 1a, the researcher analyzed each Camtasia Studio (2010) recording and made a record of what manipulations were made and when they were made. The counts were then analyzed using descriptive statistics to determine how frequently each rater used any type of manipulation as well as how frequently he/she used particular manipulations such as rewind and pause.

Research Question 1b: Raters’ reasons for reviewing recordings. The think-alouds and exit interviews provided the data to answer Research Question 1b. The researcher transcribed each think-aloud and interview and then coded them according to the reasons the raters gave for pausing or reviewing segments of the recordings. The results were then aggregated in an attempt to identify patterns or common reasons why the raters manipulated the recordings.

Research Question 1c: Time spent reviewing recordings. The Camtasia Studio (2010) recordings were used to determine how much additional time each rater spent engaged in reviewing segments. The length of each pause was timed and the time added from reviewing segments was recorded. Any time that was saved from fast-forwarding was also taken into consideration. These three factors on time were combined to determine the additional time added on to each recording by each rater.

Research Question 1d: Changes in reviewing behavior across criteria. The researcher hoped to answer Research Question 1d using the data from the think-alouds and exit interviews. Unfortunately, the raters were not as specific in their think-alouds concerning why they made each manipulation. Using the data that were provided, the researcher used the coded think-alouds to determine if reviewing behavior varied from one criterion to another.

Research Question 2: Variability attributable to each source of variance. Missionaries were the object of measurement and raters, rating conditions, and any interaction between the sources of variability were the sources of error. Rating condition was a fixed facet since it only had two distinct levels—the CRC and URC. Users of G theory have two options in dealing with fixed facets: they can either average scores across the levels of a fixed facet or conduct a separate G-study for each level. The researcher chose to conduct separate G-studies for each rating condition so that the two rating conditions could be compared to one another. Therefore, the G-study had a one-facet design. There were three sources of variance:

1. Missionaries (m)
2. Raters (r)
3. The interaction between missionaries and raters plus any unmeasured or unsystematic error (m × r, e)

Before conducting the G-study, another obstacle had to be addressed. The study design was unbalanced because two of the raters rated 32 missionaries each while the other four raters rated 16 missionaries each. Unbalanced designs are usually caused by nesting or missing data. G-studies are unable to handle these designs. Unbalanced designs are common among large scale performance assessments. Raters often have different time constraints and vary in the number of assessments they rate. Many studies have been conducted to find a way to deal with this issue (see Chiu, 1999). One method of dealing with unbalanced designs is to implement the subdividing method.

In the subdividing method, a data set is broken into smaller subsets that have designs that are conducive to G-studies (i.e., crossed, nested, and modified balanced incomplete block (MBIB) designs). A separate G-study is conducted for each of the subsets. The variance components from the separate G-studies are then synthesized using the following weighted mean equation:

$$\bar{\sigma}_{f}^{2} = \frac{\sum_{t=1}^{r}\sum_{r=1}^{s_t} n_{p,t,s}\hat{\sigma}_{f,s}}{\sum_{t=1}^{r}\sum_{r=1}^{s_t} n_{p,t,s}} . \qquad (1)$$

where f = the variance components for missionaries (m), raters (r), or the interaction between the two (m × r, *e*)

s = the s^{th} data subset

t = the t^{th} dataset (criteria and rating condition combination)

$n_{p,t,s}$ = number of examinees in the s^{th} data subset of the t^{th} dataset

Overall indices like the generalizability and dependability, or phi, coefficient as well as D-studies can then be calculated from these aggregated variance components.

The subdividing method is based on a framework used in meta-analyses (Hedges & Olkin, 1985) where an overall outcome is estimated based on data from several disparate empirical studies. Studies have determined this method to be unbiased, consistent, and accurate. Chris Chiu and Edward Wolfe have published a number of studies that have investigated the properties of variance components when the subdividing method is used. These two men conducted a study using a Monte Carlo simulation and found that results from the subdividing method were similar to those produced by balanced datasets (Chiu, 1999; Chiu & Wolfe, 1997, 2002).

To apply the subdividing method, the data were first separated by criteria and rating condition resulting in 10 different datasets. Because of the unbalanced design, the 10 datasets were further divided into four subsets. Figure 3 provides an example of how each of the datasets was subdivided. Each subset contained eight missionaries and two raters and the two variables were fully crossed. A G-study was conducted for each of the four subsets resulting in a total of 40 analyses. Variance components were calculated using a weighted mean of the variance components from the individual subsets (see Equation 1). Calculating a weighted mean is not necessary when all data subsets have the same sample size. A standard calculation of the mean would provide the same results as a weighted mean. Although each of the data subsets was supposed to have eight missionaries, this was not always the case due to missing data. Because of technical difficulties like glitches in the recordings, some raters felt unable to give ratings on some of the criteria for particular missionaries and so they reported a zero for them indicating there was no basis for judgment. In these cases, the researcher used the listwise deletion method where she removed the entire record that contained the missing data point. Therefore, some subsets only had ratings for seven missionaries instead of eight.

		Rater					
Subset	Missionary	1	2	3	4	5	6
1	1	X		X			
	2	X		X			
	3	X		X			
	4	X		X			
	5	X		X			
	6	X		X			
	7	X		X			
	8	X		X			
2	9		X	X			
	10		X	X			
	11		X	X			
	12		X	X			
	13		X	X			
	14		X	X			
	15		X	X			
	16		X	X			
3	17				X		X
	18				X		X
	19				X		X
	20				X		X
	21				X		X
	22				X		X
	23				X		X
	24				X		X
4	25					X	X
	26					X	X
	27					X	X
	28					X	X
	29					X	X
	30					X	X
	31					X	X
	32					X	X

X = Rating collected

□ = No rating collected

Figure 3. Division of four subsets for each criterion and rating condition combination.

The subsets were analyzed using GENOVA software (Crick & Brennan, 1984). Each G-study produced a variance component for each of the three sources of error (missionaries, raters, missionaries by raters combined with any additional error). Each source of error was divided by the total amount of error to produce the proportion of variance attributable to each source.

The MFRM analysis was conducted using Facets version 3.66.0 software (Linacre, 2010). The model included four facets: (a) missionaries, (b) raters, (c) rating conditions, and (d) rating criteria. The unbalanced study did not pose a problem in the MFRM analysis since the data were sufficiently connected to one another. The following statistics were reported for the missionary and rater facets:

1. Individual- and group-level logit measures
2. Individual- and group-level infit mean squares
3. Individual- and group-level outfit mean squares
4. Separation reliabilities
5. Separation ratios
6. Fixed chi-squares
7. Random chi-squares

Research Question 2a: Reliability across criteria. G-study and MFRM statistics were used to answer this question. G-studies provide two different overall reliability coefficients. One is the generalizability coefficient, or the *g*-coefficient, and the other is the dependability coefficient, or the phi (Φ) coefficient. *G*-coefficients are used when a relative decision is made meaning that examinees are being compared to one another. Phi coefficients are used when absolute decisions are being made meaning that examinees are being compared to established criteria and not to one another. Because the purpose of the MTA is to determine how missionaries as a whole are doing compared to the criteria established in *Preach My Gospel*, the researcher chose to use the phi coefficient in her analyses.

Phi coefficients were calculated for each criterion using the variance components estimated from the G-study. In order to determine if there was a statistically significant difference between the reliabilities of the various criteria, the researcher needed an appropriate statistical test. Feldt (1969) devised a hypothesis test to assess whether or not Cronbach's alpha reliability coefficients on a single test for two groups were the same. The test statistic is derived using the following equation:

$$F = \frac{1-\mathrm{alpha}_1}{1-\mathrm{alpha}_2}. \qquad (2)$$

The equation is a ratio of 1 minus the alpha coefficient for the first group to 1 minus the alpha coefficient for the second group. The group with the largest variance is always placed on top. The test statistic is distributed as *F* with $n_1 - 1$ degrees of freedom in the numerator and $n_2 - 1$ degrees of freedom in the denominator.

G theory does not calculate alpha coefficients, but it does calculate generalizability coefficients which are the G theory equivalent to alpha coefficients. Phi coefficients are a variation of the generalizability coefficient. Thus, it was logical to use Feldt's test in this study to determine if the phi coefficients from the various criteria differed significantly.

The following statistics from the MFRM analysis were reported for each criterion:

1. Difficulty measure
2. Infit and outfit mean squares
3. Separation reliability

4. Separation ratio
5. Fixed chi-square

The researcher was also interested in looking at the statistics for the missionaries and raters facets. In order to analyze the criteria independently of one another, the researcher had to run a hybrid model in Facets where only one criterion was considered at a time. This caused a problem with the connectedness of the data. Connectedness of the data is an essential part of conducting an MFRM analysis. It is not essential for every rater to rate every missionary on all criteria, but it is essential for each facet to be linked to one another through connecting observations (Linacre & Wright, 2002). For instance, if one judge rated all missionaries on one criterion and another judge rated all missionaries on a different criterion, there would be no way to determine if differences between the two criteria were a result of varying severity levels of the judges or if the criteria differed in difficulty. Therefore, more than one judge should rate each criterion so that comparisons can be made across criteria and judges.

When running the analysis on the individual criterion, Facets indicated that four disjointed, or disconnected, subsets existed in the data. Each missionary was rated by four different raters—two using the CRC and two using the URC. The disconnectedness occurred because there was no way to link judges and rating conditions across each missionary. Raters 1 and 4 never rated the same missionaries as Raters 2 and 5 within the same rating condition. It was therefore impossible to determine if a difference in ratings from one rating condition to the next was due to the affect of the rating condition or the differences among the raters since Raters 1 and 4 were not connected to Raters 2 and 5.

In order to resolve the lack of connectedness, one of the disconnected facets had to be anchored (Linacre, 2010). The researcher chose to anchor the rating conditions facet at zero since she was interested in drawing conclusions about the reliability of the criteria and not rating conditions.

Five separate MFRM analyses were run, one for each of the five different criteria. The missionary reliability estimates for the criteria were then compared using Feldt tests to determine if there was a statistically significant difference between any of them.

Research Question 2b: Impact of number of raters on reliability. A D-study was conducted using the GENOVA software. The D-study provided phi coefficient estimates for a varying number of raters.

Research Question 2c: Reliability across rating conditions. The phi coefficients from the G-study for the two rating conditions were compared using Feldt tests to determine if the rating conditions affected the reliability differently.

Linacre (2010) provides guidelines to determine if the data fit the Rasch model. The standardized residuals (StRes) are expected to be near 0.0. Standardized residuals are the residuals divided by their standard errors. When the data fit the model, no more than 5% of the absolute value of the standardized residuals is greater than 2.0 and no more than 1% of the absolute value of the residuals is greater than 3.0. The standardized residuals from each rating condition were assessed to determine if they fit these criteria.

The following statistics from the MFRM analysis where all criteria were included in the model were assessed:

1. Logit measures
2. Infit and outfit mean squares

3. Separation reliability
4. Separation ratio
5. Fixed chi-square

In order to analyze the effect of rating conditions for a particular criterion, a hybrid model had to be analyzed where only one rating condition was considered at a time. Again, connectedness was a problem. The researcher had to create connectedness by anchoring another facet besides the rating condition facet. The researcher chose to anchor raters to zero. When a disconnected data set is analyzed in Facets, the output indicates how many disconnected subsets exist. Next to each element in each facet, the Facets software indicates which subset it belongs to. Facets determined that Raters 1 through 3 were in one subset and Raters 3 through 6 were in a second subset. Using this information, the researcher indicated in her input which group each rater belonged to and anchored those groups to zero.

Missionary separation reliability estimates were obtained for each criterion and rating condition and were compared using Feldt tests to determine if rating condition had an effect on reliability within the individual criterion.

Research Question 2d: Impact of the use of the CRC on reliability. Only MFRM was used to answer this question since G-studies are unable to determine the reliability of each individual rater. A hybrid model was used in order to consider each rater separately in the analysis. Missionary separation reliability estimates were gathered for each rater as well as infit and outfit mean squares. The missionary separation reliability estimates were compared with Feldt tests to determine if there were any differences in reliability among the raters. The separation reliability for each rater was also compared with how frequently they manipulated their recordings to determine if there was any relationship between them.

Another hybrid model was used to determine the missionary reliability for each rater and rating condition. Ten separate MFRM analyses were conducted. The missionary reliabilities for the two rating conditions within a rater were compared to see if there were any raters who had more reliable ratings in one rating condition than the other.

Research Question 3: Performance of rating scale categories. A separate MFRM analysis was conducted for each criterion. Facets output provides a graph that plots the probability of occurrence for each category.   These graphs as well as the rating scale statistics provided by Facets allowed the researcher to understand how each rating scale performed.

Chapter 4: Results

The intent of Research Question 1 was to explore how raters used the CRC. To what extent do raters use the capability to manipulate the recordings and why do they use it? The following four sections answer the subquestions of this research question.

Research Question 1a: Raters' Usage of the Controlled Rating Condition

Because of technical difficulties and/or mistakes on the part of the raters in using the Camtasia Studio (2010) screen capture software, not all of the ratings using the CRC were recorded for later evaluation. Of the 64 screen captures that should have been created, only 59 were actually recorded. The following statistics are based on these 59 screen captures.

During the CRC raters took advantage of the capability to manipulate the recordings more often than not. Overall, raters manipulated the recordings in one way or another in 81% of the recordings with an average of 3.69 manipulations per recording.

Individual raters varied in how frequently they manipulated the recordings. Table 2 shows some descriptive statistics pertaining to the frequency of their manipulations. Column 2 provides the number of screen captures obtained from each rater. Column 3 documents the total number of times each rater used any type of manipulation in the recorded screen captures. Column 4 contains the average number of times each rater manipulated a recording per video and column 5 is the associated standard deviation. Column 6 is the percent of times some type of manipulation was used in a recording (number of recordings where a manipulation was actually used divided by the number of recordings where a manipulation could have been used). Column 7 is the minimum and

Table 2

Raters' Usage of Video Manipulations

Rater	Missionaries	Total manipulations	*M*	*SD*	% of recordings where one or more manipulations were used	No. of manipulations in a single recording	
						Minimum	Maximum
1	8	22	2.75	1.83	88%	0	6
2	8	33	4.13	3.27	88%	0	10
3	15	108	7.20	6.37	100%	2	27
4	6	25	4.17	3.54	83%	0	8
5	7	14	2.00	1.29	86%	0	4
6	15	16	1.07	1.16	53%	0	3

column 8 is the maximum number of times a manipulation was used in any one of their particular screen captures.

The raters varied in how frequently they took advantage of their ability to manipulate the recordings. Rater 3 manipulated the recordings most frequently with an average of 7.20 manipulations per recording. This rater used some sort of manipulation in 100% or his/her recordings and used it a maximum of 27 times in a single recording. Rater 6 manipulated his/her recordings the least with an average of 1.07 manipulations per recording. This rater only used this functionality in 53% of the recordings and never made more than 3 manipulations in any given recording. Rater 3 manipulated his/her ratings nearly 7 times more frequently than Rater 6.

Use of various types of manipulation. In this section, the researcher will disaggregate the data into the various types of manipulations (pausing, rewinding, and fast-forwarding) and explore how each of these functions was used in the rating process.

Pausing. Raters used the pause function a total of 28 times in the 59 recordings. These 28 pauses took place within 21 (36%) of the recordings. The raters also varied in how frequently they utilized this function. Figure 4 illustrates the percentage of recordings where a particular rater paused one or more times. Rater 5 paused more frequently than the other raters using this function in 57% of his/her recordings. Rater 2 paused least frequently using it in only in 13% of his/her ratings. Raters 1, 3, 4, and 6 varied little from each other in their use of pause with 33% to 40%.

Rewinding. Rewind was used much more frequently than pause. Among the six raters, 139 rewinds were documented in the 59 screen captures that were analyzed

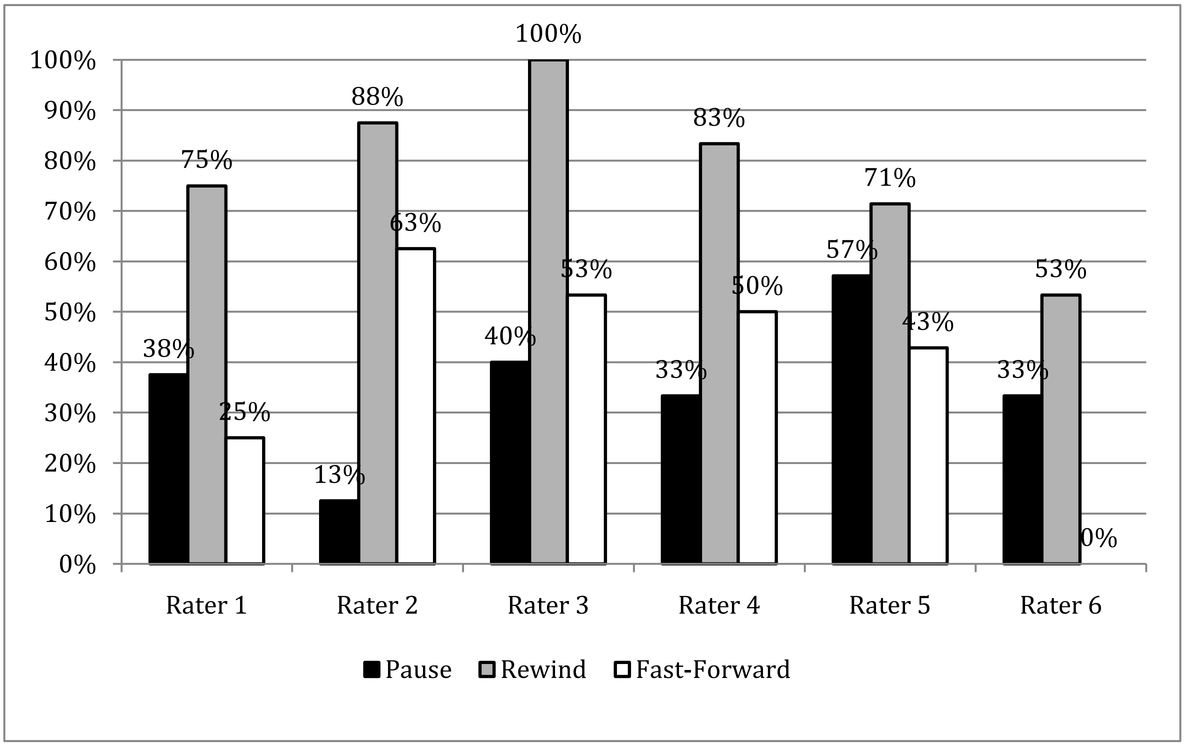

Figure 4. Percentage of recordings where pause, rewind, or fast-forward function used one or more times.

resulting in an average of 2.4 rewinds per recording. The rewind function was used in 78% of the screen captures.

Figure 4 reports the percentage of recordings in which each rater used the rewind function one or more times. Rater 3 used the rewind function most frequently. This rater used it in 100% of his/her recordings. Rater 6 used rewind the least with use in only 53% of his/her recordings.

Fast-forwarding. The frequency with which the fast-forward function was used was more similar to the pause function than rewind. A total of 51 fast-forwards were used in 21 or 36% of the recordings. Figure 4 reports the percentage of recordings in which each rater used the fast-forward function one or more times. Raters 2 through 5 used fast-forward in 43% to 63% of their recordings. Rater 6 never touched the fast-forward button and Rater 1 only used it in 25% of his/her recordings.

A different way of aggregating the video manipulation data. The above figures on how frequently each rater manipulated the recordings can be slightly misleading because in many instances, raters used the rewind or fast-forward function multiple times concurrently in an effort to find the appropriate starting point to review a segment. The frequency with which rewind or fast-forward was used could be attributed to the fact that a rater made poor judgments concerning how far back or forward a particular point of interest was or that they were unwilling to review portions of the recording that were not pertinent to what they were searching for. For example, Rater 3 used pause 1 time, rewind 9 times, and fast-forward 17 times when rating missionary 25 for a total of 27 manipulations. This rater used an exorbitant number of manipulations because he/she would rewind, watch the video for a few seconds, realize that he/she was not in the appropriate location, rewind again, watch the video for a few more seconds, realize that he/she had rewound too far, fast-forward, etc. The majority of the rewinds and fast-forwards were used merely to locate a particular segment of the recording. When the researcher combined multiple rewinds and fast-forwards together when they were used to locate a single segment, the manipulations were reduced to only 10 manipulations in this particular rating for Rater 3. Because of this, the researcher collapsed multiple rewinds, fast-forwards, and/or a combination of the two when they were used together to locate a particular segment to see how this would affect the picture of how the various functions were used.

In this section, the researcher used the same categories of rewind, pause, and fast-forward, but she aggregated the data in the rewind and fast-forward categories differently. Instead of counting each individual rewind or fast-forward, she combined multiple rewinds, fast-forwards, or a combination of the two if they occurred within 6 seconds of each other. She chose 6 seconds because from her experience in reviewing the screen captures, most raters were able to determine within 1 to 6 seconds of watching a recording whether or not they were in the correct place. Six seconds is not long enough to review a segment, but it is long enough to get an idea of where one is at in a recording. If a rater used a series of rewinds, fast-forwards, or a combination of the two to ultimately reach a position before the point where they began the manipulation, then the researcher counted the multiple manipulations as a single rewind. If a rater used a series of rewinds, fast-forwards, or a combination of the two to ultimately reach a position after the point where they began the manipulations, then the researcher documented it as being a single fast-forward.

In order to distinguish between these collapsed categories of rewind and fast-forward and the original categories, the researcher will refer to the collapsed categories as rewind′ and fast-forward′ from this point forward. Table 3 shows how the overall frequencies were affected by this new method of categorizing the manipulations. The statistical categories followed by a prime symbol (" ′ ") are the statistics that were calculated using rewind′ and fast-forward′. For some raters, this new method of counting rewinds and fast- forwards dramatically affected the frequency with which they used any type of manipulation. Rater 3 still had the most manipulations, but he/she went from 108 total manipulations to 71. His/her average fell from 7.20 to 4.73.

The researcher believes the picture portrayed by rewind′ and fast-forward′ is ultimately a more accurate picture of the rater behavior. When raters wanted to shift the video to review a particular segment, some only used rewind or fast-forward once while others used them numerous times. Although they had different methods of getting there, their ultimate goal was to back up or move forward to review a particular segment. The following two sections report the data for rewind′ and fast-forward′.

Rewind′. Figure 5 shows the frequencies for both the rewind and rewind′ categories for each rater. For four of the six raters, rewind′ was significantly less than rewind. The rewind′ categories were about 25% to 33% less than rewind. This means that 25% to 33% of their rewinds were used in conjunction with other rewinds or fast-forwards to ultimately locate a previous position in the recording. Raters 5 and 6 were unaffected by collapsing the categories. When they rewound, they only used it once to locate a previous position in the recording.

Table 3

Raters’ Usage of Video Manipulation

						% of recordings where one or more manipulations used		Number of manipulations in a single recording					
Rater	Missionaries	Total manipulations	Total′ manipulations	*M*	*M*′	%	%′	Min.	Min.’	Max.	Max.’	*SD*	*SD*′
1	8	22	15	2.75	1.88	88%	88%	0	0	6	3	1.83	1.13
2	8	33	21	4.13	2.63	88%	88%	0	0	10	7	3.27	2.33
3	15	108	71	7.20	4.73	100%	100%	2	2	27	10	6.37	2.69
4	6	25	15	4.17	2.50	83%	83%	0	0	8	5	3.54	1.87
5	7	14	13	2.00	1.86	86%	86%	0	0	4	3	1.29	1.07
6	15	16	16	1.07	1.07	53%	53%	0	0	3	3	1.16	1.16

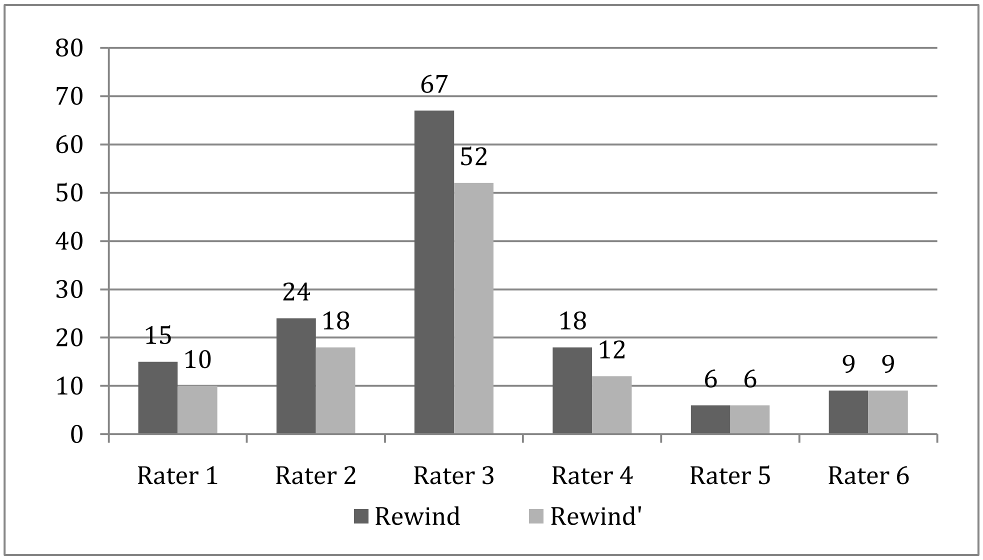

Figure 5. Frequency of rewind and rewind′.

Figure 6 shows the average use of rewind and rewind′ per recording. The averages for rewind and rewind′ fell for Raters 1 through 4 by 22% to 42%. Again, the averages for Raters 5 and 6 were unaffected.

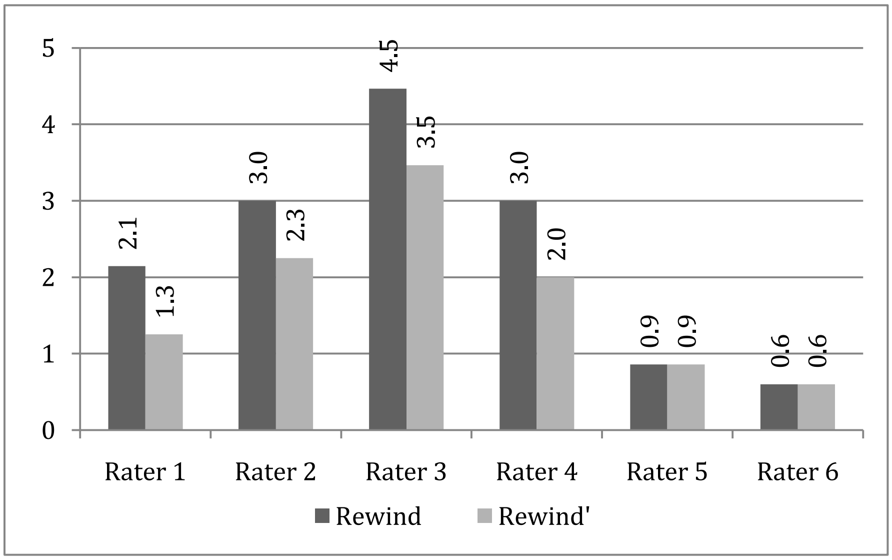

Figure 6. Average use of rewind and rewind′ per recording.

Fast-forward′. Figures 7 and 8 show the difference in the fast-forward and fast-forward′ categories. The difference between the two categories is much more dramatic than the differences between rewind and rewind′. This indicates that the majority of fast-forwards were not used to locate a segment after an initial position in the recording. They were most likely used with other rewinds to locate an earlier position but that the desired position was overshot by the rater. Rater 6 was unaffected since they never used fast-forward while rating. Raters 1 through 4 all had a 66% to 100% reduction in frequency from the fast-forward to the fast-forward′ category. Rater 5 only had a 33% reduction in frequency.

Summary. Although there was a large amount of variance in how frequently raters manipulated the video recordings, they used this capability the majority of the time when they were able to do so. One rater used it just a little more than half of the time while another rater used it 100% of the time. The other four raters used it anywhere from 83% to 88% of the time.

Although the frequency of rewind′ and fast-forward′ were generally lower than rewind and fast-forward, the relative standing of the raters in how frequently they used rewind or fast-forward remained the same. Rater 3 still used these functions most frequently and Rater 6 used them the least.

Research Question 1b: Raters' Reasons for Reviewing Recordings

Table 4 shows the various reasons raters gave for rewinding. The second column shows the percentage of raters who reported rewinding for that particular reason. The most frequently reported reason was the volume. Five of the six raters reported going back and reviewing a segment because the recording was too quiet or the missionary or

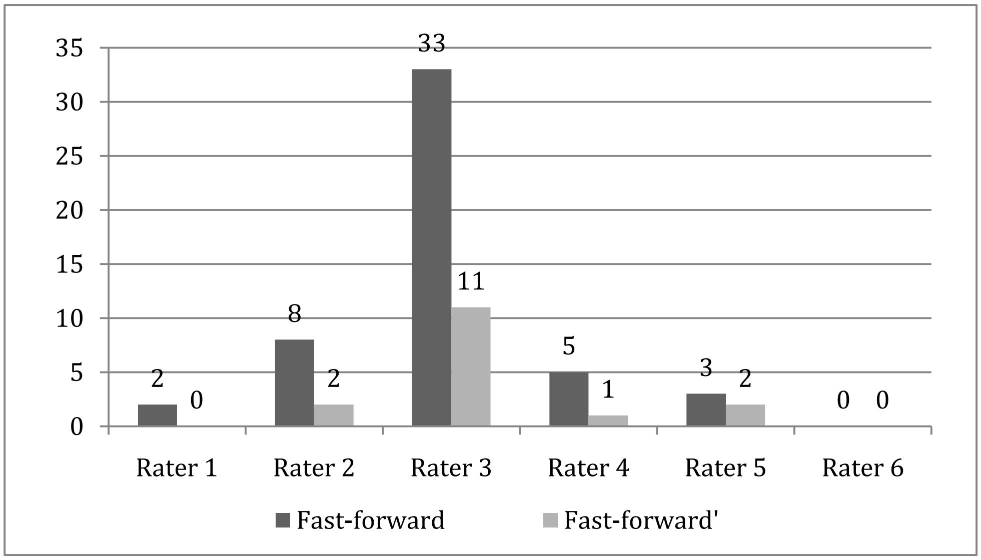

Figure 7. Frequency of fast-forward and fast-forward′.

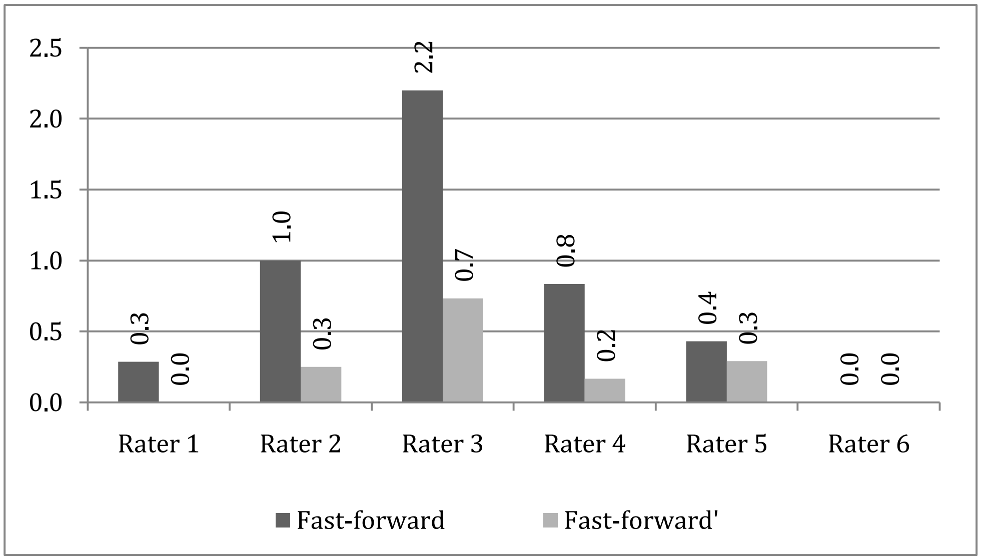

Figure 8. Average use of fast-forward and fast-forward′ per recording.

investigator spoke unclearly or quietly. Four of the raters reviewed recordings to collect more data in order to make more informed decisions using the rubric. At the end of a recording, raters realized they were not sure if any questions were asked or any invitations were extended leaving them unable to give a rating for those criteria. Therefore, they had to take some time to search through the recording for further evidence.

Table 4

Reasons Raters Reported for Reviewing Recordings

Reason for reviewing recording	% of raters who reported reviewing video for this reason
Volume/could not hear	83%
Collect more information to make a rating decision	67%
Technical difficulty/glitch	50%
Distracted/tired/busy writing down a rating	50%
Review something that was unexpected or confusing	50%
Reviewed something they thought was funny or interesting	33%
Interruptions	17%

Three raters, although not the same raters in each instance, reported technical difficulties, distractions, and reviewing something confusing as reasons for rewinding. Some of the recordings contained slight glitches in them. They did not skip over any of the content, but they caught many of the raters off guard and made them think that part of the recording had been cut out. Distractions were a problem for a few of the raters. One rater did not get much sleep the night before and was consumed with thoughts about proposing to his girlfriend that night. One of the raters who rated all 32 recordings became very fatigued and found himself/herself often losing focus. Some raters got caught up on something a missionary had said or done causing them to miss subsequent portions. For instance, one missionary appeared to be nervous and was popping his knuckles and stuttering a bit. This precluded one of the raters from paying attention to what was actually being said. The other rater who reported being distracted took many notes while watching the recordings and found him/herself missing portions of them at times.

Two of the raters enjoyed reviewing the recordings because something funny or interesting occurred. As is the case with any novice practicing a new skill, the missionaries sometimes made mistakes that were humorous. Only one rater reported needing to rewind because of an interruption. A person came in while he/she was rating and asked a question.

Research Question 1c: Time Spent Reviewing Recordings

Missionaries were given 10 minutes to demonstrate their teaching skills. Some missionaries spent more time and some spent less but on average their recordings were 10 minutes 28 seconds long. Allowing raters to manipulate the recordings added on an average of 1 minute 45 seconds to each recording. Therefore, the time spent watching each recording was increased by an average of 16%.

There was significant variance among the additional time spent on the ratings by the various raters. One rater spent an additional 13 minutes 19 seconds on one recording. The standard deviation for additional time spent rating was 1 minute 59 seconds. Table 5 provides statistics on how much additional time was added by manipulating the recordings by rater. Column 2 contains the average amount of additional time each rater spent rating when using the CRC. Column 3 is the standard deviation among the added time. Columns 4 and 5 contain the minimum and maximum amount of additional time each rater spent when using the CRC.

Table 5

Additional Rating Time Added by Raters due to Video Manipulation

Rater	Average time added	*SD*	Min	Max
1	2:53	4:19	0:00	13:19
2	1:38	1:12	0:00	3:38
3	2:29	1:27	0:35	5:17
4	1:21	1:14	0:00	3:22
5	1:11	0:47	0:00	2:12
6	0:55	1:01	0:00	2:58
Total	1:45	1:59	0:00	13:19

In the previous section on how frequently each rater used any sort of manipulation, it was apparent that Rater 3 manipulated his/her ratings the most and Rater 6 the least. In Table 5, the average additional time for Rater 6 is the lowest, but the average for Rater 3 is not the highest. Rater 1 had the highest average additional time, but this is due to the fact that he/she had an unusually high amount of additional time during one rating. Rater 1 rewound and reviewed a nearly 11 minute portion of a recording. If this outlier was removed, his/her average additional time would drop to only 1 minute 23 seconds and Rater 3 would have the highest average.

Research Question 1d: Changes in Reviewing Behavior Across Criteria

There were 31 instances where raters indicated why they reviewed the recording. Table 6 provides the descriptive statistics concerning the criteria associated with each rewind and the reason they gave for needing to rewind. Those that gave volume as their reason for rewinding could not hear either because the recording was bad or the missionary was too quiet. When raters became distracted it was generally because they were thinking about something other than the rating task or they were thinking about things the missionaries had previously said. It was not because what was said was so complex that it led to cognitive overload. Raters reported being distracted because the missionary appeared nervous and stuttered or the missionary was doing something out of the ordinary that caught the rater off guard. Rewinds were categorized under *Recollection* when a rater rewound to confirm whether or not something did or did not happen. The *Confused* category indicates that the rater had to review a segment because something the missionary said was unclear. Rewinds associated with a glitch in the recording were categorized under *Glitch*.

Of the 31 reports from raters concerning why they reviewed a segment, 15 rewinds were not associated with any particular criterion. The large number of rewinds in the *None* category could be partially due to the fact that raters were not specific enough concerning why they rewound in the think-alouds, but in most cases, it appears that the rewinds were not associated with any criteria. A problem with the volume or a glitch occurred randomly in the recording. Only 1 rewind was associated with Shows Warmth and Concern. No raters reviewed a segment to gather more information on the Listens criteria. Asks

Questions had 2 rewinds and Adjusts to Needs and Invites Others to Make Commitments both had 5.

Table 6

Criteria Associated With Each Rewind and Reason

Criteria	Volume	Distracted	Recollection	Confused	Glitch	Total	Row %
None	5	5	0	3	2	15	54%
Shows Warmth	0	0	1	0	0	1	4%
Listens	0	0	0	0	0	0	0%
Asks Questions	1	1	0	0	0	2	7%
Adjusts to Needs	3	1	1	0	0	5	18%
Invites Others	4	0	1	0	0	5	18%
Total	13	7	3	3	2	28	100%
Column %	46%	25%	11%	11%	7%	100%	

From the data reported in Table 6, it appears that most of the rewinds occurred due to factors unassociated with any particular criteria. No single criterion caused cognitive overload. When a rewind was associated with one of the criteria, they were generally associated with the more objective rating criteria. In the exit interviews, the raters categorized the criteria as either objective or subjective. There was a consensus among the raters that Shows Warmth and Concern and Listens were subjective while Asks Questions, Adjusts to Needs, and Invites Others to Make Commitments were more objective. Evidences for the two subjective criteria were woven throughout the recordings and judgments came from raters' general impressions of the missionaries. Therefore, if a rater momentarily tuned out or was unable to hear what was said, this generally did not affect their ability to rate the missionary on these categories and they did not feel the need to review those segments. Evidences for the objective criteria were more distinct. An invitation either was or was not extended. If an invitation was extended, it either included the appropriate elements of being clear, direct, and appropriate, or it did not. Questions can be counted and they are simple and clear or they are not. When a rater knew that a question was asked or an invitation had been extended but they were unable to hear it, they almost always reviewed those segments. Raters would rewind a recording if they could not hear a missionary's response to an investigator's question. One rater did not realize that a missionary had invited an investigator to be baptized until he/she heard the investigator respond to the invitation. The rater had to back up to see if he/she could catch what the missionary had actually said. Sometimes raters would get to the end of a recording and realize that they were not sure if the missionary had asked any questions so they reviewed portions of the video to determine this.

For the most part, there was not much of a difference in viewing behavior from one rating criterion to the next. Most rewinds were not associated with any particular criterion. Raters reviewed segments throughout the teaching performances because they were unable to hear, confused, distracted, or they experienced a glitch in the recording. When a rater did review a segment that was directly connected to one of the criterion, it tended to be one of the more objective criteria which included Asks Questions, Adjusts to Needs, and Invites Others to Make Commitments.

Research Question 2: Variability Attributable to Each Source of Variance

Generalizability study findings. Because of the unbalanced study design, the researcher used the subdividing method. Forty separate G-studies were conducted (four for each criterion and rating condition). The variance components within each criterion and rating condition combination were synthesized using a weighted mean. The results from the G-studies are contained in Table 7. The estimated variance components are contained in columns 3 and 5. The variance components are reported as a percentage of the total variation for each criterion and rating condition. The percentages provide the relative magnitude of each variance component and allow for comparisons to be made across criteria and rating conditions.

Variance component for the missionary facet. The variance component for missionaries, the object of measurement, should ideally be larger than the variance from other sources. A large variance component would indicate that the majority of the variance in test scores was due to actual differences in the teaching ability of the missionaries and not from measurement error such as inconsistencies among raters. The variance from missionaries is considered true score variance while all other sources of variance are classified as error variance. The variance component for missionaries is larger than the variance from other sources for three of the five criteria. Exceptions include the CRC for Shows Warmth and Concern and Adjusts to Needs.

Invites Others to Make Commitments has the largest variance components for missionaries with 68% for the CRC and 75% for the URC. Adjusts to Needs has the smallest variance components with 36% and 57%. Overall, the variance components for missionaries for the URC were larger than the CRC. The percentages were 48% and 57% respectively.

Variance component for the rater facet. Unlike the object of measurement, raters should have as small a variance component as possible. A small variance component

Table 7

Amount of Variability in the Ratings Attributed to Each Source

	Controlled		Uncontrolled	
Source of variation	Variance component	Percent of variance	Variance component	Percent of variance
		Shows Warmth & Concern		
Missionary	0.884	36%	1.621	58%
Rater	0.250	10%	0.272	10%
M × R, *e*	1.344	54%	0.888	32%
		Listens		
Missionary	1.940	57%	2.071	51%
Rater	0.534	16%	0.554	14%
M × R, *e*	0.944	28%	1.468	36%
		Asks Questions		
Missionary	1.357	41%	1.108	42%
Rater	0.580	18%	0.433	17%
M × R, *e*	1.335	41%	1.072	41%
		Adjusts to Needs		
Missionary	1.446	36%	2.676	57%
Rater	0.536	13%	0.272	6%
M × R, *e*	2.027	51%	1.721	37%
		Invites Others to Make Commitments		
Missionary	2.502	68%	3.141	75%
Rater	0.317	9%	0.356	8%
M × R, *e*	0.862	23%	0.694	17%

indicates that very little of the difference in scores among missionaries is attributable to differences in the severity/leniency of raters. The variance components for raters in the MTA were relatively small. They ranged from 6% to 18%. Asks Questions had the largest variance components for raters with an average of 17.5%. Invites Others to Make Commitments had the smallest components with an average of 8.5%. The difference between the percent of variance for raters for the CRC and URC was minimal with 13% and 11% respectively.

Variance component for the residual. The residual consists of the interaction between missionaries and raters plus any additional unmeasured or unsystematic variance. Again, this component should be minimized as much as possible. The percent of variation among the residuals ranged from 17% to 54%. Adjusts to Needs had the highest variance components for the residual with an average of 44%. This indicates that raters were not consistent in their ratings across missionaries in this particular criterion. The URC outperformed the CRC in the level of variance from the residual. The magnitude of the variance for the CRC was 39% while that of the URC was 33%.

Because the ratings for the CRC and URC were analyzed in separate G-studies, the researcher was not able to determine what percent of the variance was attributable to rating conditions.

Many-facet Rasch measurement findings. The MFRM model included four facets: (a) missionaries, (b) raters, (c) rating conditions, and (d) rating criteria. Figure 9 contains the calibrations for each of these four facets. All facets are reported in a common logit scale which has equal intervals. The first column in Figure 9 contains the logit scale. The second column represents the ability levels of the various missionaries. Each missionary is

```
+--------------------------------------------------------------------------------------------+
|Measr|+Persons|-Rater          |-Rating Condition|-Criteria                           |Scale|
|-----+--------+----------------+-----------------+------------------------------------+-----|
|   2 +        +                +                 +                                    + (7) |
|     |        |                |                 |                                    |     |
|     |        |                |                 |                                    |     |
|     |        |                |                 |                                    |     |
|     |        |                |                 |                                    |     |
|     |        |                |                 |                                    |  6  |
|     | *      |                |                 |                                    |     |
|     |        |                |                 |                                    |     |
|     | **     |                |                 |                                    |     |
|   1 +        +                +                 +                                    + --- |
|     | **     |                |                 |                                    |     |
|     | ***    |                |                 |                                    |     |
|     | ****   |                |                 |                                    |     |
|     | ***    |                |                 |                                    |  5  |
|     | **     | Rater5         |                 |                                    |     |
|     | *      | Rater2         |                 |                                    |     |
|     |        | Rater3         |                 | Invites Others to Make Commitments | --- |
|     |        |                |                 | Asks Questions                     |     |
|     | **     |                | Uncontrolled    |                                    |     |
*   0 * ***    *                *                 * Adjusts to Needs                   *  4  *
|     | **     | Rater1  Rater6 | Controlled      |                                    |     |
|     | *      |                |                 | Listens                            |     |
|     | *      |                |                 | Shows Warmth & Concern             |     |
|     |        |                |                 |                                    | --- |
|     | *      |                |                 |                                    |     |
|     | **     |                |                 |                                    |     |
|     |        |                |                 |                                    |  3  |
|     |        | Rater4         |                 |                                    |     |
|     |        |                |                 |                                    |     |
|  -1 + **     +                +                 +                                    + --- |
|     |        |                |                 |                                    |     |
|     |        |                |                 |                                    |     |
|     |        |                |                 |                                    |  2  |
|     |        |                |                 |                                    |     |
|     |        |                |                 |                                    |     |
|     |        |                |                 |                                    |     |
|     |        |                |                 |                                    |     |
|     |        |                |                 |                                    |     |
|     |        |                |                 |                                    | --- |
|  -2 +        +                +                 +                                    + (1) |
|-----+--------+----------------+-----------------+------------------------------------+-----|
|Measr| * = 1  |-Rater          |-Rating Condition|-Criteria                           |Scale|
```

Figure 9. Facets map displaying calibrations of missionaries, raters, rating conditions, and criteria.

represented by an asterisk. Those missionaries at the top of the table have higher logit measures meaning that they have a higher ability level across the five rating criteria. The third column contains the rater calibrations. Higher logit measures indicate more severe raters. The calibrations of the two rating conditions are contained in the fourth column. The fifth column shows the criteria calibrations. Those with higher logit measures are more difficult. Missionaries tend to receive lower ratings on these scales. The last column contains the 7-point rating scale. Each level or category of the scale is aligned with its corresponding logit.

Missionary facet. Table 8 presents the estimated teaching ability measures for each of the 32 missionaries included in the analysis. They are sorted according to their ability measures in column 2 which are reported in logits. The missionary with the highest ability level was missionary 32 with an ability measure of 1.31. Missionary 20 had the lowest ability measure of -1.11. Infit and outfit mean squares should ideally fall between 0.5 and 1.5. Mean squares less than 0.5 or between 1.5 and 2.0 are unproductive to the construction of a measure, but they are not degrading. Mean squares greater than 2.0 distort or degrade a measure (Linacre, 2002). Three of the 32 missionaries had infit and/or outfit mean squares less than 0.5 and 2 had mean squares greater than 1.5. Only one missionary had a mean square greater than 2.0. Infit and outfit mean squares outside the acceptable boundaries are indicated in the table with an asterisk. Overall, the data had a good fit to the model.

The person separation reliability index ranges from 0 to 1.0 and is the Rasch analogue to Cronbach's coefficient alpha in classical test theory. It is a ratio of true variance to observed variance. Ideally, the separation reliability index should be high (i.e.,

Table 8

MFRM Analysis of Missionaries

Missionaries	Ability measure	Standard error	Infit mean square	Outfit mean square
32	1.31	0.20	1.17	1.13
3	1.10	0.21	0.45[a]	0.42[a]
12	1.08	0.19	1.38	1.34
9	0.93	0.19	1.11	1.11
21	0.86	0.20	0.55	0.56
7	0.84	0.20	0.68	0.69
5	0.80	0.20	0.85	0.80
23	0.78	0.20	1.06	1.10
28	0.73	0.18	3.03[a]	2.97[a]
22	0.70	0.20	0.76	0.78
24	0.70	0.20	0.39[a]	0.40[a]
6	0.69	0.20	0.92	0.89
31	0.59	0.18	1.58[a]	1.54[a]
10	0.59	0.18	0.83	0.83
27	0.56	0.18	0.66	0.66
11	0.47	0.19	1.34	1.31
19	0.46	0.18	1.71[a]	1.68[a]
17	0.36	0.21	0.55	0.56
8	0.12	0.18	0.56	0.56
2	0.09	0.18	0.73	0.77
30	−0.02	0.18	1.33	1.31
16	−0.02	0.18	1.04	1.04
25	−0.05	0.20	1.36	1.38
1	−0.08	0.18	0.89	0.90
4	−0.14	0.18	1.16	1.15
14	−0.23	0.18	1.19	1.18
26	−0.27	0.18	1.01	1.03
29	−0.45	0.18	0.77	0.77
13	−0.58	0.19	0.58	0.57
15	−0.58	0.18	0.73	0.74
18	−1.01	0.19	0.39[a]	0.41[a]
20	−1.05	0.19	0.90	0.88

[a] Infit or outfit statistics are less than 0.5 or greater than 1.5.

close to 1.0) for the missionary facet and low (i.e., close to zero) for the other facets. This is because it is desirable for the majority of the variance to come from actual differences among the missionaries and not measurement error from raters, rating conditions, etc.

The missionary separation reliability index across all criteria for the MTA was .90. This indicates a high level of variance among the ability levels of the examinees. The results obtained on the MTA would be replicable if taken by another random sample of missionaries.

The fixed chi-square was statistically significant (χ^2 (31, N = 32) = 325.2; p < .01). The fixed chi-square for the missionary facet tests if these missionaries can be thought of as equally able (Linacre, 2010). With a p value < .01, this hypothesis can be rejected. The MTA ratings successfully distinguished between the ability levels of various missionaries. Facets also calculates a random chi-square which tests if this set of missionaries can be regarded as a random sample with a normal distribution (Linacre). The random chi-square was nonsignificant (χ^2 (30, N = 32) = 28.4; p = .55). Thus, the null hypothesis was not rejected and we can assume that this sample of missionaries was random with a normal distribution.

Rater facet. Table 9 reports the output from the MFRM analysis for raters. Column 2 reports the measure of relative severity/leniency for each rater. Those with high measures were more severe while those with lower measures were more lenient. The rater severity measures ranged from -0.82 to 0.45.

The average infit mean square was 1.01 and the average outfit mean square was 0.99 indicating that overall, the observed scores were very close to the expected scores as predicted by the model. At an individual level, five of the six raters had infit and outfit mean squares within the acceptable range of 0.5 to 1.5. Rater 1 had an infit and outfit mean square of 0.42 indicating that this rater's observed scores were closer to the expected scores than the many-facet Rasch model would predict.

Table 9

MFRM Analysis of Raters

Rater	Severity/ leniency	Standard error	Infit mean square	Outfit mean square
5	0.45	0.09	1.35	1.34
2	0.37	0.09	1.28	1.28
3	0.28	0.07	1.07	1.06
4	−0.14	0.09	0.42[a]	0.42[a]
6	−0.14	0.07	0.86	0.85
1	−0.82	0.11	1.06	1.00

[a] Infit or outfit statistics are less than 0.5 or greater than 1.5.

The rater separation reliability index was .97. This high number indicates that there are real differences among the raters. Unlike the missionary separation reliability index, it is more desirable for the rater separation reliability to be as close to zero as possible. Raters should be equally lenient/severe and interchangeable with one another. There was a high level of unwanted variance in the severity/leniency among these raters.

Raters had a statistically significant fixed chi-square (χ^2 (5, N = 6) = 125.0; $p < .01$) and a nonsignificant random chi-square (χ^2 (4, N = 6) = 4.8; $p = .31$). Therefore, we can conclude that there are distinct differences in severity/leniency among the raters and that they are random and normally distributed.

Research Question 2a: Reliability Across Criteria

Generalizability study findings. Table 10 reports the phi coefficients in columns 2 and 3 when two raters rate each missionary for each criteria and rating condition. The closer the coefficient is to 1.00, the more reliable the ratings are. In column 4, the phi coefficients for the two rating conditions are averaged to produce a mean phi coefficient. The mean phi coefficients ranged from .590 to .833.

Table 10

Phi Coefficients by Criteria and Rating Condition for Two Raters

Criterion	Controlled Φ	Uncontrolled Φ	Mean Φ
Shows Warmth & Concern	.526	.736	.631
Listens	.724	.672	.698
Asks Questions	.586	.595	.590
Adjusts to Needs	.530	.729	.645
Invites Others to Make Commitments	.809	.857	.833

Many-facet Rasch measurement findings. Table 11 presents the difficulty measures as well as the fit statistics for the five criteria. Considerable variance exists among the difficulty levels of the criteria. Shows Warmth and Concern had the lowest difficulty measure of -0.32 and Invites Others to Make Commitments had the highest with 0.28 logits. The infit and outfit statistics indicate that these criteria were consistent with the model predicted by the MFRM. They also indicate that the various criteria were rated in a consistent manner among the judges. In general, missionaries were rated lower on more difficult scales and higher on less difficult scales. The separation reliability index was .95 and the separation ratio was 3.95 indicating that there were significant differences among the criteria and they are not interchangeable with one another. This conclusion is supported by a statistically significant fixed chi-square (χ^2 (4, N = 5) = 82.9; p < .01). Thus, some criteria demand higher levels of teaching ability than others.

Table 11

MFRM Analysis of Criteria

Criteria	Difficulty	Standard error	Infit mean square	Outfit mean square
Invites Others to Make Commitments	0.28	0.07	1.31	1.30
Asks Questions	0.25	0.07	0.96	0.96
Adjusts to Needs	−0.01	0.07	0.98	0.94
Listens	−0.20	0.08	0.96	0.97
Shows Warmth and Concern	−0.32	0.08	0.78	0.76
Mean	0.00	0.07	1.00	0.98
Standard deviation	0.26	0.00	0.19	0.19

As was described in the Method section of this dissertation, a hybrid model had to be run in Facets to determine how each criterion performed. The hybrid model introduced disconnectedness in the data, so the researcher chose to anchor the rating conditions facet at zero. The results for the following five criteria were calculated after anchoring the rating condition facet.

Missionaries. Table 12 contains the missionary group-level statistics for each of the five criteria. The average missionary separation reliability indexes ranged from .58 to .82. The four criteria with reliability estimates between .71 and .82 indicate that judges were able to distinguish between missionaries with high and low ability levels and central tendency was not a problem. These results would be replicable with a similar set of raters. However, the Adjusts to Needs criterion had a worrisome missionary separation reliability estimate of .58. This indicates that the raters were less able to distinguish between the performance of the various missionaries in this category and that a central tendency problem exists. The infit mean squares ranged from 0.93 to 1.02 and the outfit mean squares ranged from 0.91 to 1.02 thus fitting into the recommended range of 0.50 to 1.50.

Table 12

MFRM Missionary Group-Level Statistics for Five Criteria

Criteria	Missionary reliability	Infit mean square	Outfit mean square
Shows Warmth & Concern	.76	0.95	0.91
Listens	.71	0.98	0.96
Asks Questions	.75	0.95	0.95
Adjusts to Needs	.58	0.93	0.96
Invites Others to Make Commitments	.82	1.02	1.02

Using a Feldt test the missionary separation reliability indexes were compared to determine if any of the criteria were significantly more or less reliable than the others. The results from the Feldt test corroborated the findings from the G-study that the Invites Others to Make Commitments criterion was statistically significantly higher than the Listens criterion at an alpha level of .05 ($F_{\mathrm{C2C5}}(93, 93) = 1.61$, $p = .01$) and the Adjusts to Needs criterion ($F_{\mathrm{C4C5}}(93, 93) = 2.33$, $p < .01$). However, Invites Others to Make Commitments did not differ statistically significantly from Shows Warmth and Concern

($F_{C1C5}(93, 93) = 1.33$, $p = .08$) or Asks Questions ($\mathrm{F}_{C3C5}(93, 93) = 1.38$, $p = .06$). Another finding from the MRFM analysis that differed from the G-study was that the missionary separation reliability of the Adjusts to Needs criterion was significantly lower than not only the Invites criterion, but also the other three criteria ($F_{C1C4}(93, 93) = 1.75$, $p < .01$; $F_{C2C4}(93, 93) = 1.45$, $p = .04$; $F_{C3C4}(93, 93) = 1.68$, $p = .01$).

Raters. Table 13 contains the summary rater statistics for the five criteria. For each criterion, the rater reliability was high. They ranged from .83 to .95. These high reliabilities indicate that there were real differences among the raters in their levels of severity/leniency for each criterion. This is unwanted variance. Raters should be as similar as possible in their levels of severity so that they can be interchangeable with other raters.

Table 13

MFRM Summary Rater Statistics for Five Criteria

Criteria	Rater reliability	Infit mean square	Outfit mean square
Shows Warmth & Concern	.92	0.99	0.94
Listens	.83	1.00	0.96
Asks Questions	.95	1.01	1.00
Adjusts to Needs	.87	0.86	0.86
Invites Others to Make Commitments	.84	1.02	0.98

Research Question 2b: Impact of Varying Number of Raters on Reliability

Increasing the number of raters in an assessment increases the generalizability of an assessment. Averaging across a higher number of raters creates more stability and precision in the mean rating for each missionary. Using information about the variance of each facet from the G-study, a D-study projects how the reliability will increase as the sample sizes of the facets are increased. The phi coefficient was used in this study to measure overall test reliability since absolute decisions were being made. Figure 10 shows how averaging an examinee's score across an increasing number of raters increases the reliability for each criterion and rating condition. Each criterion is represented by a different line. Increasing the number of raters from two to three dramatically increases the reliability but this change diminishes after five raters. Increasing the number of raters to four would produce phi coefficients of .70 or greater for each criterion.

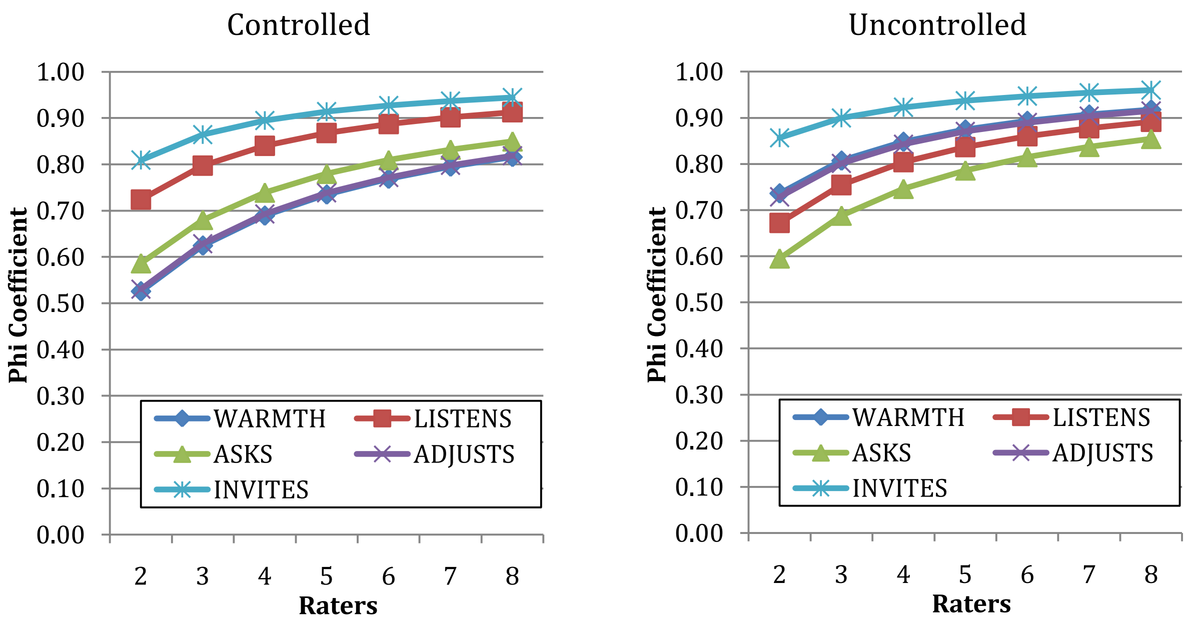

Figure 10. Projected reliability of absolute decisions obtained by varying the number of raters rating each examinee.

Research Question 2c: Reliability Across Rating Conditions

Generalizability study findings. In order to determine if there was a statistically significant difference between the two rating conditions, a Feldt test was conducted on the phi coefficients for the two rating conditions. The weighted mean phi coefficients across all five criteria were compared for the CRC and URC. This test resulted in a nonsignificant *F* statistic ($F(128, 133) = 1.284$, $p = .08$) thus failing to reject the null hypothesis which indicates that there was not a significant difference between the reliabilities of the two rating conditions.

The weighted mean phi coefficients for the rating conditions within each individual criterion were compared using the Feldt test to see if there were any significant differences (see Table 14). The only criterion that produced significant results between the two rating conditions was Shows Warmth and Concern ($F_{\mathrm{CU}}(31, 31) = 1.795$, $p = .05$). The URC proved to be more reliable than the CRC in this criterion.

Table 14

Phi Coefficients for Two Rating Conditions Across Five Criteria

Criteria	Controlled	Uncontrolled
Shows Warmth & Concern	.526	.736
Listens	.724	.672
Asks Questions	.586	.595
Adjusts to Needs	.530	.729
Invites Others	.809	.857
Average	.639	.719

Many-facet Rasch measurement findings. When the data fit the Rasch model, no more than 5% of the absolute value of the standardized residuals will be greater than 2 and no more than 1% of the absolute value of the residuals will be greater than 3 (Linacre, 2010). The data gathered from both rating conditions fit these requirements. The CRC produced 9 residuals (3%) outside the ±2 boundary and 2 residuals (1%) outside the ±3 boundary. The URC produced 10 residuals (3%) outside the ±2 boundary and 0 residuals (0%) outside the ±3 boundary. Therefore, both rating conditions met the stipulations laid out by Linacre for model fit using standardized residuals.

Table 15 presents the results from the MFRM analysis of the two rating conditions across the 5 criteria. The conditions differed from each other by 0.10 of a logit. The rating condition separation reliability index was .23 and the separation ratio was 0.54. The small separation ratio signifies that the variance between the rating conditions was less than the measurement error. This indicates that the variance introduced by the different rating conditions was negligible.

Table 15

MFRM Analysis of Rating Conditions

Rating condition	Measure	Standard error	Infit mean square	Outfit mean square
Uncontrolled	0.05	0.05	0.92	0.91
Controlled	−0.05	0.05	1.08	1.06
Mean	0.00	0.05	1.00	0.98
S.D.	0.05	0.00	0.08	0.08

The associated fixed chi-square was nonsignificant (χ^2 (1, N = 2) = 2.6; p = .08). This indicates that the two rating conditions were not significantly different from one another and the raters were consistent across rating conditions. Rating conditions did not have a significant impact on the variability of the ratings. Because rating conditions was a fixed facet, there was no random chi-square.

Although the variance introduced by rating conditions was negligible when all criteria were considered at once, the researcher was interested in knowing the impact of rating conditions when each criterion was considered separately. Table 16 contains the missionary separation reliability estimates for the two rating conditions by criteria. The reliabilities ranged from .57 to .92. Using Feldt tests, the researcher compared the two rating conditions to see if there was a significant difference between the reliabilities they produced. The only criterion that had a statistically significant difference between the two rating conditions was Listens ($F_{\mathrm{CU}}(31, 31) = 2.263$, $p = .01$). The CRC proved to be more reliable than the URC.

Table 16

MFRM Group-Level Missionary Reliability Estimates for Rating Conditions by Criteria

Criteria	Controlled	Uncontrolled
Shows Warmth & Concern	.70	.80
Listens	.81	.57
Asks Questions	.83	.81
Adjusts to Needs	.69	.51
Invites Others to Make Commitments	.92	.92
Average	.89	.88

Research Question 2d: Impact of the Use of the CRC on Reliability

Table 17 contains the missionary group-level statistics for each rater. Column 2 contains the missionary reliabilities that range from .75 for Rater 2 to .91 for Rater 1.

These high reliability estimates indicate that each rater had a sufficiently high level of reliability in their individual ratings. Columns 3 and 4 contain the fit statistics. Both the infit and outfit mean squares fell within the acceptable range of 0.50 to 1.50. This indicates that overall, the ratings awarded by each rater fit the model well.

Table 17

MFRM Missionary Group-Level Statistics by Rater

Rater	Missionary reliability	Infit mean square	Outfit mean square	% of recordings where one or more manipulations used	Average manipulations per recording
1	.91	0.95	0.93	88%	1.88
2	.75	0.98	0.99	88%	2.63
3	.88	0.98	0.99	100%	4.73
4	.87	0.76	0.86	83%	2.50
5	.87	1.02	1.02	75%	1.63
6	.89	1.00	0.97	53%	1.07

Column 5 contains the percent of recordings where they used one or more manipulations and column 6 contains the average number of manipulations each rater made per recording. The figures in column 5 and 6 are based on the calculations of rewind′ and fast-forward′. The correlation between the missionary reliabilities and the frequency with which raters manipulated the videos was -.14. This correlation is very weak and it cannot be concluded that any real relationship exists between the raters' use of the digital recordings and the reliability of their ratings.

Using a Feldt test, the researcher compared the missionary separation reliability indexes for each rater to one another to determine if there was a statistically significant difference between the intrarater reliabilities. Raters 1, 3, 4, 5, and 6 did not differ significantly from one another. However, the reliability for Rater 2 was significantly lower than the reliabilities from each of the other five raters ($F_{R2,R1}(60, 60) = 2.78$, $p < .01$; $F_{R2,R3}(60, 124) = 2.08$, $p < .01$; $F_{R2,R4}$ $(60, 60) = 1.92$, $p = .01$; $F_{R2,R5}(60, 60) = 1.92$, $p = .01$; $F_{R2,R6}$ $(60, 124) = 2.27$, $p < .01$). With the exception of Rater 3, Rater 2 manipulated the ratings more than any other rater.

Table 18 parses the missionary group-level statistics by rater and rating condition. These data allow us to determine if there was a difference in the reliability of ratings within raters according to the rating condition they used. Overall infit and outfit statistics for all raters and rating conditions lie between 0.50 and 1.50. Table 18 also provides the standardized residuals. Only Raters 1 and 5 were able to meet the standardized residual requirements in both rating conditions. Raters 2 and 4 had poor fit in the URC because more than 1% of Rater 2's data had an absolute standardized residual greater than 3 and more than 5% of Rater 4's absolute standardized residuals were greater than 2. Raters 3 and 6 had poor fit in the CRC because more than 5% of their absolute standardized residuals were greater than 2. From these results, the way the raters manipulated the digital recordings does not appear to have any systematic affect on the reliability of the data.

In order to determine if there was a statistically significant difference between the reliabilities for the CRC and URC for each rater reported in Table 18, the researcher conducted Feldt tests. Table 19 contains the results of these tests. The only raters that had a significant difference between the two rating conditions were Raters 1 and 4. For Rater 1, the URC produced more reliable results. Rater 4 had more reliable results in the CRC.

Table 18

MFRM Missionary Group-Level Statistics by Rating Condition and Rater

Rater	Measure	Infit mean square	Outfit mean square	% |StRes| ≥2	% |StRes| ≥3	Reliability	χ^2 sig
			Controlled				
1	2.14	0.92	0.92	0%	0%	.85	.00
2	−0.18	0.87	0.87	2%	0%	.72	.00
3	0.30	1.01	1.00	8%	0%	.89	.00
4	1.79	0.76	0.72	4%	0%	.90	.00
5	0.08	1.17	1.17	4%	0%	.89	.00
6	0.59	1.25	1.27	6%	1%	.85	.00
			Uncontrolled				
1	0.96	0.95	0.92	0%	0%	.92	.00
2	0.01	1.07	1.10	2%	2%	.77	.00
3	−0.22	0.88	0.85	4%	0%	.86	.00
4	2.14	0.95	1.17	7%	0%	.62	.00
5	−0.46	0.70	0.71	0%	0%	.80	.00
6	0.68	0.72	0.69	3%	0%	.87	.00

Table 19

Statistics by Rater from Feldt Test Comparing Rating Conditions

Rater	*df*	*F*	*p*-value
1	28, 28	1.875	.051
2	28, 28	1.217	.303
3	60, 60	1.273	.176
4	28, 28	3.800	<.005
5	28, 28	1.818	.060
6	60, 60	1.154	.291

There appears to be no pattern between how frequently a rater manipulated the recordings and the reliability of their ratings. Rater 4 was the only rater to have more reliable ratings for the CRC than the URC. The frequency with which this rater manipulated the ratings falls in the middle of the raters. This rater ranked fourth in the percent of recordings where manipulation occurred and third in the average number of times they occurred per recording. Rater 1's URC ratings were more reliable. Again, this rater's frequency of manipulating the recordings fell into the middle of the rankings among raters. Rater 3, who manipulated the recordings more than any other rater, had no significant difference between the two rating conditions. Rater 6 who manipulated the recordings the least also failed to have a significant difference in the reliability of the ratings between the two rating conditions. Therefore, it appears that a rater's use of the CRC had no systematic affect on the reliability of their ratings.

Research Question 3: Performance of Rating Scale Categories

MFRM provides a report on how each category within a rating scale functions. Table 20 contains the rating scale category statistics and Figure 11 presents the category probability curves for Shows Warmth and Concern. Column 4 in Table 20 contains the step calibrations for the different categories. A step calibration is the location on the logit scale where a category and the category preceding it are equally probably. For example, using the category statistics from Table 20, if a missionary had a logit score of -2.52, the model predicts they would have a 50-50 chance of being rated a 2 or a 3. The step calibrations should always proceed in order. A category should never contain a lower step calibration than the previous category. In Table 20, disordering exists in the step calibration between categories 6 and 7 indicating the raters were not always able to make a clear distinction between these two categories. In the future, MTA administrators may want to consider collapsing categories 6 and 7 or making them more distinctive. Also, raters never used category 1. The description for category 1 states, "Behaves in a disrespectful or disinterested manner." Missionaries rarely behave in this manner but the researcher believes it is important to include this category in case this behavior is ever displayed.

To test the hypothesis that collapsing categories 6 and 7 would improve the scale, the researcher ran another MFRM analysis with the two categories collapsed. Doing so resolved the disordering between the categories, but it unfortunately led to a lower separation reliability for missionaries. Based on these finding, other solutions should be considered for resolving the disordering.

Table 20

MFRM Rating Scale Category Statistics for Shows Warmth and Concern

Category	Instances	Percent	Step calibration	Standard error
1	0	0%		
2	5	4%		
3	21	16%	−2.52	0.50
4	25	20%	−0.57	0.28
5	36	28%	−0.01	0.25
6	14	11%	1.99[a]	0.26
7	27	21%	1.11[a]	0.29

[a] Step calibrations disordered.

```
-4.0             -2.0              0.0              2.0              4.0
     ++----------------+----------------+----------------+----------------++
   1 |                                                                     |
     |                                                             77777777|
     |222                                                      7777        |
     |   222                                                777            |
     |      22                                            77               |
   P |        222                                        7                 |
   r |           2                                      7                  |
   o |            22                                  77                   |
   b |              22                               7                     |
   a |                2                             7                      |
   b |                 22   3333                   7                       |
   i |                  3*33    333                                        |
   l |                33  2        33       55    7                        |
   i |              33     22        3   555  5557                         |
   t |           333        2     444***44     755                         |
   y |         33             22 44   553  44  7   5                       |
     |      333               44*2  55   33  4*  666**                     |
     |   333                44    **       37*6**     **666                |
     |333               4444    55  22    ***33  44     55566666           |
     |            444444   55555     6****     333 4444    555556666666666 |
   0 |*******************************77   22222222*************************|
     ++----------------+----------------+----------------+----------------++
    -4.0             -2.0              0.0              2.0              4.0
```

Figure 11. Category probability curves for Shows Warmth and Concern.

Table 21 and Figure 12 contain reports for the Listens scale categories. Disordering exists between categories 2 and 3 as well as among categories 5, 6, and 7. The step calibration for category 7 is lower than the step calibration for category 5. Again, this indicates that raters were not able to make clear distinctions among these categories. The categories should be collapsed or the category descriptions should be analyzed to determine if there is ambiguity. The researcher collapsed categories 2 and 3 as well as categories 6 and 7. The collapse resolved the disordering, but again, it led to a lower missionary separation reliability estimate.

Table 21

MFRM Rating Scale Category Statistics for Listens

Category	Instances	Percent	Step calibration	Standard error
1	1	1%		
2	3	2%	−1.94[a]	1.04
3	25	20%	−2.44[a]	0.55
4	32	26%	0.07	0.26
5	29	23%	1.02[a]	0.23
6	8	6%	2.70[a]	0.26
7	27	22%	0.59[a]	0.27

[a] Step calibrations disordered.

Also, categories 1, 2, and 6 were underutilized. The description for category 1 states, “Ignores, interrupts, or fails to listen to investigator.” Although it is possible, missionaries rarely demonstrate this behavior. For this reason it is clear why categories 1 and 2 were underutilized.

```
 -4.0             -2.0              0.0              2.0              4.0
  ++----------------+----------------+----------------+----------------++
1 |                                                                    7|
  |                                                             7777777 |
  |11111                                                    7777        |
  |     11                                                77            |
  |       11                                             7              |
P |         11                                         77               |
r |           11                                      7                 |
o |             1                                    7                  |
b |              1                                  7                   |
a |               1                                7                    |
b |                1       33333333                                     |
i |                 1    33        33             7                     |
l |                  1  3            3           7                      |
i |                   *3            44**4444    7                       |
t |                  3 11        444    3 55**5*55                      |
y |               2**22221     44       5*3   *4  55                    |
  |           22223      2*2244       55   3 7  4   55                  |
  |       2222  33         **222    55      *3   44   55                |
  |  22222   333        444  11 22*5      77 6**666**666**6             |
  |22   33333       4444     55***12222***666   333  4444 5****6666666  |
0 |*************************************1*******************************|
  ++----------------+----------------+----------------+----------------++
 -4.0             -2.0              0.0              2.0              4.0
```

Figure 12. Category probability curves for Listens.

Table 22 and Figure 13 contain the scale category reports for Asks Questions. There was a better distribution of scores across the 7 categories, but disordering did exist between categories 3 and 4 and categories 6 and 7. The researcher collapsed categories 3 and 4 as well as categories 6 and 7. This resulted in an ordered scale without any negative impact on the missionary separation reliability estimates. The missionary separation reliability increased by .01 and the rater separation reliability decreased by .01. Figure 14 shows the probability curves for the categories of the collapsed scale containing five categories.

Table 22

MFRM Rating Scale Category Statistics for Asks Questions

Category	Instances	Percent	Step calibration	Standard error
1	8	6%		
2	19	15%	−2.16	0.41
3	16	13%	−0.64[a]	0.28
4	35	28%	−1.02[a]	0.26
5	24	19%	0.76	0.25
6	11	9%	1.77[a]	0.30
7	14	11%	1.30[a]	0.35

[a] Step calibrations disordered.

```
 -4.0             -2.0              0.0              2.0              4.0
  ++----------------+----------------+----------------+----------------++
1 |                                                                     |
  |                                                              7777777|
  |1111                                                      7777       |
  |    1111                                               777           |
  |        11                                           77              |
P |          11                                        7                |
r |            11                                    77                 |
o |              11                                 7                   |
b |                1                               7                    |
a |                 11                            7                     |
b |                   1                          7                      |
i |                    1  2                     7                       |
l |                  222*2 222       444       7                        |
i |                22    11   22  444   44                              |
t |             222        1    2*       5**557                         |
y |           22            1  44 2    55   4755                        |
  |        222               **3333*355    77**6**6666                  |
  |     222              333*4 1   5**33  *66  4  555 6666              |
  |22222              333 44    **5   2***3     44   55   666666        |
  |             3333334444   555  1****72223333   4444 555555   66666666|
0 |********************************77 11111*****************************|
  ++----------------+----------------+----------------+----------------++
 -4.0             -2.0              0.0              2.0              4.0
```

Figure 13. Category probability curves for Asks Questions with original scale.

```
 -4.0             -2.0              0.0              2.0              4.0
  ++----------------+----------------+----------------+----------------++
1 |                                                                     |
  |                                                                     |
  |                                                                     |
  |                                                                   55|
  |111                                                              55  |
P |   11                                                          55    |
r |     11                           33                         55      |
o |       1                      3333  3333                   55        |
b |        11                  33          333               5          |
a |          1               33               3             5           |
b |           1            33                  33         55            |
i |            1122222222 3                      3       5              |
l |           2221      3*22                      3*4444*               |
i |         22    11   3    22                  444 3 55 4444           |
t |      222        133       22             444     *       44         |
y |    22           311         222        44       5 33       444      |
  | 222           33   11          22   444       55    33        444   |
  |2            33       11          ***        55        33         444|
  |          333           111   4444   2222 555            333         |
  |    333333              444***11     5555*22222             33333    |
0 |************************55555555*****1111111111**********************|
  ++----------------+----------------+----------------+----------------++
 -4.0             -2.0              0.0              2.0              4.0
```

Figure 14. Category probability curves for Asks Questions with collapsed scale.

Table 23 and Figure 15 contain the category reports for the Adjusts to Needs scale. Categories 5 and 6 contain a disproportionate number of ratings showing a restriction of range toward the upper end of the scale. The only disordering occurs between categories 4 and 5. The researcher collapsed the two categories resolving the problem with disordering but the reliabilities were negatively impacted indicating that this was not a good solution.

Table 23

MFRM Rating Scale Category Statistics for Adjusts to Needs

Category	Instances	Percent	Step calibration	Standard error
1	6	5%		
2	13	10%	−1.72	0.45
3	19	15%	−0.99	0.31
4	15	12%	−0.40[a]	0.26
5	32	25%	−0.54[a]	0.24
6	35	28%	0.51	0.22
7	6	5%	2.69	0.43

[a] Step calibrations disordered.

```
 -4.0             -2.0              0.0              2.0              4.0
  ++----------------+----------------+----------------+----------------++
1 |                                                                     |
  |                                                                     |
  |1111                                                                 |
  |    111                                                              |
  |       111                                                         77|
P |          11                                                     77  |
r |            11                                                 77    |
o |              1                                              77      |
b |               1                               66          77        |
a |                11                         6666  6666     7          |
b |                  1                      66          66677           |
i |                   1                    6              766           |
l |                    122               66             77   66         |
i |                 222211222         55*555          77       66       |
t |              222      1 3**333  55 6    55       7           66     |
y |           222         3*   22 **  6       555  77              66   |
  |        222          33  11   *44**           **                  666|
  |     222          333      ***42*64**4      77  555                  |
  |22222          333      44*51166 22  3*44*77       555               |
  |          33333     444*55 666111  2**7**3*4444       5555555        |
0 |***************************777777***11*******************************|
  ++----------------+----------------+----------------+----------------++
 -4.0             -2.0              0.0              2.0              4.0
```

Figure 15. Category probability curves for Adjusts to Needs.

Table 24 and Figure 16 contain the category reports for the Invites Others to Make Commitments scale. Table 24 shows that a disproportionate number of ratings fell in categories 4 and 5 manifesting a central tendency effect. Disordering occurs between categories 2 and 3 and between categories 6 and 7. The researcher analyzed the data after collapsing categories 2 and 3 as well as 6 and 7 creating a 5-point scale. This resolved the disordering without negatively impacting the reliabilities. Based on these findings, the researcher recommends changing this scale from a 7-point scale to a 5-point scale. Figure 17 contains the probability curves for the collapsed 5-point scale.

Table 24

MFRM Rating Scale Category Statistics for Invites Others to Make Commitments

Category	Instances	Percent	Step calibration	Standard error
1	10	9%		
2	4	4%	−0.84[a]	0.45
3	11	10%	−2.16[a]	0.41
4	30	27%	−1.35	0.33
5	36	32%	0.34	0.25
6	12	11%	2.24[a]	0.28
7	10	9%	1.78[a]	0.38

[a] Step calibrations disordered.

```
 -4.0             -2.0              0.0              2.0              4.0
  ++----------------+----------------+----------------+----------------++
1 |11111111                                                             |
  |        1111111                                                 77777|
  |               111                                         77777     |
  |                  11                                     77          |
  |                    1                                  77            |
P |                     1                               77              |
r |                      1                            77                |
o |                       1                          7                  |
b |                        1                        7                   |
a |                         1                      7                    |
b |                                               7                     |
i |                          1          55       7                      |
l |                           1      555  555   7                       |
i |                              444*4       557                        |
t |                            *4 55  44      755                       |
y |                           4 15      44   *666*66                    |
  |                          4  51        ***     556666                |
  |                        4*3**33*    6667 4       55  6666            |
  |                     3***2*     **** 77   444      555   66666       |
  |          222222****** 555 22********333     4444     55555   6666666|
0 |*****************************7777222*********************************|
  ++----------------+----------------+----------------+----------------++
 -4.0             -2.0              0.0              2.0              4.0
```

Figure 16. Category probability curves for Invites Others to Make Commitments.

```
 -4.0             -2.0              0.0              2.0              4.0
  ++----------------+----------------+----------------+----------------++
1 |                                                                     |
  |                                                                     |
  |                                                                     |
  |1111                                                                5|
  |    11                                                           555 |
P |      11                                                       55    |
r |        11                                                   55      |
o |          11                                               55        |
b |            1                                             5          |
a |             11                                         55           |
b |               1             333333333      444444444  5             |
i |                1 22222    33         33 444         **              |
l |               22*1    22**            4*3          5  444           |
i |            222    1    3  22        44   33      55      44         |
t |          22        1 33     22    44       33   5          44       |
y |       222          3*1        22 4           3*5             44     |
  |    222           33   11       4*2          55 33              444  |
  | 222            33       11  444   22      55     33               44|
  |2            333          4**1       222*55         333              |
  |       333333        44444    1111 55555 22222         333333        |
0 |*********************5555555555555*11111111111***********************|
  ++----------------+----------------+----------------+----------------++
 -4.0             -2.0              0.0              2.0              4.0
```

Figure 17. Category probability curves for Invites Others to Make Commitments for collapsed scale.

Chapter 5: Discussion

Reflections on Findings for Each Research Question

Research Question 1a: Raters' usage of controlled rating condition. This study provides evidence that when raters have the ability to manipulate the recordings as they observe a ratee they use it more often than not. Few studies have documented rater behavior in regards to how they manipulate video recordings. Ryan et al. (1995) found that the controlled observational group (CRC) paused (0-45 times) more than they rewound (0-13 times). Raters also indicated they found pausing more helpful than rewinding. Rating behavior for the MTA differed from rating behavior found in the Ryan et al. study. Raters rewound (0-8 times) far more frequently than they paused (0-2 times) and they believed that the rewind function was much more useful than the pause function.

From comparing the two studies, it appears that differences in a performance assessment may have an effect on the reviewing behavior of raters. Although the researcher does not know the exact cause of the differences in rating behaviors, there are a few potential causes worth noting. There were many differences between the two performance assessments. The performance assessment in the Ryan et al. study was a group-discussion exercise that is common in assessment centers. Because of the group discussion, Ryan et al. stated that there was a lot of noise that may have kept raters from focusing on the ratee. The noise may have caused more cognitive overload thus causing the raters to need to pause to record observations. In the current study, only one investigator and missionary were a part of each performance. There was likely much less noise in this study thus causing raters to not have the same cognitive demands that raters had in the Ryan et al. study.

Differences in rating behaviors may be indicative of differences in cognitive demands in rating PAs. This study was purposely simplified to make it less cognitively demanding on the raters. Missionaries typically teach in pairs and they take turns teaching. For this study, missionaries taught without their companions in order to give each missionary the opportunity to teach uninterrupted and to display the necessary teaching skills. The raters could rate without being negatively or positively influenced by another missionary. The number of criteria included in this study was also reduced from nine to five to ease the cognitive demands on the raters. Future studies should explore how rating behaviors change across different PA's and how the rating behaviors are connected to the cognitive demands of the rating situation.

Another major difference between the two studies was the experience of the raters. The raters in the Ryan et al. study were undergraduate introductory psychology students. The students rated six different criteria. They participated in a 1.5 hour training where the criteria were defined and they were given the opportunity to practice rating but it does not appear that the students had ever had experience rating others based on these criteria prior to the training. The raters in this dissertation had all had ample experience with the criteria. The raters learned and applied the criteria as they served as missionaries. As teachers, they taught them to other missionaries and gave the missionaries formative feedback on how well they applied them on a regular basis. The more experience an individual has with performing a particular task the less cognitive resources they need to use to perform that task (Best, 1992). Although many of the MTC raters did not have formal experience with using the MTA, they were very experienced with the criteria and with informally assessing missionaries on their performance of the criteria. Perhaps the rating task was less cognitively taxing on the MTC raters because of their experience thus causing them to need to manipulate the videos far less frequently.

Research Question 1b: Raters' reasons for reviewing recordings. The most common reasons raters backed up and reviewed portions of the recordings were to try to hear something that was said that they could not hear, to review something they missed due to being unfocused or distracted, not having enough information to make a rating decision, glitches, something that was said that was unexpected or confusing, something funny or interesting, and interruptions. Raters from the Ryan et al. (1995) study differed in the reasons they gave for pausing and rewinding. The study reported that the purpose for pausing was to give them more time to record observations and the purpose for rewinding was to observe something they may have missed or to ensure that they had not missed anything. Again, differences in reviewing behavior may be due to differences in the cognitive load of the rating task and/or the experience of the raters.

Issues such as the volume or reviewing something that was funny are obviously not associated with the rating task being too cognitively demanding. When raters did review segments because they did not fully capture something or to write down notes, it is difficult to determine if the raters used this capability because rating was too cognitively demanding or if they used it because they knew they could review the video so they failed to focus on all pertinent behaviors the first time through. Raters admitted that they had a tendency to tune out more frequently when they knew they could review the recording. The URC kept raters constantly focused because they knew they only had one opportunity to watch the recording.

Because volume was the most common issue causing raters to review segments, the MTC should consider resolving this problem when recording missionary teaching samples in the future. The microphones used in this study were placed on the opposite side of the room from where the missionaries were sitting. This would seem to make sense as they were attached to the camera. The MTC should consider placing the microphones closer to the missionaries. Also, the MTC could give instruction to the missionaries to speak loudly and clearly so that what they say can be captured and understood. Volume is an issue that all institutions implementing PAs where video recordings are used should take into consideration. If the rater cannot hear what the ratee is saying, then they cannot give them an accurate or reliable rating.

Affectively, raters appreciated having the ability to manipulate the recordings. This finding is in agreement with findings from other studies. Ryan et al. (1995) cited a study conducted by Lepard et al. (1990) where raters observed examinees through direct observation or through a videotape where they were allowed one rewind. Those who used the videotape reported less fatigue, less stress, and more confidence in the accuracy of the behaviors they recorded. The raters in the current study also shared similar feelings in their exit interviews. Although they did not always manipulate the recordings, raters appreciated having it available in case they needed it. This may be something to consider for other PA creators and users.

Research Question 1c: Time spent reviewing recordings. On average, raters spent an additional 1 minute 45 seconds reviewing the recordings when they were able to control them. This is not an exorbitant amount of time in terms of the additional cost incurred from allowing raters to review the recordings. Few, if any, studies have sought to know how much additional time manipulating recordings adds on to the rating process. This question was important because the researcher was interested in knowing the costs and benefits of allowing raters to control the recordings. If reviewing the recordings doubled the time it took to rate, using the CRC would become infeasible. Therefore, according to this study, additional costs in terms of rater time should not be a huge factor when institutions are deciding whether or not to use the CRC.

Research Question 1d: Changes in reviewing behavior across criteria. For the most part, there was not much of a difference in viewing behavior from one rating criterion to the next. Most rewinds were not associated with any particular criteria. Raters reviewed segments throughout the teaching performances because they were unable to hear, confused, distracted, or they experienced a glitch in the recording. When a rater did review a segment that was directly connected to one of the criterion, it tended to be one of the more objective or analytic criteria which included Asks Questions, Adjusts to Needs, and Invites Others to Make Commitments.

Perhaps the CRC is more suitable for PAs where the criteria are more analytic. The URC may be more suitable for holistic criteria. The G-study from Research Question 2c gives partial support to this hypothesis. The G-study found that ratings from the Shows Warmth and Concern criterion were more reliable when raters used the URC than the CRC. Unfortunately, the data from MFRM analysis in this same research question contradict this hypothesis. The ratings for the Listens criterion were more reliable when raters used the CRC. Further research would need to be conducted to understand whether or not ratings for holistic/subjective and analytic/objective criteria are affected differently by the CRC and URC.

Research Question 2: Variability attributable to each source of variance. In the G-study the majority of the variation among ratings came from actual differences among the missionaries which is desirable. An average of 48% of the variation came from the missionaries in the CRC across the five criteria and an average of 57% came from the missionaries in the URC. The raters contributed to 13% of the variance for the CRC and 11% for the URC. The interaction between missionaries and raters in addition to any unmeasured or unsystematic error contributed to 39% of the variance for the CRC and 33% for the URC. Because of the nested and unbalanced design of the dataset, the researcher was not able to include both rating conditions in the same G-study and was therefore not able to determine the percent of the variance attributable to this facet.

The MFRM analysis confirmed the results from the G-study. It revealed that there was a high level of variance among missionaries. The missionaries varied from each other in their ability levels and the raters were able to distinguish those differences. The variance for raters was higher than desired. There were statistically significant differences among the severity/leniency levels of the raters.

Reliability is always one of the main concerns test makers must focus on when creating a high quality PA. The MTA did a good job of discriminating among the missionaries but MTA administrators should look at reducing the measurement error that comes from raters and the interaction between raters and missionaries. Administrators could do this by adding more raters and/or providing additional training (including on-the-job practice with feedback) for the raters. Data from the MFRM can be used to analyze individual-level statistics. It would be advantageous for MTA administrators to analyze future ratings using MFRM to determine how the raters are performing and which ones may need additional training.

Research Question 2a: Reliability across criteria. Variance existed among the phi coefficients of the five criteria. The Asks Questions criterion had the lowest reliability with an average phi coefficient of .59. Invites Others to Make Commitments had the highest reliability with a phi coefficient of .83. This criterion was also the only one to differ statistically significantly from the other criteria at an alpha level of .05. The other criteria all had reliabilities that were comparable to one another.

The MFRM separation reliability index of .95 for criteria as well as the statistically significant fixed chi-square indicated that there were definite differences among the difficulty levels of the criteria. Invites Others to Make Commitments was the most difficult and Shows Warmth and Concern was the least.

When the five criteria were analyzed separately using MFRM, missionary separation reliability indices were obtained for each of them. This analysis revealed different conclusions about differences in the reliability among the criteria than the G-study. It found Adjusts to Needs to have the lowest missionary separation reliability and Invites Others to Make Commitments to have the highest. The reliability for Invites Others to Make Commitments differed significantly only from the Listens and Adjusts to Needs criteria. Additionally, the Adjusts to Needs criteria was significantly lower than the other four criteria. Test makers should consider revising this scale to increase its ability to discriminate among the various ability levels of missionaries.

If G theory and MFRM both measure reliability, why would they produce differing results? A MFRM separation reliability statistic for missionaries controls for all variance that comes from any other facet such as raters, rating conditions, etc. The phi coefficient is a measure of the reliability of an assessment that includes all sources of variation. This is the most likely cause of the differences between the two methods.

Examining the reliability of the MTA by criteria level helps MTA administrators focus in on the criteria that are contributing the most to poor reliability. They may want to consider taking measures to increase the reliability of the criteria that performed poorly. They can do this by revising the scales and/or providing better rater training to help raters rate more consistently.

Research Question 2b: Impact of number of raters on reliability. The D-study revealed that the reliability increased as additional raters were added to the model. There is no predefined criterion for how high a phi coefficient should be in order to produce sufficiently reliable ratings. It is up to test creators and implementers to determine the level of reliability they are willing to accept. Part of the decision is also dependent on how costly it is to achieve a particular level of reliability. The MTA could have a phi coefficient of approximately .80 across all five criteria if they were willing to have seven raters rate each missionary using the CRC and six raters using the URC. Six to seven raters rating each missionary would be far too costly for the MTC. In consulting with them, the MTC has indicated that they would ideally like to use no more than two raters to rate each missionary. Using two raters for each missionary would produce phi coefficients above .70 for only two to three of the five criteria depending on the rating condition used. In order to produce higher phi coefficients across all five criteria, MTA administrators need to consider other alternatives besides adding more raters. Again, alternatives include improving the rating scales and rater training.

Research Question 2c: Reliability across rating conditions. A past study on the use of the CRC found that it only had a marginal impact that was neither large nor consistent on the observational accuracy of the ratings when compared to the URC (Ryan et al., 1995). This study looked at the reliability of the ratings as opposed to accuracy. The G-study determined that there was not a statistically significant difference between the reliabilities of the two rating conditions when a weighted mean was computed across all five rating criteria. When the two rating conditions were compared within each criterion, only the Shows Warmth and Concern showed a statistically significant difference between the rating conditions. The URC produced more reliable results than the CRC.

The MFRM analysis also indicated that there was not a significant difference between the two rating conditions across all five criteria. However, when each criterion was considered separately, the missionary reliability for rating conditions proved to be statistically significantly different for the Listens criterion. In this case, the CRC provided more reliable ratings. Again, this difference in the results of the G-study and MFRM analysis may be caused by the fact that MFRM controls for variance from all other facets.

It is unclear why the Shows Warmth and Concern and the Listens criteria had significant differences between the two rating conditions. As suggested in the above discussion on Research Question 1d, the Shows Warmth and Concern criterion may better lend itself to a more holistic rating method. Ratings based on an overall impression may be more reliable than those based on distinct behaviors. The only problem with this hypothesis is that the Listens criterion is subjective as well but raters using the CRC produced more reliable ratings. Future studies should analyze whether certain criterion are better analyzed using one rating condition over the other and if so, what characteristics make it so.

Ryan et al. (1995) did not find that the CRC had a significant impact on the accuracy of ratings. Just as the rating conditions in this study produced statistically significant differences in the reliability of some criteria and not others, perhaps Ryan et al. may have found differences in accuracy between the two rating conditions if different criteria had been used.

Research Question 2d: Impact of the use of the CRC on reliability. Overall, no patterns were found between how a rater used the digital recordings and the reliability of their ratings. All raters proved to have sufficiently reliable ratings with missionary separation reliabilities ranging from .75 to .91. The missionary reliability from Rater 2, who had the lowest reliability, was the only one to differ significantly from the reliabilities of the other raters. The correlation between the reliability of the ratings for each rater and the frequency with which they manipulated the recordings was only -.14. Therefore, the researcher cannot conclude that there was a relationship between these two variables.

Few, if any, studies have sought to analyze the relationship between rating behavior and reliability. This study does not provide any evidence that manipulating video recordings one way is more effective than another. Raters were not given any instruction concerning how they should manipulate the recordings or what kinds of manipulations were effective and which were not. Future research should examine whether certain types of manipulating behavior produces better ratings. Perhaps the CRC would significantly increase reliability if raters were trained how to implement it properly. Perhaps the raters should also be given some training on how to best take advantage of the ability to stop and rewind the recording. When should they be encouraged to use it and for what purposes?

Research Question 3: Performance of rating scale categories. The rating scales for all five of the criteria contained disordering among the categories. All the scales with the exception of the Adjusts to Needs scale had disordering between categories 6 and 7. This indicates that raters had a difficult time deciphering between these two categories. The Adjusts to Needs scale had a disproportionate number of ratings toward the upper end of the scale and the Invites Others to Make Commitments scale showed a central tendency effect. It is unclear whether or not these rater effects were caused by the majority of the missionaries actually displaying those levels of behavior or if the scales were not well enough defined to allow raters to make distinctions among the missionaries.

MTA administrators should consider clarifying the meaning of the statements in the rating scales and consider collapsing some of the categories to resolve problems with disordering. The Adjusts to Needs scale should be revised in order to aid the raters in discriminating among the various missionaries.

Comparison of G Theory and MFRM

Both of these two methods of analyzing reliability had their strengths and weaknesses in analyzing the reliability of the MTA. In agreement with other studies that have compared G theory and MFRM (Alharby, 2006; Lynch & McNamara, 1998; Smith & Kulikowich, 2004; Sudweeks et al., 2004), the researcher believes that both methods contributed insight and understanding into the reliability of the MTA. It was valuable to have the output from both studies. The G-study was useful because it showed how much each facet contributed to the total variance. It was helpful to see how much measurement error there was relative to the true score. MFRM allowed the researcher to determine that there was too much variance among the raters, but it did not allow her to get a picture of how much measurement error raters contributed relative to the other facets. Understanding the relative sizes of the measurement error contributed by each facet helped to determine which facets should be focused on first in order to improve the reliability of the MTA. Another advantage of G theory was the fact that a D-study could be conducted. The D-study allowed the researcher to determine how the reliability would be affected by adding raters. MFRM does not have this functionality.

Because of the unbalanced study design as well as the missing data, it was very difficult to analyze the data using a G-study. G-studies cannot handle either of these issues. MFRM is able to handle abnormalities in the data set much better than a G-study. MFRM just requires that all the data be connected. The output from the MFRM provided a wealth of detail that could not be obtained from the G-study. The researcher was able to analyze how the individual raters and missionaries performed. This information will be especially useful for the MTC as they implement interventions to increase the reliability of the MTA. They can determine which raters may need further training.

The researcher found it useful to conduct the G-study first to get an overview of how the MTA was performing and then conduct the MFRM in order to understand more of the details. The researcher would recommend that other researchers also use both methods when analyzing reliability in order to analyze reliability from different vantage points.

Implications for Microteaching

Few studies on microteaching have explored their reliability. Although they are used mainly for formative purposes, it would still be insightful to assess how reliable peers and supervisors are in giving feedback. Are some supervisors more severe than others? Are some supervisors more critical with some students than others? Do supervisors or peers award more reliable ratings? Research into this topic would improve the quality of microteaching experiences.

Limitations

This study was conducted in an environment that is foreign to many teacher education programs. Missionaries are required to teach quite differently than typical classroom teachers. They teach in small group settings. Generally, they teach individuals or families in their homes and not in a formal classroom. Missionaries must engage on a very personal level with those taught. They must know an investigator's fears, doubts, belief system, lifestyle, etc. The content of what they teach is of a spiritual nature and is taught through the heart and spirit and not just the mind. A missionary's ultimate purpose in teaching is not just to increase cognitive knowledge, but also to change individuals' affective characteristics, beliefs, and behaviors. Missionaries are to help investigators not only cognitively understand the content but also to believe and embrace it. Their teaching style is more similar to tutoring than to classroom instruction. All of these teaching differences should be taken into account when generalizing the findings from this study. Similar studies should be conducted in other performance assessment and teacher education settings to determine if these results are transferrable.

Recommendations for Future Research

This study was exploratory. Not much research has been conducted in the realm of rater behavior in regards to the CRC. Because this study was exploratory, the data gave a picture of what happens when raters used the CRC and how this affects reliability but future studies should now delve deeper into why raters behave the way they do when using the controlled rating condition and why it impacts reliability the way it does.

How do rating behaviors change with different PAs? There was a difference between this study and the Ryan et al. study in how raters manipulated the recordings. It would be interesting to study how and why rating behaviors change when different PAs are being administered. Is the rating behavior affected by the characteristics of the rater or the nature of the criteria being assessed? Does the CRC contribute more to the reliability of the ratings when the assessment is more cognitively demanding? For instance, the CRC may have more of an impact on ratings if raters were required to assess more complex rating criteria and/or rate a larger number of criteria simultaneously. It would also be important to know if there were some rating behaviors that would produce more accurate and reliable ratings than others. If so, raters should be taught how to effectively manipulate the recordings while rating.

One question that was not assessed in this study was whether or not the CRC has more of an impact when a rater is fatigued. In their exit interviews, many of the raters reported manipulating the recordings more often after rating for a significant amount of time because they became fatigued. It would be insightful to analyze ratings toward the end of a rater's schedule to determine if there was any impact on the ratings.

How would the reliability of the ratings be affected if a behavioral checklist was used as opposed to rating scales? The CRC may have more of a positive impact on ratings if it was more important for raters to identify distinct behaviors. This is what occurs when a behavioral checklist is used.

In order for the MTC to gain a full understanding of the reliability of the MTA, further tests should be conducted with different tasks, different teaching occasions, and different rating occasions. The MTC should also conduct studies looking into the validity of the instrument.

References

Alharby, E. R. (2006). *A comparison between two scoring methods, holistic vs. analytic, using two measurement models, the generalizability theory and the many-facet Rasch measurement, within the context of performance assessment*. Retrieved March 20, 2010, from Dissertations & Theses: Full Text. (AAT 3236860)

Allen , D. W., & Ryan, K. (1969). *Microteaching*. Reading, MA: Addison-Wesley.

Allen , D. W., & Wang, W. (2002). Microteaching. In J. W. Guthrie (Ed.), *Encyclopedia of education* (pp. 1620-1623). New York: Macmillan Reference USA.

Arter, J. (1999). Teaching about performance assessment. *Educational Measurement: Issues and Practice, 18*(2), 30-44.

Arvey, R. D., & Murphy, K. R. (1998). Performance evaluation in work settings. *Annual Review of Psychology*, *49*, 141-168.

Benton-Kupper, J. (2001). The microteaching experience: Student perspectives. *Education*, *121*, 830-835.

Best, J. B. (1992). *Cognitive psychology* (3rd ed.). St. Paul, MN: West Publishing Company.

Bolton, C. (1996). *Preservice teachers' sense of efficacy and the influence of performance assessment*. Aiken: University of South Carolina at Aiken. (ERIC Document Reproduction Service No. ED406366)

Braden, J. P. (2005). Performance-based assessment. In S. W. Lee (Ed.), *Encyclopedia of school psychology* (pp. 380-381). Thousand Oaks, CA: Sage Publications. Retrieved February 15, 2010, from Gale Virtual Reference Library.

Camtasia Studio (version 7.0.0) [computer software]. (2010). Okemos, MI: TechSmith.

Chiu, C. W. T. (1999). *Scoring performance assessments based on* judgments: *Utilizing meta-analysis to estimate variance components in generalizability theory for unbalanced situations*. Ph.D. dissertation, Michigan State University, East Lansing. Retrieved September 28, 2010, from Dissertations & Theses: Full Text. (AAT 9948082)

Chiu, C. W. T., & Wolfe, E. W. (1997). *Generalizability theory: A new approach to analyze non-crossed performance assessment data*. Paper presented at the annual meeting of the American Educational Research Association. Chicago, IL.

Chiu, C. W. T., & Wolfe, E. W. (2002). A method for analyzing sparse data matrices in the generalizability theory framework. *Applied Psychological Measurement*, *26*, 321.

Crick, J. E. & Brennan, R. L. (1984). GENOVA: a general purpose analysis of variance system (Version 2.2) [Computer software]. Iowa City, IA: American College Testing Program.

Cruickshank, D. R., & Metcalf, K. K. (1993). Improving preservice teacher assessment through on-campus laboratory experiences. *Theory into Practice*, *32*(2), 86-92.

DeCarlo, L. T. (2005). A model of rater behavior in essay grading based on signal detection theory. *Journal of Educational Measurement*, *42*, 53-76.

Delandshere, G., & Petrosky, A. R. (1998). Assessment of complex performances: Limitations of key measurement assumptions. *Educational Researcher*, *27*(2), 14-24.

DeNisi, A. S. (1996). *Cognitive approach to performance appraisal*. New York: Routledge.

Feldman, J. (1981). Beyond attribution theory: Cognitive processes in performance appraisal. *Journal of Applied Psychology*, *66*, 127-148.

Feldt, L. S. (1969). A test of the hypothesis that Cronbach's alpha or Kuder-Richardson coefficient twenty is the same for two tests. *Psychometrika*, *34*, 363-373.

Haertel, E. (1988). Assessing the teaching function. *Applied Measurement in Education*, *1*, 99-107.

Haertel, E. H. (1991). New forms of teacher assessment. *Review of Research in Education*, *17*, 3-29.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. New York: Academic Press.

Joe, J. N. (2009). *Using verbal reports to explore rater perceptual processes in scoring: An application to oral communication assessment*. Retrieved March 20, 2010, from Dissertations & Theses: Full Text. (AAT 3338640)

Kallenbach, W. W., & Gall, M. D. (1969). Microteaching versus conventional methods in training elementary intern teachers. *Journal of Educational Research*, *63*(3), 136-141.

Kolk, N. J., Born, M. P., van der Flier, H., & Olman, J. M. (2002). Assessment center procedures: Cognitive load during the observation phase. *International Journal of Selection & Assessment*, *10*, 271-278.

Kpanja, E. (2001). A study of the effects of video tape recording in microteaching training. *British Journal of Educational Technology*, *32*, 483-486.

Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions*, *16*, 878.

Linacre, J. M. (2010). Facets (version 3.67.0) [Computer software]. Chicago, IL: Winsteps.com.

Linacre, J. M., & Wright, B. D. (2002). Construction of measures from many-facet data. *Journal of Applied Measurement*, *3*, 486-512.

Lumley, T. (2002). Assessment criteria in a large-scale writing test: What do they really mean to the raters? *Language Testing*, *19*, 246-276.

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*, 158-180.

Madaus, G. F., & O'Dwyer, L. M. (1999). A short history of performance assessment. *Phi Delta Kappan*, *80*, 688-695.

McKnight, P. C. (1971). Microteaching in teacher training: A review of research. *Research in Education*, *6*, 24-38.

Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement*, *4*, 386-422.

Orr, M. (2002). The FCE speaking test: Using rater reports to help interpret test scores. *System*, *30*, 143-154.

Palm, T. (2008). Performance assessment and authentic assessment: A conceptual analysis of the literature. *Practical Assessment, Research & Evaluation*, *13*(4), 1-11.

Pecheone, R. L., & Chung, R. R. (2006). Evidence in teacher education: The Assessment for California Teachers (PACT). *Journal of Teacher Education*, *57*(1), 22-36.

Peterson, T. L. (1973). Microteaching in the preservice education of teachers: Time for a reexamination. *The Journal of Educational Research*, *67*(1), 34-36.

Popham, W. J. (1973). *Alternative teacher assessment strategies*. Retrieved March 20, 2010, from ERIC database.

Popham, W. (1990). *Modern educational measurement: A practitioner's perspective*. Englewood Cliffs, NJ: Prentice-Hall.

Preach My Gospel (2004). Salt Lake City, UT: The Church of Jesus Christ of Latter-day Saints.

Ryan, T. G. (2006). Performance assessment: Critics, criticism, and controversy. *International Journal of Testing*, *6*(1), 97-104.

Ryan, A. M., Daum, D., Bauman, T., Grisez, M., Mattimore, K., Nalodka, T., et al. (1995). Direct, indirect, and controlled observation and rating accuracy. *Journal of Applied Psychology*, *80*, 664-670.

Sanchez, J. I., & De la Torre, P. (1996). A second look at the relationship between rating and behavioral accuracy in performance appraisal. *Journal of Applied Psychology*, *81*(1), 3-10.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Newbury Park, CA: Sage.

Shore, B. M. (1972). *Microteaching: A brief review*. Retrieved March 10, 2010, from ERIC database.

Smith, Jr., E. V., & Kulikowich, J. M. (2004). An application of generalizability theory and many-facet Rasch measurement using a complex problem-solving skills assessment. *Educational & Psychological Measurement*, *64*, 617-639.

Stiggins, R. J. (1994). *Student centered classroom assessment*. Toronto, Canada: Maxwell Macmillan Canada.

Stiggins, R., Tziriel, D., Pellegrino, J. W., Silver, E. A., Herman, J. L., & Zuniga, S. A. (2003). Assessment. In J. W. Guthrie (Ed.), *Encyclopedia of education* (2nd ed., pp. 123-139). New York: Macmillan Reference USA.

Sudweeks, R. R., Reeve, S., & Bradshaw, W. S. (2004). A comparison of generalizability theory and many-facet Rasch measurement in an analysis of college sophomore writing. *Assessing Writing*, *9*, 239-261.

Suto, W. M. I., Greatorex, J. (2008). A quantitative analysis of cognitive strategy usage in the marking of two GCSE examinations. *Assessment in Education: Principles, Policy, and Practice*, *15*, 73-89.

Trent-Wilson, V. (1990). *The effects of a microteaching program upon the critical thinking skills of preservice teachers*. Retrieved March 20, 2010, from Dissertations & Theses: Full Text. (AAT 9030647)

Waxman, H. C. (2003). Classroom observation. In J. W. Guthrie (Ed.), *Encyclopedia of education* (pp. 303-310). New York: Macmillan Reference USA.

Wolfe, E. W., & Feltovich, B. (1994). *Learning to rate essays: A study of scorer cognition*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA. Retrieved March 10, 2010, from ERIC database.

Zenisky, A. L. (2007). Performance-based assessment. In N. J. Salkind (Ed.), *Encyclopedia of measurement and statistics* (Vol. 2, pp. 757-760). Thousand Oaks, CA: Sage Reference. Retrieved March 20, 2010, from Gale Virtual Reference Library.

Appendix A

Missionary Teaching Assessment Rating Scale

	0	1	2	3	4	5	6	7	**Score**
Shows Warmth and Concern	No basis for judgment	Behaves in a disrespectful or disinterested manner	Shows skills in between levels 1 & 3	Respectful and polite but distant or detached	Shows skills in between levels 3 & 5	Shows interest and concern; respects agency and beliefs	Shows skills in between levels 5 & 7	Warm and friendly; sincere interest and; words and actions show compassion	
Listens	No basis for judgment	Ignores, interrupts, or fails to listen to investigator	Shows skills in between levels 1 & 3	Attentive but doesn't demonstrate understanding or importance of investigator's thoughts	Shows skills in between levels 3 & 5	Focuses on investigator comments and non-verbal communication; demonstrates	Shows skills in between levels 5 & 7	Listens attentively; responds to non-verbal communication; seeks clarity if needed; investigator's feelings important	
Asks Questions	No basis for judgment	No questions to discover needs and interests	Shows skills in between levels 1 & 3	Missionary asks uncomfortable, irrelevant, or difficult questions; unsuccessful at discovering needs and interests	Shows skills in between levels 3 & 5	Clear questions that help identify investigator needs and interests	Shows skills in between levels 5 & 7	Simple, thought-provoking questions that allow reflection and motivate investigator to express thoughts and feelings; follow-up questions used when needed	
Adjusts to Needs	No basis for judgment	Focuses on lesson, not investigator; doesn't address needs or interests.	Shows skills in between levels 1 & 3	Attempts to address needs but doesn't understand or respond in helpful ways	Shows skills in between levels 3 & 5	Shows basic understanding of needs; adjusts teaching but adjustments are too limited or too extensive	Shows skills in between levels 5 & 7	Understands needs; content, pace, and sequence of teaching consistent with needs	
Invites Others to Make Commitments	No basis for judgment	Doesn't extend commitment invitations	Shows skills in between levels 1 & 3	Invitations tentative, unclear, pushy or untimely	Shows skills in between levels 3 & 5	Invitations clear, direct, and appropriate	Shows skills in between levels 5 & 7	Invitations tailored to help investigator move toward conversion; expresses follow-up plans and offers support	

Figure A1. Missionary Teaching Assessment rating scale for five criteria included in the study.

	0	1	2	3	4	5	6	7	Score
Begins the Lesson	No basis for judgment	Makes no effort to prepare investigator (e.g., build trust, explain purpose or share expectations); launches immediately into the lesson	Shows skills in between levels 1 & 3	Launches into lesson with minimal effort to learn about investigator; does not use an introduction; beginning segment is too shallow or too extensive	Shows skills in between levels 3 & 5	Asks simple questions about investigator's background; uses an introduction to the message such as those listed *PMG (*pp. 176-177)	Shows skills in between levels 5 & 7	Begins in a warm, respectful manner; genuinely seeks to understand the background of the investigator; discusses mutual expectations; uses an introduction that fits the circumstances	
Teaches for Understanding	No basis for judgment	Message is unclear or includes incorrect doctrine	Shows skills in between levels 1 & 3	Message is too complex or disjointed; rambles or teaches doctrine unrelated to investigator needs; uses unfamiliar religious terms without defining them	Shows skills in between levels 3 & 5	Message follows a logical pattern; attempts to present doctrine based on the needs of the investigator; teaches correct doctrine; usually defines unfamiliar religious terms	Shows skills in between levels 5 & 7	Message follows a logical pattern and is taught in a clear and concise manner; teaches doctrine relevant to investigator's needs; provides definitions for unfamiliar religious terms; checks investigator's understanding	
Uses Scriptures	No basis for judgment	Does not use the scriptures	Shows skills in between levels 1 & 3	Reads or briefly refers to scriptures without providing context or application; use of scriptures does not contribute meaningfully to the lesson	Shows skills in between levels 3 & 5	Gives context and reads scriptures; use of scriptures contributes to understanding the message	Shows skills in between levels 5 & 7	Gives background and context; reads and explains scriptures; helps the investigator understand and apply principles to bring about conversion	
Testifies	No basis for judgment	Does not testify	Shows skills in between levels 1 & 3	Testifies too infrequently or in a way that is mechanical, irrelevant, or repetitive	Shows skills in between levels 3 & 5	Frequently testifies (i.e. shares simple, direct declarations of personal belief) in a sincere and believable manner	Shows skills in between levels 5 & 7	Frequently testifies in heartfelt, convincing, and personalized manner (e.g. shares appropriate, faith-promoting personal experiences)	

Figure A2. Missionary Teaching Assessment rating scale for four criteria not included in the study.

Appendix B

Informed Consent Form for Raters

The Effect of Raters and Rating Conditions on the Reliability of the Missionary Teaching Assessment
Consent to be a Research Subject

Introduction

This research study is being conducted by Abigail Ure, a doctoral candidate, at Brigham Young University to determine the reliability of the Missionary Teaching Assessment (MTA) as well has how the reliability is effected by various raters and rating conditions. You were invited to participate because of your past rating experience for the MTC Research and Evaluation department.

Procedures

If you agree to participate in this research study, the following will occur:

- You will attend a two-hour rater training.
- You will rate sixteen missionary teaching performances that will be assigned to you. Each one will take approximately 15 minutes to rate.
- You will rate all sixteen missionaries in the MTC Research and Evaluation department in the same day.
- Software will record when you perform any kind of manipulation of the digital recording such as pausing, rewinding, etc.
- You will be interviewed immediately following your rating session for approximately thirty minutes concerning your experience rating and how you utilized the ability to control the pace with which you viewed the digital recordings.
- The interview will be audio recorded to ensure accuracy in reporting your statements.
- The interview will take place in a private room located in the MTC.
- The researcher may contact you later to clarify your interview answers for approximately fifteen minutes.
- The total time commitment will be approximately six and one half hours.

Risks/Discomforts

There are minimal risks for participation in this study. However, you may feel mentally and/or physically fatigued from the length of time that is required for you perform the ratings. To minimize the fatigue, you are welcome to stretch, get a snack, or use the restroom at your discretion during your rating session.

You may also feel uncomfortable with someone analyzing your ratings. The purpose of the study is not to draw conclusions about your effectiveness as a rater, but rather to gather information about the effectiveness of the MTA. The data resulting from your ratings will be used to refine how the MTA is administered and rated in the future.

Benefits

There will be no direct benefits to you. However, it is hoped that through your participation researchers will learn more about the Missionary Teaching Assessment and will be able to improve the instrument.

Confidentiality

Strict confidentiality will be maintained. No individual identifying information will be disclosed. All data collected in this research study will be stored in a secure area and access will only be given to personnel associated with the study.

Compensation

You will be compensated your regular MTC hourly wage for all time spent in training, rating, or being interviewed.

Participation

Participation in this research study is voluntary. You have the right to withdraw at anytime or refuse to participate entirely without affecting your employment or standing at the MTC.

Questions about the Research

If you have questions regarding this study, you may contact:

Abigail Ure
6071 Village Bend Dr. #212
Dallas, TX 75206
801-372-7119
Abbey.ure@gmail.com

Questions about your Rights as Research Participants

If you have questions regarding your rights as a research participant, you may contact:

BYU IRB Administrator
A-285 ASB, Brigham Young University
Provo, UT 84602
801-422-1461
irb@byu.edu

I have read, understood, and received a copy of the above consent and desire of my own free will to participate in this study.

Signature:____Date:_____________

Printed Name:____

Appendix C

Protocol for Rater Think-Alouds

Rater Think-aloud Protocol

INTERVIEWER RESPONSIBILITIES AND INSTRUCTIONS
Important: Please make sure that the participant has a signed informed consent form on file. If not, he or she will need to complete the form BEFORE beginning the interview.

As the interviewer, your primary responsibility is to direct the think-aloud process and then ask the additional questions. Aside from the occasional probing you, the interviewer, should interject as little as possible in this process. Please try to refrain from leading the participant in any way.

You will have a digital camcorder. Please make sure the camcorder is working properly before you begin the interview. It is recommended that you take notes while the participant is talking. Jot down notes about the environment and what the respondent is saying and doing. After the interview is complete, take about 5 minutes to reflect (e.g., note your overall impressions).

INTERVIEWER SCRIPT
{Read the following to the participant.] Thank you for agreeing to participate in this study. Your responses will help us understand your experiences as a rater. How you respond and what you share with me will remain confidential in that your name and any other identifying information will not be linked together.

This exercise should take no longer than 15 minutes. If at any time you feel like taking a break or wish to not continue, you may do so without penalty.

We will watch the teaching performance you just watched and rated. While you are watching the recordings, I would like for you to talk freely about what you noticed during the initial observation and how you arrived at the particular scores you gave them. Please don't feel as though you have to filter your comments. The purpose of this exercise is to capture as much of the thoughts you originally had as you watched and rated the performance. When we reach the points where you paused or reviewed the video, I would like you to freely explain why you did so. Again, the things you share will remain confidential. You may stop and replay parts if you want to expand on your thought process.

Do you have any questions?

[Play teaching performance recording]

Appendix D

Exit Interview Questions

Rater Exit Interview Protocol

You will have a digital camcorder. Please make sure the camcorder is working properly before you begin the interview. It is recommended that you take notes while the participant is talking. Jot down notes about the environment and what the respondent is saying and doing. After the interview is complete, take about 5 minutes to reflect (e.g., note your overall impressions).

INTERVIEWER SCRIPT
{Read the following to the participant.] Thanks again for agreeing to participate in this study. Your responses will help us understand your experiences as a rater. How you respond and what you share with me will remain confidential in that your name and any other identifying information will not be linked together.

This exercise should take no longer than 30 minutes. If at any time you feel like taking a break or wish to not continue, you may do so without penalty.

I just have a few questions for you about the rating process.

1. On a scale from 1 to 5, 1 being *not at all helpful* and 5 being *very helpful*, how much did rewinding help you make a better rating decision?

2. Do you have any additional information that would help me understand why you replayed the video during the rating process?

3. On a scale from 1 to 5, 1 being *not at all helpful* and 5 being *very helpful*, how much did pausing help you make a better rating decision?

4. Do you have any additional information that would help me understand why you paused the video during the rating process?

5. Which rating condition did you prefer more—the controlled or uncontrolled access?

6. Why?

7. How did the controlled rating condition affect the difficulty of the rating process?

8. How did the controlled rating condition affect the ratings you arrived at?

9. How would you rank the 5 criteria in order of easiest to most difficult to rate?

Thank you for taking the time to participate in the study today. Do you have any questions for me?

After the interview is complete, take about 5 minutes to reflect (e.g., note your overall impressions).