The Pennsylvania State University
The Graduate School
College of Education

ORAL PERFORMACE SCORING USING GENERALIZABILITY THEORY AND MANY-FACET RASCH MEASUREMENT: A COMPARISON STUDY

A Dissertation in
Educational Psychology
by
Saif F. Alkahtani

© 2012 Saif F. Alkahtani

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

August 2012

The dissertation of Saif F. Alkahtani was reviewed and approved* by the following:

Jonna M. Kulikowich
Professor of Education
Dissertation Adviser
Chair of Committee

Hoi K. Suen
Distinguished Professor of Educational Psychology

Pui-Wa Lei
Associate Professor of Education

Edgar Yoder
Professor of Agricultural Economics and Rural Sociology

Rayne Sperling
Associate Professor of Education
Graduate Program Coordinator

*Signatures are on file in the Graduate School.

ABSTRACT

The principal aim of the present study was to better guide the Quranic recitation appraisal practice by presenting an application of Generalizability theory and Many-facet Rasch Measurement Model for assessing the dependability and fit of two suggested rubrics. Recitations of 93 students were rated holistically and analytically by 3 independent raters for their implementation of Quranic rules and proficiency of reading.

Although, the relationship of raw scores of holistic and analytic revealed high estimates, suggesting that, on average, rank ordering of students was consistent across the scoring rubrics, a paired-sample t-test revealed statistically significant differences between their means. Furthermore, individual and overall comparisons of holistic and analytic scoring rubrics of Quranic recitations using MFRM showed that analytic scoring rubric is associated with better individual and overall fit statistics for all measurement facets. Likewise, G-theory analysis showed that analytic scoring rubrics were associated with lesser measurement errors and with higher coefficients of dependability (i.e., G-coefficients, and D-indices). The introduction of analytic rubrics might have helped guide raters to evaluate students' recitations more consistently and bring raters to a common understanding of the scoring scales. Such findings might lend more support to the introduction of analytic scoring to the Quranic assessment practice, as it guides raters to rate consistently and similarly.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENT

All praise is due to Allah, and Allah's Peace and Blessings be upon His Final Messenger, his pure family, his noble companions, and all those who follow them with righteousness until the Day of Judgment. "My Lord, enable me to be grateful for Your favor which You have bestowed upon me and upon my parents and to work righteousness of which You will approve and make righteous for me my offspring. Indeed, I have repented to You, and indeed, I am of the Muslims." (Quran, Chapter 46; Verse 15).

This work is dedicated to my family; Mom, wife, brother and children: Razan, Fahad and Khalid. I'm also appreciative of all the help and support I have received from my professors, colleagues, friends and department's staff. "And if you should count the favors of Allah, you could not enumerate them. Indeed, Allah is Forgiving and Merciful" (Quran, Chapter 6, Verse 18).

Chapter One

Introduction

Teachers assign students grades based on their performance on some types of assessment tools (e.g., tests). Yet, the teachers' interest goes beyond the immediate outcome of the assessing procedures or tasks presented to the students (e.g., Kane, 2006). They rather are interested in the extent to which their students' performance on this set of items or tasks would generalize to a pool of other interchangeable sets of items or tasks not used in the assessment procedures. In theory, such inference would give a good picture of what a student can or cannot do. Presumably, teachers teach a wide content, but however; due to time restrictions they just test students on a sample of items or tasks assuming that these tasks are an excellent representation of the rest. Tests are samples of information, and based on scores important decisions are made such as students' enrollment into programs, promotion, and graduation. Yet, measurements always have error. There are theories, models, and procedures within the framework of psychometrics that are very well established to address (e.g., Lane & Stone, 2006) the degree of measurement error in a set of test scores and indicate directions for deletion of items/tasks, revision, and modification to improve reliability and validity.

These psychometric models help monitor such imperfection of measurements and keep it to the lowest level possible. The imperfections are modeled and used as an aiding component that gives great insight into the reproducibility of the scores. Measurement error may introduce fewer questions when test items are objective and students select responses as in the case of multiple-choice items. Measurement error is likely to increase with performance assessment formats due to the inclusion of raters (e.g., Rangel, 1990),

and the need for subjective evaluations of performance using rating scales or rubrics. As a consequence, there has been a great need for measurement models that extend beyond objective scoring to handle situations in which raters or judges use rating scales or rubrics as part of their assessment procedures.

With such feasibility of modeling raters' behavior, different questions concerning quality of ratings could be raised. For instance, a researcher could assess how consistent the raters are in assigning scores. Yet, a different researcher might rather assess the accuracy of using a rating scale or rubric. This study addresses the first question in the context of oral performance assessments.

The construct to be measured is oral recitation of Quran which is defined as the reading of the holy Quran they way the prophet Muhammad PBUH read it. Allah says "And recite the Quran in a slow, pleasant tone and style" (Translation of Quran chapter 73, verse 4). Such recitation is governed by a set of coherent rules called rules of Tajweed. In general, Tajweed in Arabic means to beautify, improve and master. Yet, according to the science of Quranic studies it specifically means to pronounce and articulate each letter from its right place of utterance and to give it its due length and accidental and permanent characteristics (Al-Hussary, 1999; Musri, 2005).

Purpose of the Study

The Quran subject in Saudi Arabia (SA) is assessed orally, that is the students are asked to recite one or more segments of the Holy Quran, and their recitations are judged for correct pronunciation of words, correct implementation of Tajweed rules (i.e., especial rules of reading Quranic passages), adequate speed and/or good memorization.

Therefore, an assessment that would yield reliable and valid ratings that are used for grading, or pass-fail decisions would be of great value.

Quran is a required subject of all students of elementary through high school, and according to a recent statistics on general education in Saudi Arabia for year 2009/2010, there are 4,803,966 students. Yet, Quranic recitations assessments are not limited to general education as they are also of great relevance to many post-secondary educational and non-educational institutions.

As described, oral recitation of Quranic passages requires speaking skills that are similar to constructs assessed orally such as delivering a presentation, participating in a classroom discussion or a debate. These constructs allow the students to demonstrate sufficient understanding and knowledge and adequate application in a verbal form (Nitko, 2004). Students are judged for both correctness and communication skills. However, it is more structured and specific. Every student reads the same passage and is assessed on the same dimensions.

It is more closely related to reading texts in Arabic in that they both involve coding and decoding of words, and ability to articulate words. Yet, it is also different in many respects. For instance, a high priority is placed on understanding and applying the rules of Tajweed. Explicitly, a reciter needs to pay attention not only to the letter to be read but also to those before and after as they can change the way it is pronounced. For instance, letter "ن", pronounced NOON, can be pronounced differently when it is free from minor vowels depending on the next letter. In such a case it can fall into one of four categories: 1) To be pronounced clearly called The-Making-Clear rule, 2) emerged with next letter called Merging, 3) hidden or muted called Hiding, 4) or changed into another

letter "م", pronounced MEEM called The Changing. In addition, these rules require adding a nasal sound called "Ghunna" depending on the letter that follows immediately. Furthermore, these set of rules or guidelines dictate for the sake of clarification the proper way and time of pausing and starting to convey the message(s) of Quran.

Tajweed rules are Quran-specific and need to be learned, internalized and applied properly in Quranic recitation. These coherent guidelines include for the sake of example, dimensions of vibrating or stressing on particular letters and lengthening vowel letters. These dimensions are fully elaborated and taught and assessed in evaluation of oral recitation of Quran and highly emphasized in Saudi Education.

Toward this end, different scoring schemes are being used to convert such judgments or performances into ratings. However, little is known about the efficacy of such scoring methods for Quran recitation appraisal. Although the efficacy of scoring methods has been researched for different contexts, oral context in general (Bonk & Ockey, 2003) and Quranic assessment in specific receives little attention. The rather most frequently researched context is that of the written-format (Klein et al, 1998). Nevertheless, oral performance scoring has certain defining characteristics that make it distinguishable from that of written responses. To be more explicit, in written responses, raters have the ability to place responses in piles, score questions one at a time, look again at the responses, and to create comprehensive and stable scoring-standards. On the contrary, oral performance assessments and evaluations are temporal, take place in what can be considered "real-time", and following such established guidelines as those for written responses may be unattainable.

For instance, for the constructed response format those who assign scores are encouraged to conceal the respondent's identity, mark one question at a time, give initial grading, revise, and review grades. Such guidelines are impractical or perhaps impossible to use with oral performance, hence the resulting scores or ratings are more prone to subjectivity.

To objectively score students performance, raters may use different scoring schemes (i.e., rubrics). These scoring schemes are meant to guide the rater to award ratings that represent how much of the targeted construct, or skill to be measured, the students possess. Such regulatory devices (i.e., rubrics) would assist the raters and the students as to what to look for in a set of responses and what to assess. Good rubrics can give insight into weaknesses and strengths of the teacher, students and instruction (e.g., Nitko, 2004).

Most commonly used scoring rubrics are holistic and analytic. Holistic rubrics result in one global summary index (Carr, 2000; Chi, 2001; Klein et al, 1998; Nitko, 2004). On the contrary, analytic schemes give different descriptive subscale indices that may or may not be averaged across such subscales (Carr, 2000; Chi, 2001; Klein et al, 1998; Nitko, 2004). Such rubrics have their advantages and disadvantages. Compared to analytic scoring rubrics, holistic scoring schemes have the advantage of allowing lesser time in scoring (White, 1984), and of being easier to construct, use(Linn & Gronlund, 2000) and train on which makes it of crucial importance for large scale assessments where a cost-effective means is considered (Carr, 2000; White, 1984). However, it may be more prone to raters' errors and hard to justify the resultant value (Klein et al., 1998). On the other hand, analytical rubrics are more detailed and thought to increase accuracy by directing

the rater's attention to specific parts of the dimension to be assessed. Hence, analytic scoring procedures are thought to reduce subjectivity.

A main concern, however, is that in the context of oral performance, analytic schemes could lead to greater dependency on memory to judge specific dimensions. Specifically, the oral performance is temporal and impractical to repeat leading to greater dependency on memory of the evaluator. As a result, such advantage of analytic rubrics over holistic is not clear. For instance, with a detailed rating scale the raters would be forced to either record their ratings as the student performs the task or after the student has finished. So, it is possible that the rater if recording at the same time could miss important parts of the performance. Yet, if the rater waits until the student is done then good memory of the performance demonstrated is a key to reliable and valid score assignment.. In other words, concise or perhaps instant recording of specific dimensions could cause other dimensions go missing and delayed marking is memory-dependent. Such sought balance might make it difficult for the raters to give consistent and accurate ratings. A holistic rubric, alternatively, allows the opportunity to pay more attention to the behavior in general, and not to be distracted by reporting. The disadvantage though is potentially more subjectivity in score assignment.

Holistic rubrics may better suited Quranic appraisal since it gives general ratings of the performance in whole if the purpose is to give a pass/failing grades. Yet, analytic rubric would be of immense benefit to the teacher as he or she points to the students' strengths and weaknesses. However, given the dilemma previously stated a study investigating the best scoring scheme for Quranic recitation assessment seems highly important.

Significance of the Study

The present study could further our knowledge of the best scoring scheme for oral assessments in general and for Quranic recitation assessment in particular. It could assist and aid the general practice of assessing Quranic recitations as it gives some insight into the consistency of oral ratings of different scoring schemes. It explores the scoring schemes efficacy in a widely used context yet rarely or perhaps never been studied (i.e., Quran recitation).

Quranic recitation assessment relies heavily on the use of raters and passages. They both help give more information concerning students' ability to recite Quran properly. They however are a source of measurement error (e.g., Cardinet, Tourneur & Allal, 1976; Harrison, McAfee, & Caldwell, 2002; Rangel, 1990). Such a need for measurement models to monitor these two sources explains the movement toward Generalizability Theory (Brennan, 2001; Cronbach, Gleser, Nanda, & Rajaratnam, 1972), and Many Facet Rasch Measurement Model, abbreviated MFRM (Linacre, 1994)

The Generalizability Theory (G-Theory) has come to light to overcome the limitations of Classical test theory (CTT). In the CTT model the observed scores are composed of two independent parts; true and undifferentiated error sources (e.g., Thissen & Wainer, 2001). This notion of undifferentiated source of errors and the other barely achieved assumption of parallel tests has led to the inception of a librating measurement theory namely the G-theory. It significantly avoids the restrictive assumptions of CTT (Suen, 1990) and uses different statistical analyses (i.e., Analysis of Variance, ANOVA).

The ability of G-theory to handle different source of errors, to simultaneously consider different decision studies, and to be of use with relative and absolute

orientations (Shavelson & Webb, 1991) make it of value especially for performance assessment that utilizes raters. In a sense, depending on the assessment scenarios (e.g., number of sources of error) G-Theory analyses result in estimates of the degree to which a generalization from a sample of data to a larger domain is accurate (i.e., dependable).

MFRM (Linacre, 1994) shares great similarity with G-theory. In a sense, they both, for example, differentiate error sources, treat more than one facet, and handle missing observations. Yet, it evolves around the central idea of objective measurement where invariance is emphasized. MFRM goes beyond raw data, and models a response vector, and the degree to which such data at hand behave in accordance to the expectations by the model is a key factor. It extends the basic Rasch model (Rasch, 1960) to model effects beyond those of persons and items (Bond & Fox, 2007). Rasch model, conceive of probability of arriving to a correct response as a function of person latent ability and item difficulty. Yet, MFRM add more wealth to the assessment by investigating the impact of other facets (e.g., raters) on the ratings. Consequently, the dependability of the scores or ratings is well approximated when relevant and important factors that have great impact on scores are modeled.

Several attractive features of MFRM include its ability to produce separate estimates for each facet and its elements that are invariant, of equal interval scale and within a common frame of reference (Bond & Fox, 2007; Linacre, 1994). However, such features are conditioned on securing a good fit of the data to the model. Therefore, different quality control statistics are produced to judge the degree to which such measures and estimates can be trusted.

This study explores facets other than students' ability to recite Quran properly and assesses their impact on the dependability of ratings. Equally essential, it may shed lights on the use of Many-Facet Rasch Measurements (MFRM) and Generalizability Theory (G-Theory) for evaluating consistency in Quranic assessments and thus could guide the Quranic assessment practice.

Additions of the study

Although oral assessment in general and Quranic assessment are used in a regular basis in the Saudi educational institutes, Quran assessment is an under-researched topic. In addition to the above, Quran assessments are not limited to or used only in educational settings but also in grand and prestigious contests at both national and international levels (e.g., King Abdul-Aziz International Contest of Quran). Such institutions might find such line of research of immense benefit as they plan and revise their policies and practices. Equally important, this study introduces holistic and analytic scoring rubrics to the Quranic appraisal practice and assesses their effectiveness.

Research Questions

1. Are scaled ratings of oral performance using analytic scoring rubrics more consistent than ratings using holistic methods?
2. Within the Quranic recitation assessment context, do analytic scoring rubrics improve the consistency of the raw scores analyzed by Generalizability Theory significantly and beyond those achieved by holistic scoring rubrics? In other words, are raw ratings of oral performance using analytic scoring rubrics more consistent and have higher coefficients of generalizability (reliability) and dependability indices than ratings using holistic rubrics?

3. Do raters find it difficult to be consistent for some scoring criteria than for others?
4. Do calibrated measures of Quran passages statistically differ in difficulty within each scoring rubric?
5. Do calibrated measures of raters statistically differ in leniency as they rate students' recitations within each scoring rubric?
6. Do analytic scoring method results in more distributed measures of students' scores?

Key terms introduced and referenced in this dissertation include:

Constructed-response assessments: it refers to any performance assessment that goes beyond merely selected response format like MC or true false (Osterlind, 2006).

Rubric: A set of coherent and explicit rules to award particular performance a certain score, and it usually provides examples or description of different levels of performance (Linn & Gronlund, 2000; Osterlind, 2006).

Holistic Scoring: a rubric that gives an overall judgment or score to a performance on different central dimensions, it includes typically a description of different levels of performance in order to help the rater chose which scoring criterion best matches the performance (Linn & Gronlund, 2000; Osterlind, 2006).

Analytic Scoring: A rubric that yields a separate score for each dimensions or criterion of the performance. In analytic rubrics each dimension is associated with different level of performance and thus rated separately. (Nitko, 1994; Osterlind, 2006)

Oral Recitation: the reading of the holy Quran they way the prophet Muhammad PBUH read it. In other words, reciting the Quran according to a set of rules called Tajweed that

assure the letters are articulated properly and given its due length and accidental and permanent characteristics (Al-Hussary, 1999; Musri, 2005).

Classical True Score Theory: It is the oldest scoring theory of reliability. It postulates that any observed score can be portioned into true and error components. The true component is the average of all measurements taken over repeated parallel forms or occasions and independent of the error component. The error part however, is random and it is expected value is zero. That is over repeated measures such a component has an average value of zero and has a variance that is independent and homogenous across subjects (Thissen & Wainer, 2001). Such a theory addresses the percentage of true variation in the observed score (i.e., reliability), and provides an estimate of the degree of measurement error in the data. Yet however, despite its wide use, it suffers from several limitations. It cannot handle more than one source of error at the same time, and the resulting statistics are not sample-free (Shultz & Whitney, 2005).

Item Response Theory: a group of different models that relate the probability of a certain response to a latent trait (e.g., intelligence) and item characteristics (Hambleton, Swaminathan & Rogers, 1991; Ostini & Nering, 2006). It relies heavily on the assumption of a complete latent space and local independence (i.e., any relationship between responses can be accounted for by the latent trait) and produces invariant estimates (i.e., sample independent statistics), and conditional error of measurement (Embretson & Reise, 2000; Hambleton & Swaminathan, 1990; Hambleton, Swaminathan & Rogers, 1991). However, all IRT models need to be judged for fit to the data before estimates can be trusted (Hambleton, Swaminathan & Rogers, 1991)

Rasch Modeling (RM): A model developed by Rasch (1960) that relates the probability of endorsing a correct item to a person's latent ability and severity of the item. It shares many features and assumptions of IRT (e.g., local independence). Yet, it differs in several important respects. One difference is mathematical. Specifically, items in RM are allowed to vary in difficulty only (Embretson & Reise, 2000). Whereas item discrimination and guessing are modeled in other IRT models, RM addresses them in infit and outfit statistics (Wright, 1995). A second difference lies in the sufficiency of total scores to estimate parameters of RM (Fischer, 1995). Another difference however is theoretical and has to do with specific objectivity; refers to the ability to compare people for proficiency without any reference to items (e.g., Thissen & Wainer, 2001).

Many-Facets Rasch Modeling: Many-Facets Rasch Modeling, MFRM (Linacre, 1994), is an extension of RM that treats facets other than persons and items.

Chapter Two

Literature Review

This chapter is organized into two main sections; performance assessment scoring and its quality controls. It discusses first the imperfect nature of observed scores, the assessment formats, and why raters differ. Then, it elaborates on the ways to evaluate the scores worthiness and to quantify the reliability and raters' agreements. Finally, the nature of oral performance appraisal and the scoring methods are discussed, and a section on the importance and the need for theories of measurement is provided.

Measurements are Never Perfect

There has been a vast body of research on the study of reliability and validity of scores. Students' scores be it on a test, a rating scale or the like are not a pure reflection of their ability or the construct under probe but however, a mixture of true or pure dimensions (e.g., ability in the case of achievement test) and some other factors. Consideration of these factors and partitioning them is a major distinction between two measurement theories: Classical Test Theory (CTT) and its extension, Generalizability Theory (G-Theory). While CTT analyzes any observed score into true and undifferentiated error terms, G-Theory further partitions the error term into different sources. These sources are handled in many ways ranging from equalizing (standardization) them (e.g., CTT) to modeling them (e.g., MFRM, G-theory).

Any one of these factors or sources could be pertinent to the assessment situation depending on the purpose of assessment (i.e. formative vs. summative), the extent of generalization, or scoring orientation (relative vs. absolute). Of these factors that have

been frequently investigated, raters, tasks, scoring criteria, and test formats were of immense value in the study of error patterns (e.g., Lane & Stone, 2006).

Assessment Formats and Use of Raters

Teachers have many assessment formats to choose from when assessing their students' performance (e.g., Smith Jr. & Kulikowich, 2004). Commonly used formats are Multiple Choice (MC), essays and Rating Scales (RS). These commonly used formats have their strengths and weaknesses. One of the strengths in MC formats is that they are easier to administer and to score (Haladyna, 1997; Popham, 1993). Another strength is that their ability to cover a wide range of topics in lesser time when compared to the essay format (Haladyna, 1997). However, they are inherently prone to guessing (Nitko, 2004), and cheating. A major criticism is centered on the low ability in appraising high critical thinking (HCT) (Chase, 1999; Haladyna, 1997). Similarly, test-taking skills could and contaminate the test-results. Thus, students who have much exposure to MC format tend to use certain strategies to answer questions of which they know not (Shultz & Whitney, 2005; Westgaard, 1999).

For the above reasons, performance assessment was seen as a more valid approach. It is valid since it provides the opportunity for the assessor to evaluate the participants given real-life contexts which probably enhances the validity of the scores. Put differently, the evaluaee is granted a better chance to demonstrate his or her knowledge and reasoning skills, and so is the evaluator to asses in-depth content (Nitko, 2004), and to assess HCT (Chase, 1999). Yet, such strengths come at cost.

Going beyond objectively scored tests necessitates the use of raters or scorers. Once the raters are allowed to be part of the assessment, subjectivity is an inevitable reality. In

other words, the inclusion of raters in the measurement process would probably introduce some variations in the scores of the examinees that are irrelevant to the examinees real differences. In some cases, such variation might be small and trivial and its effect could be ignored, yet in other cases raters could show differences as large as the performance shown by the students (Ruggles, 1911 in Linacre, 1994). Part of such variation, for instance, could be attributed to being consistently severe or lenient (e.g., O'Neill & Lunz, 2000). It has been long documented that raters differ in severity when scoring students' performance (e. g., Harrison, McAfee, & Caldwell, 2002), and these differences in severity account for approximately fifty percent of the variation among raters (Edgeworth, 1890 in Linacre, 1994). Most importantly, such differences in severity may still existent even after thorough training of how to assign scores using rubrics or rating scale (Bonk & Ockey, 2003; Smith & Kulikowich, 2004).

The predicament is that such variation could cloud or contaminate the true differences among examinees, and as a result, important and critical decisions (e.g., certification, diagnosis) could improperly be made. Thus, performance assessment formats are frequently characterized as being tools that yield low reliable scores (Linn & Gronlund, 2000; Nitko, 2004; Parkes, 2000).

Ways to Handle Raters' Differences

Training of raters has been one major way to protect against such unwanted variation in scoring and a safeguard to ensure that an examinee receives his or her deserved grades independently of who happens to grade the responses. Ideally, scores should be rater-independent and only represent examinee's level on the construct being measured. Thus, the training usually seeks to impose a common perception of excellence and to bring

raters to close agreements so that the examinee would be awarded a similar if not the same rating by each rater.

Such agreement could be improved by using a scoring rubric (e.g., Linn & Gronlund, 2000). A rubric is a guide for how assigning scores numerically given overall or parts of a performance. Therefore, training programs train raters on using rubrics as they are generally effective in achieving a common understanding of the scale and reducing the variation among raters. Yet, another simple way to handle rater differences is to require raters to give justifications for their ratings (Mero & Motowidlo, 1995), for knowing that someone else is rating the same project helps rating accuracy (Dennis, 2007). This way, raters may become aware of the possibility that some subjectivity could be introduced into the ratings.

Sources of Raters' Variation

Raters' differences may come from different sources. One source of differences might be the use of different standards (e.g., O'Neill & Lunz, 2000,), and thus crafting an agreed upon rubric is necessary (Nitko, 2004). In other words, a vague definition of the targeted behavior may introduce such a variation of raters. For instance, some raters might consider some behavior as a reflection of the construct being measured while others do not (Nunnally & Bernstein, 1994; Suen & Ary, 1989). Another reason for such differing standards might be related to employing different sources or groups of raters (e.g., new raters vs. experts), where different groups of raters would see quality of performance differently and as such assign score differently (Murphy & Cleveland, 1995, as cited in Facteau & Craig, 2001; Greguras & Robie, 1998). Raters' expertise could also contribute to their perception of varied dimensionality in the responses.

Closely related to inconsistent adherence to one standard is the rater drift problem (Nunnally, 1959). This problem occurs when a rater drifts from the standard due to fatigue. Rater drift could perhaps be related to the change of definition of the correct answer through time and after being exposed to different answers (Moskal & Leydens, 2000). Such an issue is widely observed and some approaches were already developed to handle it. Explicitly, the observer's definition of the targeted behavior needs to be frequently checked (Paul & Lentz 1977, as cited in Suen & Ary, 1989) and the rater needs to be kept motivated (Reid, 1982, as cited in Suen & Ary, 1989).

Another plausible reason for raters' variations might be related to the Halo error as named by Thorndike (1920). Halo effect refers to the impact of impression a rater may have about an examinee on his or her appraisal (Bingham, 1939). It is a distortion (i.e., error) that renders ratings inaccurate by thinking of the evaluaee on the whole and failing to consider different aspects of the ratee as separate and independent (Thorndike, 1920). Such constant error could contaminate any judgmental rating and as a safety-guard measure raters should provide not only independent ratings but evidence to (support it) such ratings (Thorndike, 1920).

Carryover effect also is a source of differences in grading. This happens when a rater's grading of a response influences his/her next grading of the same person on another response. For instance, if the ratee performed well on the first task, the rater might be inclined to give high ratings on the second task even if the performance was poor. Another source of differences in grading is the range restriction. It has been widely observed that some raters tend to give middle scores and avoid extreme scores. This error could be attributed to the fear of giving extremely high or low scores, and it is frequently

observed with new raters or those exposed to training programs that over-emphasize close agreement.

There are viable and valuable suggestions for handling and overcoming such obstacles (e.g., training, rubric developing). However, following such suggestions; although, extremely important and necessary, does not entirely eliminate sources of error (Suen & Ary, 1989). For the above reasons and others, utilizing statistical and measurement techniques for further checking proves essential.

Quantification of Raters' Behaviors:

Given that variation among raters does exist, quantifying such variation is necessary for the respective assessing situations. There are statistical procedures to address questions about measurement error. These procedures differ in a) how to quantify b) what to quantify c) the nature of outcome and d) the level of scale.

Quality of Ratings

Since ratings are not always pure and accurate, and that they are used for different levels of decisions, it appears important that ratings continue to undergo increasing scrutiny regarding quality. From a psychometric point of view, quality of measurement is by and large checked for its reliability (i.e., dependability or reproducibility) and validity.

Reliability refers to the degree of "absence of random error variance" (Suen & Ary, 1989, p. 118). The reliability coefficient is a lower estimate of how strongly a true score correlates with an observed score. The true score is operationally defined as the average of scores over repeated measures of the same person under similar conditions (CTT). By comparison, the observed score is the resultant score from the current

administration of a measure. Such score is a combination of two independent entities; true and error. The error score will average out to zero over repeated measures.

Validity, on the other hand, concerns the accumulative evidences on the degree of adequacy with respect to using and interpreting the measurements (AERA, APA, & NCME, 1999). Only indices of reliability as indicators of psychometric qualities are addressed in this study and hence, discussed in the ensuing sections.

Indices of Reliability

Measures of observational reliability can be classified, although not exclusively, into interobserver agreement, intraobserver reliability, or interclass generalizability (Suen, Ary, & Covalt, 1990). According to Suen, Ary and Covalt, interobserver agreement indices are pure statistical methods that assess the degree to which two or perhaps more independent observers yield similar ratings (i.e., agreements on whether certain behaviors occurred or not). They include, naming a few, Kappa Coefficient, occurrences and nonoccurrence agreement indices, and Scott's Pi coefficient (Suen, Lee, Prochnow-LaGrow, 1985). From yet different perspective and depending on certain assumptions a different class of indices, the intraobserver reliability indices, assesses the dependability of the score given by a single observer. In other words, it assesses how consistent an observer is in ratings the same behavior on repeated occasions. It is represented by Pearson's ρ or Phi Coefficient (φ), and only good for relative interpretation. More important, it depends on the assumptions of parallelism (CTT) and any violation of this assumption would risk their interpretation (Suen & Ary, 1989).

The third yet more encompassing category as in Suen, Ary and Covalt (1990) is the interclass generalizability, where the true variation is that of the subject (i.e., the

object of measurement) and the rest of variations are considered error (in criterion-referenced assessments). An exception to this is when a facet of measurement is treated as fixed. Under such a scenario the interaction of that facet with the object of measurement turns out to be a part of the true variation. Examples of interclass indices are Hartmann's coefficient ($r_n 2$), Berk's $r_1 1$ and $r_2 2$ and Cronbach's alpha (α_n) recommended by Bakeman and Gottman (Suen, 1988).

Measures of Inter-Rater Agreement

- ***Smaller /Larger Index.***

Smaller/Large index ranges from .00 to 1.00 and it is obtained by having two independent raters rate a specific construct and then dividing the smaller rating by the larger one. Although it is simple and popular it does not have strong mathematical properties or sufficient theoretical basis (Suen & Ary, 1989; Suen, Lee & Prochnow-LaGrow, 1985). Suen, Lee and Prochnow-LaGrow (1985) stressed that the extent of inflation due to chance agreement this index might have is unknown, and its interpretation is sometimes misleading. Neither is it conceptually equivalent to any of the recognized measures of reliability nor does it correlate with them, and as a result, its use should be discontinued (Suen, Lee & Prochnow-LaGrow, 1985). Hence, they suggested that its use should be allowed only when its nature has been substantiated as measure of reliability.

- ***Interrater Agreement Index.***

  - **Proportion agreement index:**

$$P_0 = \sum P_{ii} .$$

$$= [(\text{No. of agreements}/\text{No. of agreements} + \text{No. of disagreements})]$$

Such an index or its variant the percentage agreement index ($P_0\%$), has been the most frequently used index of interobserver agreement indices for behavioral data (Suen & Lee, 1985).

However, such index might be inflated by mere chance-agreement. Interestingly, such mere chance-agreement is likely found when the observed behavior's prevalence is toward either extreme end (.00 or 1.00) (Costello, 1973; Hartmann, 1977; Hopkins & Herman, 1977; Johnson & Bolstad, 1973; Mitchell, 1979, as cited in Suen & Ary, 1990). The existence of chance agreement and its dependency on behavior prevalence complicate the interpretation of percentage agreement index and make identical values incomparable (Suen & Ary, 1988).

As a solution, some researchers suggested the use of other measures like Occurrence / Nonoccurrence agreement indices in lieu of P_0 (Suen & Ary, 2005). Yet, others have suggested the use significance testing. However, the result of such tests would be in many cases misleading given that observational data are often autocorrelated. Therefore Kelly (as cited in Suen and Ary 1988) suggested that two conditions are to be met before percentage agreement index can be trusted; prevalence of behavior should be at most 0.8 or at least 0.2 and the resultant index is at least 0.9. Such stipulations are met when the observations are not autocorrelated (Ary & Suen, 1985). Although, such

conditions would solve the autocorrelation and prevalence problem, such an index is warned against, and considered inappropriate to use as an Interrater agreement statistic (Suen & Lee, 1985).

Despite the wide use of P_0, Suen and Lee (1985) found that using a relaxed standard, 25% to 75% of a sample of published studies would have had unreliable indices had Kappa coefficient been used. Similarly, using a rigorous standard 50% to 75% of the sample would have had unreliable indices had Kappa coefficient been used. Therefore, alternative statistics were suggested such as, Scott's π coefficient or Cohen's Kappa Coefficient (e.g., Suen & Lee, 1985), or G-index (Green, 1981).

- ***PI Coefficient (π-statistic)*:**

$$\pi \text{ Coefficient } = \frac{P - e(\pi)}{1 - e\ (\pi)},$$

Scott (1955) recommended π-coefficient as measure of agreement between raters that adjust for chance agreement. However, it is not appropriate for marginal distributions that are not identical across raters (Burton, 1981).

- ***G-index*** (Holley & Guilford, 1964):

$$G = 2(P_0) - 1.$$

G-index was developed as a solution to the problems associated with correlational indices of interobserver reliability (Green, 1981).

- ***Cohen's Kappa κ*:**

$$\text{Cohen's Kappa } \kappa = \frac{\sum_{i=1}^{I} P_{ii} - \sum_{i=1}^{I} P_{i.}P_{.i}}{1 - \sum_{i=1}^{I} P_{i.}P_{.i}}$$

It is a one-statistic summary suggested by Cohen (1960) which indicates the raters' agreement achieved beyond that of chance (i.e., chance-corrected agreement). It

can handle both dichotomous and polytomous (at the nominal level) outcome variables and accommodates two raters or more. It is adequate when raters are treated equally (Kvalseth, 1991), but inappropriate otherwise. In other words, when one rater is given importance over other raters (e.g., one rater is a standard or professional), ϰ coefficient is inappropriate. Its value ranges from -1.00 to +1.00; where +1.00 indicates a perfect agreement, .00 indicates an agreement that is not different from those made randomly, and -1.00 indicates a perfect disagreement. It is suggested as a remedy to the limitation of π-coefficient, namely the chance agreement adjustment. While the ratings of two raters classifying participants in the targeted categories are averaged and raised to the second power in the calculation of π, the ϰ-coefficient is based on multiplication of such proportions (Gwet, 2002). Worth noting, however, there is ambiguity in what is meant by chance (Green, 1981).

A major concern with ϰ-coefficient is that the overall agreement does not guarantee that the classifications into categories are equally agreed upon by raters. In other words, raters could show strong agreement in rating subjects in some categories but moderate to very weak in others (von Eye & Mun, 2005). Therefore, another supplementary statistic is suggested; Conditional Kappa.

- ***Conditional Kappa*:**

  Conditional Kappa = $\frac{P_{ii}-P_{i.}P_{.i}}{P_{i.}-P_{i.}P_{.i}}$

Conditional Kappa is a local agreement index that measures the consensus among raters in specific categories. Testing the hypothesis that the partial Kappa (i.e., Conditional Kappa) is .00 and, building a confidence around the index could be obtained based on the respective variance estimates (von Eye & Mun, 2005).

- ***Weighted Kappa*:**

$$\text{Weighted Kappa} = \frac{(\sum_{i=1}^{I} \sum_{j=1}^{J} \omega_{ij} P_{ii} - \sum_{i=1}^{I} \sum_{j=1}^{J} \omega_{ij} P_{i.} P_{.i})}{(1 - \sum_{i=1}^{I} \sum_{j=1}^{J} \omega_{ij} P_{i.} P_{.i})}$$

Since Kappa coefficient is first introduced for categorical variables, the weighted Kappa was suggested to allow for outcome at the ordinal level. Consequently, interpretation of the weighted Kappa would be valid as long as the variable is at the ordinal level, and the weighted assignment is justified (von Eye & Mun, 2005). The weighted Kappa is mainly used for hypothesis testing, but not frequently used as an index for "interobserver agreement or intraobserver reliability" (Suen & Ary, 1989, p. 113).

- ***Occurrence and Nonoccurrence Agreement Indices*:**

  Pocc = [(Occurrence agreements) / (occurrence agreements + disagreements)] *100%

While the occurrence agreements refer to the number of times both raters concurred on the occurrence of the behavior, the nonoccurrence agreement indicates the number of times both raters concurred on the nonoccurrence of the respective behavior. Disagreements are otherwise. They are appropriate for situations in which only two observers assign scores and suggested as a replacement forP_0, since they help reduce chance agreement. However, they do not entirely remove error (Suen & Ary, 1989).

Intraobserver Reliability Indices

Intraobserver reliability indices are based on restrictive CTT assumptions (i.e., Parallelism). Therefore, a brief introduction to CTT will be provided, and for more

general treatment of CTT the reader is referred to Thissen and Wainer (2001) or Nunnally and Bernstein (1994). The CTT posits that any given score (i.e., observed) is composed of two independent components; true and error. The true score is ideally the average of all resultant scores of the same person on parallel forms of a test (Nunnally & Bernstein,1994). Any difference between the observed and the true score is an error, and this error term is random and independent of the true score (Thissen & Wainer, 2001). In addition, the error variation is homogenous across subjects (e.g., Osterlind, 2006). Hence, the expected value of error is zero. On the other hand, the expected value of the observed score is the true score (the average score over repeated measures on parallel forms), and the true score of an examinee is independent of that of any other examinee.

The variance of the observed score is composed of true and error variances, and dividing the true variance by the observed variance gives a coefficient of reliability. It is the proportion of true variance to the observed. In other words, it tells how much of the variation among the examinees' scores is considered true variation and how much is random error. Yet, direct estimate of reliability is impossible before finding parallel forms of a measure. Thus, practical methods or strategies to achieve parallel forms were developed. Specifically, test-retest, equivalent forms, and split-half strategies.

Parallel forms refer to the condition where two tests that have equal means and equal total variances, and that the error variance of one form is independent of the other form; hence, error variances are equal (e.g., Osterlind, 2006). The correlation of the scores on two equivalent forms for instance is a reliability coefficient ranging from .00 to 1.00. A value of one indicates a perfect reliability; there is no random error variation. Yet, a

reliability coefficient of .00 indicates that the assignment of scores for participants is as good as random; the variation of scores is completely random error.

By extension, two observers are considered two equivalent forms (Murphy & DeShon, 2000), and the correlation between the scores in this case is a reliability coefficient, that is an interobserver reliability index (Suen, Ary, 1989; Suen, Ary & Covalt, 1990). Unlike interobserver agreement, intraobserver reliability is more complex and certain assumptions need to be met (i.e., Parallel Test Assumption, PTA) before interpretation could be trusted. When these assumptions are met, Person's r is a reliability coefficient and an estimate of the proportion of the true variation to the total variation (Murphy & DeShon, 2000; Suen, 1990; Suen & Ary, 1989).

For observational studies, having the same observer observe the same behavior is unattainable (i.e., costly), and the use of two equivalent observers seems more feasible (Suen & Ary, 1989). They explained that it is impossible to use split-half or test-retest strategies, for the former requires purposeful selection and arrangement of behaviors, and the latter is impossible (i.e., behaviors do not repeat exactly) and the alternative is costly (i.e., having the same observer watch a videotaped behavior over and over). Therefore, finding equally trained observers who share the same standards is akin to parallel forms of a measure.

It could be understood from Suen and Ary (1989) argument that different context might require different adopted parallel-form strategies. Evidences on such idea are scattered in the literature. For instance, Murphy and DeShon (2000) argue against the tenability of parallelism for inter-rater reliability for job appraisal in organizations, suggesting that raters are not equivalent or comparable, since raters bring to the

assessment situation influences that are neither true ratings nor random errors. Specifically, they stated that raters' agreement is not only due to true standing of the examinee for the measured construct but also a combination of factors of which is the common interpretation, standards and so forth. Therefore, raters might disagree because they see different facets and amount of performance which cannot be considered random error. Although the estimated coefficients are considered useful, Murphy and DeShon argue against calling them inter-rater reliability coefficients as well as deriving other estimates like standard error of measurement or the adjustment for attenuation.

Interclass Correlation Coefficients

Another way to estimate rater reliability is through interclass correlations, abbreviated as ICC. Interclass correlations are based on the analysis of variance where the true and error variances are estimated directly. Avoiding parallel assumptions, it partitions the total variance into that of systematic true variance (i.e., true differences among the subjects), systematic raters' variance, and random error variance that lump their interaction and other unaccounted for random errors (Shrout & Fleiss, 1979; Suen & Ary, 1989). Based on how the data are to be interpreted (relative vs. absolute), the systematic rater variance may be ignored (e.g., for, absolute interpretation, systematic rater variance is a source of error). In effect, it is a special case of generalizability theory to be discussed later in the following sections (Shrout & Fleiss, 1979).

The interclass correlation coefficient for when the relative model is adopted is calculated as follows (equivalent to traditional intraobserver reliability):

$$\rho_1^{\,2} = \frac{\sigma^2(subjects)}{\sigma^2(subjects) + \sigma^2(error)}$$

On the other hand if the absolute model to be adopted the calculation is as follows:

$$\rho_2^{\,2} = \frac{\sigma^2(subjects)}{\sigma^2(subjects) + \sigma^2(raters) + \sigma^2(error)}$$

Worth reiterating, intraobserver reliability is an estimate of the consistency of scores given by a single rater (Suen & Ary, 1989). Other traditional intra-rater reliability coefficients could be obtained, and depending on the unit of reliability analysis the number of raters such indices could be adjusted (cf., Shrout & Fleiss, 1979; Suen & Ary, 1989).

Summary on Consistency and Consensus

There are different proposed statistics for estimating reliability of ratings. Some are mere statistical estimates that lack measurement framework (i.e., descriptive statistics), while others, are indeed guided by measurement perspective. The latter encompasses three measurement theories, CTT, G-Theory and MFRM.

Consensus statistics are mere descriptions of the current measuring situation and thus inadequate for generalizations beyond the samples and conditions. Neither they can provide any confidence bands around the scores obtained, nor can they provide any estimates of how much measurement error is there in the measures.

Those derived with CTT orientation, to the contrary, can be generalized to similar situations and capable of providing standard error of measurements and confidence bands around the scores. They however suffer from restrictive assumptions that are rarely met

in practice. On the other hand, G-Theory lends itself usefully not only because its assumptions are more relaxed, but also because of its flexibility to handle different data collection designs. Consistency estimates for each design could be estimated, and ways to improve consistency can be manipulated until the best configuration is achieved. For instance, samples in particular facets could be added or reduced depending on their effect on the consistency estimates.

MFRM with the main focus on objective measurement adds more wealth to the assessment analysis. Specifically, it reports separate parameter estimates for each facet of the measuring situation and its elements and conveniently allows for direct comparisons between and within facets since they are placed in the same frame of reference. Notably, it gives rich and detailed diagnostic statistics that could help improve the scale and the measuring situation and give at the same time more insight into the validity of the data.

Why not Interrater Reliability?

Despite the fact that extant research has used interrater reliability indices; the use of such indices has been challenged (e.g., Bond & Fox, 2007). First, interrater reliability indices cannot be used as or be equivalent to other indices of score reliability such as test-retest or KR-20 (Reckase, 1995). As Reckase points out, the Interrater reliability does not include variability due to tasks embodied in all other forms of classical reliability estimators. Second, interrater reliability overestimates the reliability of the ratings since it treats the tasks as fixed (Brennan, 1995). Therefore, failing to include the task error variance means technically allowing such variation to be treated as true variation and not error (Brennan, 1995). Finally, Interrater reliability is of no value for criterion-referenced

assessments as it only detects consistency of rank ordering (e.g., Bond & Fox, 2007), and hence cannot detect systematic differences between judges (Nunnally & Bernstein, 1994).

Why are G-Theory and MFRM Important?

G-Theory is a versatile data collection and measurement model. It fits most of the assessment designs and provides useful indices about the quality of the resulting scores. For instance, standard error of measurement (SEM) could be obtained to set a confidence interval (CI) around the score to obtain the lower and upper bounds of the score. Such CI could be of great assistance when making crucial decisions. For instance, it is highly important when the scores of two or more candidates are compared.

MFRM shares great similarity with G-theory. In a sense, they both, for example, differentiate error sources, treat more than one facet, and handle missing observations. Yet, it evolves around the central idea of objective measurement where invariance is emphasized. Although MFRM goes beyond raw data, they both (i.e., MFRM & G-Theory) represent useful theories of measurement in performance assessment.

Theories of Measurement

There has been a great need for measurement models that extends beyond objective scoring to handle situations in which raters or judges are part of the assessment procedures. Educational settings are abundant of performance assessments (e.g., essays, long-term projects). Clearly, raters are not of concern in objectively scored assessment; however, going beyond such an assessment automatically contribute to raters becoming a critical part of the assessing procedures that should be modeled effectively. Such modeling is easily possible with the help of advancement in measurement theories. For instance, G-Theory can accommodate raters and model them as a facet of measurement

and inform the researcher of the number of raters he needs in order to have reliable measurement of his or her students. Likewise, MFRM not only has the ability to model raters, items, and subjects but also to adjust the resulting scores accordingly.

Such differences between the raters (e.g., severity) could be easily considered and taken into account. Thus, reliability of the scores could not be only established easily for various settings and testing conditions, but also could be checked and enhanced by taking into account such sources of variation (e.g., increasing the number of raters in the D-study). Furthermore, students' scores for instance could be adjusted for raters' severity and task difficulty. MFRM model compensates mathematically for such severity, for the raters' facet and the examinees' ability are positioned on the same continuum metric. Interestingly, MFRM can reduce the cost of assessment by providing adequate estimates without having a fully crossed design as long as a sufficient linkage is secured (Bond & Fox, 2007; Linacre, 1994). Thus, considering the inability of training to eliminate raters' overall severity (Lunz & Stahl, 1994) along with such remarkable advancement in measurement, a move toward utilizing these models seems to be of importance.

G-Theory

G-theory (Brennan, 2001; Cronbach, Gleser, Nanda, & Rajaratnam, 1972) has come to light as a solution to the limitations of Classical test theory (CTT). In the CTT model the observed scores are composed of two independent parts; true and undifferentiated error sources. This notion of undifferentiated source of errors and the other barely achieved assumption of parallel tests has led to the inception of a librating measurement theory namely the G-theory. It significantly weakened the assumption of

CTT and came with different statistical analyses. The ability of G-theory to handle different sources of errors, to simultaneously consider different decision studies, and to be of use with relative and absolute orientations are improvements in modeling that can contribute to more precise score assignment than what is achieved by CTT.

Components of the G-Theory

Because in G-Theory as in the other theories researcher uses pertinent terminologies, a brief introduction is provided, however, for complete treatment of such a topic readers are referred to (Brennan, 2001; Brennan, 1992; Shavelson & Webb, 1991). Often-mentioned terminologies of G-Theory are the universe score, facet, random vs. fixed, crossed vs. nested, and the universe of admissible observations and the universe of generalizations, hence briefly defined.

The universe score is the expected score of a person over the facets under probe, where facets denote any specified source of error in the data collection design (e.g., raters). A random facet refers to the conditions (levels of the factor) in each facet that are specified as an exchangeable (i.e., random) sample of a larger pool representing the population. Fixed effects, on the other hand, are not to be generalized beyond the sample. For instance when they are the pool itself and not a sample of it, or when the same elements will be used in measurement situations to come. A complete crossed design refers to the full data collection plan in which each object of measurement or facet condition is encountered with each other facet condition. However, in the nested design not each examinee and each element in the facets is encountered with each other facet

element. A mixed design would have both and thus only those considered cross would encounter each element of the other facets.

In G-theory, there are two universes; the universe of admissible observations and the universe of generalization (Brennan, 2001). The first specifies any acceptable conditions (e.g., potential raters or prompts) in any measurement facet, whereas the latter specifies to what degree or extent the researcher is wishing to generalize. For instance, if a researcher is interested in evaluating students on an essay-history test, the researcher might specify raters and tasks facets as admissible observations. In other words, the relative impact of the raters and the tasks on the students' scores are the only sources to be taken into account (Brennan, 2001).

G-theory lends itself profitably to the data collector. It accommodates both single and multifaceted settings. It is performed through two stages; generalizability and decision studies (G-study and D-study, respectively). In G-study, the researcher obtains the variance component estimates, in order to quantify the relative impact of such sources on the object of measurement (Brennan, 2001; Shavelson & Webb, 1991). These estimates are then used to obtain the generalizability coefficient and/or dependability index for the scores in the D-study. The aforementioned information (variance component estimates) is utilized in the D-study and different scenarios are manipulated (e.g., increasing number of tasks or raters). D-study could but does not have to be with the same sample and/or the same design of the G-study. The best cost-effective configuration is then suggested.

Contradictory to CTT where examinees are always the object of measurement, in the G-Theory the object of measurement is usually, though it does not have to be, the

examinees. So the variability of such a source is considered a true variance (i.e., universe score variance). However, any other source of variability (that is to say, facet) is considered a source of error. The interaction term of the object of measurement with any other facet is considered a part of the error term and the only source of error when the relative orientation is to be adopted and all facets are random. On the other hand, all terms other than the object of measurement is considered part of the error term when the absolute orientation is adopted and the design is fully crossed.

Specifying the design as random or fixed also has an impact on the analyses and its result. Fixed facet is considered when the conditions in any facet are not to be generalized beyond the sample (Shavelson & Webb, 1991). This necessitates that such conditions are to be used in any examination to come and therefore, this facet's interaction with the object of measurement is no longer an error term but rather part of the true variance (Suen, Ary & Covalt, 1990). It has to be mentioned that at least one facet should be random in order to make possible the evaluation of the scores reliability (Suen, 1991).

The G-coefficient is the ratio of the true variance (the variance of the object of measurement) to itself plus the relative error:

$$\mathrm{E}\rho^2 = \frac{\sigma^2(\tau)}{\sigma^2(\tau) + \sigma^2(\delta)}$$

The D-index is the ratio of the true variance to itself plus absolute error variance:

$$\Phi = \frac{\sigma^2(\tau)}{\sigma^2(\tau) + \sigma^2(\Delta)}$$

Given such flexibility, raters could be included as a facet and their variation could be assessed. Different collection designs (e.g., crossed vs. nested), and different number

of raters could be manipulated to assess the degree of dependability of ratings given by randomly similar raters in the universe of generalization.

Many-Facet Rasch Model (MFRM)

Many-Facet Rasch Model (MFRM) by Linacre (1989) is an extension of the Rasch Model (1960). It shares the same assumptions of the other forms of Rasch models (e.g., unidimensionality, and local independence), yet it adds the flexibility of treating other facets beyond that of the item difficulty (e. g., rater facet). It is a unidimensional and compensating mathematical model that places scores of the object of measurement (e.g., examinee) along with the other facets (e.g., item difficulty, rater severity) on the same scale which allows conveniently for separate and invariant parameter estimates, and equal-interval scale for the measurement. Conveniently, all Rasch models consider the total scores as a sufficient statistics. However, the attainment of such properties is conditional on securing a good model fit.

MFRM components

The MFRM components (Linacre, 1994) are:

$$\text{Log}\left(\frac{P_{nijk}}{P_{\text{nijk-1}}}\right) = B_n - D_I - C_j - F_k$$

Where

P_{nijk} is the probability of examinee n being graded on item I by the rater j a rating of k.

$P_{\text{nijk-1}}$ is the probability of examinee n being graded on item I by the rater j a rating of k-1.

B_n is the ability of examinee n

D_I is the difficulty of item i.

C_j is the severity of judge j.

F_k is the difficulty of the step up from category k-1 to category k and k=1, M.

MFRM analysis typically provides some useful statistics (e.g., reliability index, infit and outfit) which conveniently allowed the investigator to judge how reliable the variation in each facet in general, and to judge the consistency of elements in particular (i.e., consistent pattern of responses). Due to their importance they are introduced briefly.

Rasch models make strong assumptions, and as a result resultant fit statistics are judged for how off they are from their expectations (Bond & Fox, 2007). Outfit statistics are the residuals that has been standardized and averaged over persons and items. Infit statistics on the other hand are standardized residuals weighted by their respective variance (Bond & Fox, 2007, Linacre, 2002). In other words, infit statistics influenced by unusual pattern of responses to items close in difficulty to the person estimated ability, whereas outfit statistics are influenced by those far from either item or persons estimates.

These important fit statistics are usually in the form of a mean square (cf., standardized fit statistics). Put differently, it is a chi-square statistic divided by degrees of freedom which measures randomness in the data that is how predictable or unpredictable the data are (Linacre, 2002). As Linacre stated, the expected value is 1.00 and values less than one indicates less randomness and less useful information whereas values larger than one indicates unpredictability in the data. Acceptable range of productive Mean Square statistics is 0.5 to 1.5, and values larger than 2.00 distort the measurement (Wright & Linacre, 1994). [1]

Similar to the Cronbach's alpha's range from .00 to 1.00, people's reliability index measures the degree to which the persons estimates will reproduce when given a

[1] Linacre (1989) states that the acceptable range is from 0.8 to 1.2 and Fox and Bond (2007) state that the acceptable range for judged performance is from 0.4 to 1.2

similar sample of items (Bond & Fox, 2007). It is a quantification of how much of the observed variance is genuine (Wright & Masters, 1992). Similarly, passage-separation reliability assesses the degree to which the passages estimates will reproduce when given to a similar sample of persons. In other words, it indicates how much of the variation in the passages is true and not error of measurement. Yet another important reliability index is the separation index. Unlike the separation reliability which is bounded between zero and one and thus non-linear, the reliability separation index is the square root of the proportion of the adjusted variance of the sample to its squared average error of measurement (Bond & Fox, 2007; Stone, 2002). In other words, it is the division of true standard deviation of a sample by its standard error of measurement, and the resultant estimate indicates how many levels of the respective facet can be distinguished reliably (Wright, 1996).

Another attractive feature of MFRM is the ability to provide standard error of calibration for all elements in the testing situation (e.g., raters, subjects) which judges the degree to which the estimates are precise (Stone, 2002).

This ability to focus on specific elements in a facet is an advantage over traditional measurement models. Specifically, although ANOVA-based models would quantify the variance of a group in a facet (e.g., group of judges), they fail to pinpoint which particular element (i.e., a rater) is contributing more or less to such variance. Consequently, no adjustment could be done (Lunz et all. 1994). As previously stated, Given that raters significantly differ even after extensive training, their performance needs to be modeled and properly assessed; in order to achieve invariant measures of examines (Linacre, 1994).

Scoring Methods

To rate a performance, a scoring system must be used to convert judgments into scores, and this scoring system needs to be useful in guiding the grading process. Put differently, the scoring scheme should yield ever possible reliable ratings; that is reproducible. Two commonly used scoring methods are holistic and analytic. In the analytic scoring schemes the quality being evaluated or the construct being measured is divided into criteria that are rated separately and may be summed to give one index (Carr, 2000; Chi, 2001; Klein et al, 1998; Nitko, 2004). On the other hand, holistic scoring methods are used to rate the overall outcome without giving separate scores for each part or criterion, and hence result in only one global index (Carr, 2000; Chi, 2001; Klein et al, 1998; Nitko, 2004).

Holistic Methods Pros and Cons

Obvious advantage of using holistic scoring methods is the practicality. They are easier and quicker to construct and use, and consequently they lower the cost of assessment (Carr, 2000; White, 1984). Such an advantage makes the holistic schemes the preferred methods of scoring in large-scale assessments. Second advantage is the claim that some constructs are rather to be assessed on the whole. Following Gestaltism that the whole is more than its parts, some constructs are claimed to be more and beyond the parts (e.g., Klein, 1998; white, 1984).

However, holistic scoring methods have received much criticism. An objection to the holistic scoring is the (claim) argument that giving only one global summary index makes the holistic scoring methods of less value for diagnostic and formative purposes (White, 1984). Specifically, it fails to pinpoint the strengths and weaknesses of the

responses (Hamp-Lyons, 1995). In addition, the lack of justification for the resulting value might pose a challenge to the raters (Klein et al, 1998), and with the advancement in online-scoring collaborative raters might find it harder to work with holistic rubric. Equally important, the opponents of holistic scoring method maintain that its resultant ratings are more prone to rater's subjectivity.

Analytic Scoring Methods

Having more than one subcategory is seen as an answer to the shortcomings of holistic methods to provide diagnostic information (White, 1984). Yet, this complicates the scoring process by increasing the unwanted two; cost and time (White, 1984). Another disadvantage of using analytic, however, is the difficulty of constructing appropriate subcategories or criteria (Carr, 2000; White, 1984).

Reliability and Consistency of Analytic and Holistic Scores

Securing proper quality of any psychometric assessing procedure is a priority that increases in importance as the stakes grow high. Of course, scoring rubrics as they guide the evaluation process are not exempted from such requirements. Reliability and validity are needed to be assessed for the resultant ratings. However, discussion of the validity is beyond the scope of the present study and not to be discussed. Only the reliability aspect of quality is reviewed.

Two different terms are used for assessing reliability of rubrics-derived ratings; consistency and consensus. Consistency is a measure of how two raters or perhaps one rater assign scores of same rank-ordering, and consensus is measure of how two independent raters or more give exact ratings.

The literature is not definite or decisive concerning the superiority of one scoring method over the other. Whereas some studies have shown that holistic scoring has led to higher Interrater reliability, other studies have concluded just the opposite. For instance, Olson (1988) found that holistic resulted in an Interrater reliability of (.73) compared to analytic (.66). And likewise, Voskuijl and Sliedregt (2002) found that holistic produced higher reliability estimates when compared to analytic.

On the other hand, other studies concluded that analytic scoring methods result in higher reliability indices. Of relevance to the current study is the work by Chi (2001) who found higher separation reliability associated with the analytic scoring for the students' facet, and that the analytic scoring yields more consistent severity of raters when compared to holistic using MFRM. This was also supported by the findings that holistic scoring yielded ratings with higher reliability separation for the raters. Similarly, Alharby (2006) found that students' facet was associated with a higher reliability of separation. However, holistic scoring rubrics within the context of ESL writing statistically outperformed the analytic on the overall fit. A comparison of generalizability coefficients showed no statistical differences, though analytic produced higher estimates.

What Factors Moderate Reliability Estimates?

One way to understand why raters would find some scoring techniques easier than others lies in the nature of the scoring rubrics, the type of the task being rated, the quality of training, and the skills and experiences of the raters. A 2002 meta-analysis study by Voskuijl and Sliedregt on job analysis inter-rater reliability showed that rating of raters sufficiently experienced with rating scales, ratings based on detailed-scales, and ratings

based on delineated jobs yielded higher inter-rater reliability. Although, they suggested using highly proficient raters in order to have higher inter-rater reliability; they stated that thorough training could bring about reliable ratings from nonprofessionals as high as that of trained professionals under specific circumstances.

On the contrary to the literature on judgmental evaluation as Voskuijl and Sliedregt (2002) pointed out, the holistic scoring outperformed the decomposed counterparts (i.e., analytic). The reason for such findings, they suggest, is that analytic scoring might be sometimes tedious and time-consuming task for the raters (Cornelius & Lyness, 1980, as cited in Voskuijl & Sliedregt, 2002; Sanchez & Levine, 1994, as cited in Voskuijl & Sliedregt, 2002). Such findings might highlight the different functioning of the scoring rubrics and their suitability to some tasks but to the exclusion of others. In other words, analytic scoring might be suitable to one task while holistic is befitting another.

In another different meta-analysis study on job analyses, Dierdorff and Wilson (2003) found a higher inter-rater reliability for ratings on specific versus general tasks, although the opposite was also found when the number of raters and items was small. Their explanation was that specific tasks are easier to rate and interpret than general ones. However, such superiority of task specific ratings might be merely a result of the reliability assessment methods used (i.e., repeated item vs. rate-rerate), they noted.

Type of raters is suggested as a source of moderating factors of inter-rater reliability. Prior literature indicates raters from different sources are at variance when rating a performance assessment (Facteau & Craig, 2001). Such disagreements coupled

with extreme ratings lead to narrower range of ratings, thus lower inter-rater reliability coefficients (Murphy & DeShon, 2000).

In summary, scoring method, experience with rating, general versus specific tasks, nature of the scoring rubric and training are found to have potential impact on reproducibility of ratings.

Oral Assessment

There are times and situations where oral performance is the artifact or the outcome, and only direct measurement of that outcome is the most appropriate. If forms of assessments do not simulate the real life performance, and do not capture the essence of the quality, validity of such measurements is at risk.

Oral performance seems to be in this category. Oral performance arguably has distinguishing characteristics from that of written formats, and such distinctiveness might require different considerations when evaluating performance of students in Quran subject. Specifically, the rater cannot in many cases conceal the performer's identity, visit the performance again and again for assurance of good quality ratings, rate one task at a time, or compile the responses in deferring groups based on quality before the real grading can take place. Therefore, it is impractical to use the guidelines suggested for written-format contexts to handle the subjectivity of the rater (e.g., Linn & Gronlund, 2000).

Nevertheless, scoring of oral performance and the effectiveness of alternative scoring methods (Quran in particular) seem to receive too little attention if any. Rather, most of the literature on usefulness of the scoring methods is found in the context of assessments that require written responses (Klein et al, 1998). Given that backdrop and

for the above-mentioned reasons, there is a need for a study that takes on the task of contrasting two scoring techniques; holistic vs. analytic within the context of Quran recitation.

Closing the Gap

Performance assessment is increasingly being used and utilized as a direct measurement of the trait or the skills under consideration. Yet, the limitations (e.g., rater and task errors) of such tools and procedures cannot be overstated (e.g., Haris, 1997). They need to be painstakingly handled and appropriately addressed in order to obtain scores or ratings that are dependable. One important factor toward achieving trust worthy scores is a measurement model that is more sensitive, versatile, and effective and at the same time yielding statistics that could generalize to future settings. This explains well the move toward and the interest in MFRM and G-Theory as they possess such qualifications. Another but equally important factor is the scoring rubrics as they guide the rater and examinee behaviors. Such scoring rubrics although useful, they reveal different aspects of the performance, and yield different scores of different consistency. Suitability of such scoring rubric to one context or another requires a bit of philosophical and empirical reasoning, and this study seeks to close this gap by exploring the best scoring rubrics in the context of oral assessment using MFRM and G-Theory.

Chapter Three

Sample

Data were collected from a sample of high-school and junior-high school students in Riyadh (n = 93). Schools of Riyadh School District were stratified and a random sample of students was then selected. Approximately one third of the students included in the study came from junior-high school and the other two third came from high-school.

The raters who agreed to rate the students' recitations were very skilled in Quranic rules of recitation. Two of the raters have a post-Bachelor's level of education, have attended additional and special courses for Quranic recitation and have taught Quran for non-profitable organizations. The third rater has a license with a continuous chain of narrators to the prophet Muhammad PBUH.

Tasks

To be assessed for quality of reading and implementation of ruling of recitation, the students were asked to recite three different segments of the holy Quran and their recitations were audio-taped. The average time of each segment was approximately 2:15 minutes. These three segments were chosen based on several consultations with a teaching faculty in King Faisal University and various teachers of Quran. They were selected for their assumed differing difficulties, abundance of Quranic rulings and familiarity. The three selected Quranic passages were from the chapters of Fussilat, Maryam and Nooh.

Raters independently rated each student's recitation using analytic and holistic (279 Passages) with an average interval time of two weeks. The raters were not counterbalanced to prevent carryover effects of methods.

Students were asked to recite predetermined segments or passages of the Quran, and three raters rated the students' performance for each of the three segments selected.

Instrumentations

The researcher found no existing rubrics for Quranic recitation. The job of creating the rubrics was shared with a college faculty specialized in Quranic Studies to construct two rating scales; analytic and holistic. Comments and feedback were accumulated from teachers of Quran.

Both scoring rubrics ranged from 0 to 4, with 0 as the lowest category and assigned to the weakest performance and 4 indicates proficiency. The holistic rubric requires the rater to listen to the recitations and give one comprehensive score representing how proficient the students are on five dimensions. However, the analytic rubric for Quranic recitations requires a separate score on each dimension. The dimensions of the analytic scale were:

1. Correct language.
2. The ruling of vowel-free Meem and Noon and Tanween.
3. The ruling of Stressing.
4. The ruling of lengthening.
5. The ruling of vibration.

It was assumed that correct use of language is the first to be mastered and that the ruling of vibration would be less mastered. Such hierarchy can be verified by the ruler map provided by FACETS.

Procedures

After collecting the audio-taped recitations of segments of the Holy Quran selected based on differing difficulties, familiarity and wealth of Quranic rulings by a college faculty specialized in Quranic Studies, raters were trained on using the scales prior to actual rating session. Due to geographical reasons, training sessions were administered individually or over the Internet. The purpose of creating and using rubrics was presented, and the structure of each scoring rubrics (i.e., analytic and holistic) was explained. Trial ratings were provided and feedback and questions were discussed thoroughly. Random trial ratings were compared to check for common understanding of the scale or questions about the scale. After showing a satisfactory understanding and implementation of the scales, the raters then proceeded to begin the actual rating.

The observations were analyzed using Genova version 2.1 (Crick &Brennan, 1993) and FACETS (Linacre and Wright, 1994) applications. Genova is an applications used for G-theory estimates and FACETS is for MFRM. In the next chapter, results for the FACETS and GENOVA are presented, respectively. For FACETS, construct or variable maps help to interpret quality of scaling. For GENOVA, variance component estimation summarizes different design studies to consider the optimal frequency and type of facets to ensure dependability of scores. Both scales were analyzed for their appropriateness, and the resulting raw and scaled scores (i.e., ratings by raters using both

rubrics) were correlated using (Pearson's r and Spearman). Mean scores of both methods were tested for their statistical differences (paired t-test).

Chapter Four

Results

Descriptive Statistics.

This section presents descriptive statistics of raters' correlations, and holistic and analytic scoring rubrics. Inspection of the correlation matrix of raters using holistic scoring rubrics to evaluate recitations of students revealed highly correlated ratings. Table 4.1 shows that rater 1 and 2 had the higher correlational value (0.96) and that rater 1 and 3 had the lowest Pearson' r value (0.90).

Table 4.2 gives a matrix of Pearson's r correlations between raters using analytic scoring rubrics to evaluate the students' recitations. Still, rater 1 and 2 show the highest correlational estimate (0.98) and rater 2 and 3 showed the lowest estimate (0.94). These estimates of analytic correlations might suggest that rank ordering of students was more consistent than that of holistic. In other words, the use of analytic scoring rubrics helped guide the raters to yield relatively similar ratings of students. This might give some indication to the benefits of introducing the analytic scoring rubrics to the practice of Quranic recitation assessment.

Table (4.1)
The Correlation Matrix for the Three Raters for Holistic Scoring Approach

Variables	Rater 1	Rater 2	Rater 3
Rater 1 Sig. (2-tailed)	1		
Rater2 Sig. (2-tailed)	0.962** .000	1	
Rater 3 Sig. (2-tailed)	0.902** .000	0.918** .000	1

**. Correlation is significant at the 0.01 level (2-tailed).

Table (4.2)
The Three Raters Correlation Matrix for Analytic Scoring Approach

Variables	Rater 1	Rater 2	Rater 3
Rater 1 Sig. (2-tailed)	1		
Rater2 Sig. (2-tailed)	.986** .000	1	
Rater 3 Sig. (2-tailed)	.949** .000	.943** .000	1

**. Correlation is significant at the 0.01 level (2-tailed).

Many-Facet Rasch Measurement Model Analysis

This section compares holistic to analytic scoring approaches for quality of fit. Observations were analyzed using the FACETS computer software, version 3.67 (Linacre, 2010). The result of holistic data analysis is discussed first followed by a discussion of analytic data analysis. Individual and overall comparisons of the two approaches conclude this section.

Holistic Data Analysis

Students Facet:

The results of the MFRM analysis for the facet for students show an overall good fit to the model. The students showed acceptable performance patterns in reciting the Quranic passages as evaluated by the raters. In other words, students' performance on the passages presented and evaluated by the raters behaves as expected by the model.

Quality control indices suggest a good fit as the individual INFIT and OUTFIT statistics fall within the recommended range between 0.5 and 1.5. A similar conclusion can be drawn for the overall INFIT statistics as the mean was 0.81, and the respective standard deviation was 0.59. In the same way, OUTFIT statistics fall within the recommended range, and exhibit an overall mean of 0.88 and a standard deviation of 0.90. Approximately, 35% of the data had INFIT statistics outside the recommended range of 0.5 and 1.5, suggesting that the data fitted the model reasonably well. However, 48% of the data had OUTFIT statistics outside the boundaries of the interval of 0.5 and 1.5 but with only 7 sets of responses resulting in statistics exceeding 2 ZSD.

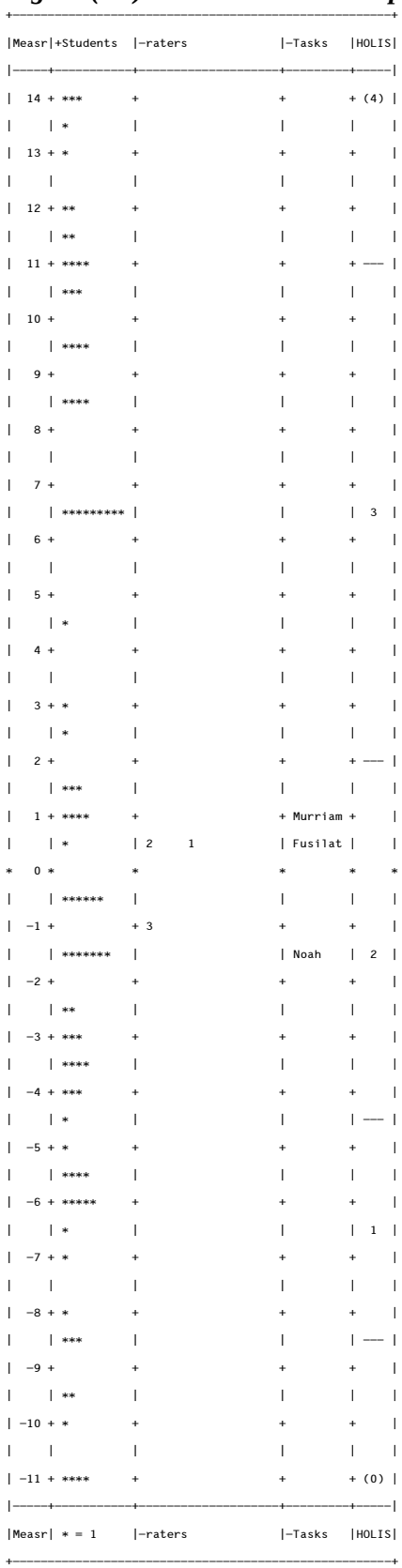
The calibrated students' measures span the logit scale of a minimum -12.47 to 15.00 with a mean of 1.08 and a standard deviation of 7.46. The student recitation ability

can be distinctively separated into 9.46 strata with a separation index of 6.84 and a reliability of 0.98. These differences in ability to recite properly were statistically significant as the respective Chi-square "fixed effect" value was significant χ^2 (92) = 6027.9; and p <.00001. The data analysis suggest a normal distribution of the students' scores as the Chi-square "random effect" value was not significant χ^2 (91) = 91.7, and p = 0.46.

Table (4.3) Holistic Rater Measures:

Observed raw score	Observed average	Rasch Logit Measure	Model SE	INFIT		OUTFIT		Rater ID
				Mean Square	Standardized	Mean Square	Standardized	
578	2.1	.73	.16	.89	-1.1	.98	0.00	Rater 1
595	2.2	.31	.16	.71	-3.2	.58	-2.7	Rater 2
651	2.4	-1.03	.15	1.10	1.00	1.08	.5	Rater 3

Figure (4.1) Holistic Variable Map



Raters Facet:

Table 4.3 shows MFRM analysis result for raters. The raters showed acceptable internal consistency in applying the holistic scoring when evaluating students' recitations. The INFIT statistics fall within the recommended range between 0.5 and 1.5. The overall INFIT statistics suggest a good fit to the model as the mean was 0.90, and the respective standard deviation was 0.19. In the same way, OUTFIT statistics fall within the recommended range, and exhibit an overall mean of 0.88 and a standard deviation of 0.26.

The calibrated measures of severity as seen from Table 4.3 show rater 3 as the most lenient (a logit of -1.03) and rater 1 as the most severe rater (a logit of .73). The mean for this facet was, as anticipated, 0.00, and the Standard Deviation was 0.92. To address the interchangeability of raters, the separation index was found to be 5.81 and the reliability index was 0.97, and that they can be reliably differentiated into 8.07 groups. The indices suggest that the raters are dissimilar in their leniency and can be placed differently along the logit scale, and the Chi-square "fixed effect" hypothesis that raters are the same was rejected χ^2 (2) = 69.8; and $p < 0.0001$. The Chi-square random effect suggests a random distribution of the data χ^2 (1) = 1.9; $p = 0.16$. To address whether raters assign values similar to those assigned by scoring machines or independent raters (Linacre, 2011), MFRM outputs reveals the raters were found to have had exact agreements of 528 (67.2%) which was in accordance with the model expected agreements of 508 (65.7%).

Task Facet:

Table (4.4) Holistic Passages Measures and Quality Control Statistics

Observed raw score	Observed average	Rasch Logit Measure	Model SE	INFIT		OUTFIT		Task ID
				Mean Square	Standardized	Mean Square	Standardized	
569	2.1	.94	.16	.92	-.8	.75	-1.4	Fusilat
593	2.2	.35	.16	.94	-.6	1.04	.2	Merriam
662	2.4	-1.29	.15	.84	-1.8	.85	-.8	Noah

Table 4.4 shows MFRM analysis result for passages. The INFIT statistics were all in the acceptable level between 0.5 and 1.5. These statistics examine the extent to which the passages were consistently evaluated. The overall INFIT statistics suggest a good fit to the model as the mean was 0.90, and associated with a standard deviation of 0.05 indicating a low spread of these statistics around their mean. In the same way, OUTFIT statistics fall within the recommended range, have an overall mean of 0.88 and a standard deviation of 0.15.

As seen from Table 4.4 calibrated measures of passages' difficulty were presented with passage 3 as the easiest (a logit of -1.29) and passage 1 as the most difficult (a logit of .94). The mean for this facet was, as anticipated, 0.00, and the standard deviation was 1.15.The separation index for the passage facet was 7.33 and the reliability index was 0.98. The indices suggest that the passages are of differing difficulties and hence different passages need different amount of Quranic skill to be recited properly. These passages can be reliably differentiated into 10.10 groups. The corresponding Chi-square "fixed effect" hypothesis that passages are the same was rejected χ^2 (2) = 109.9; $p < 0.00$. The

Chi-square random effect value suggests a random distribution of the data χ^2 (1) = 2; p = 0.16. Such findings give support to the hierarchy of difficulty of passages. It was assumed that Maryam would be the most difficult and Nooh would be the easiest.

Table (4.5) Holistic Rating Scale

Data				Quality Control			Rasch-Andrich Thresholds		
Category									
score	Used	%	Cum. %	Average meas.	Exp. Meas.	Outfit Meas.	Measure	S. E.	Rubric Labels
0	49	6%	6%	-9.06	-8.82	.7			Very Weak
1	134	17%	24%	-5.81	-5.94	1.1	-8.42	.22	Fair
2	265	34%	58%	-1.73	-1.66	.7	-4.67	.16	Good
3	252	33%	90%	6.91	6.88	.9	1.92	.19	V. good
4	47	10%	100%	11.92	11.87	.8	11.17	.19	Excellent

Figure 4.3 shows the probability characteristic curves for the five scoring categories (i.e., 0-4) of the holistic rating scale employed by the raters to appraise the recitations. The graph shows to some extent smoothed and well-behaved curves. Category 2 and 3 were matched up with the center of the calibrated-measure distribution, and least able students are assigned category 0, and high performing students were assigned category 4 as it peaks high toward the high end of the distribution. However, category 1 has a restricted range, and category 3 has appeared to spread over a rather large area of the logit scale continuum. This spread of category 3 might have resulted from category 1 not being used as much as expected, or perhaps that additional category between 3 and 4 is needed to better differentiate between students with higher Quranic

proficiency. Category 3 was probable from logit 1.92 to approximately logit 11.17. This large area on the logit continuum could be better evaluated by adding another category. Likewise, adding a category between 2 and 3 might result in a better functioning of the scale. On the whole, the rating scale was used reasonably well.

As seen from Table 4.5 the rating scale was used in accordance with its intended purpose. All response categories were sufficiently used and they have a range of logit estimates from -9.06 to 11.92. All reported OUTFIT statistics were well above 0.5 and below 1.5 and the expected measures by the model matched the respective observed average measures.

Figure (4.2) Holistic CI for Measure Relative to Item Difficulty

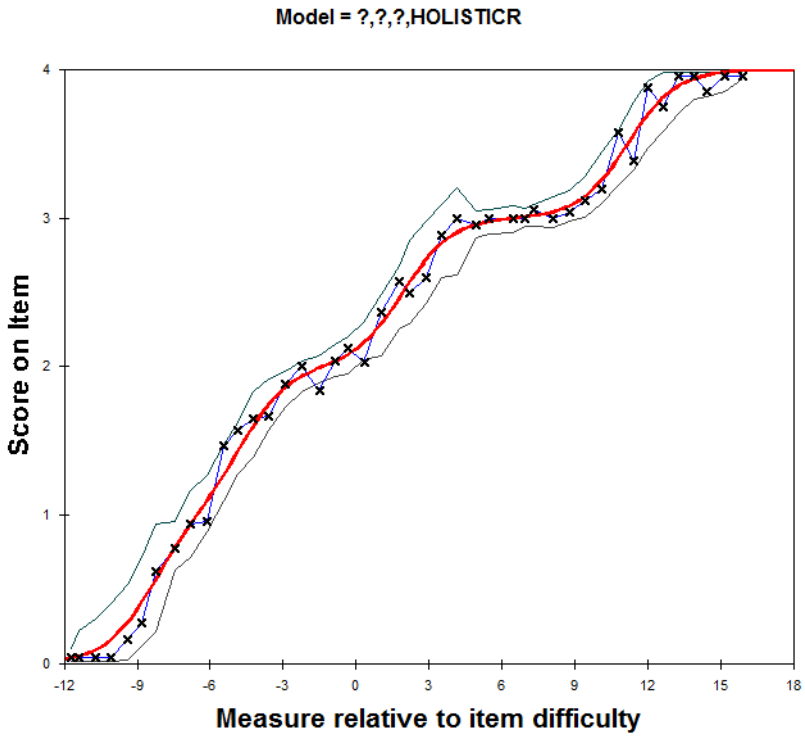
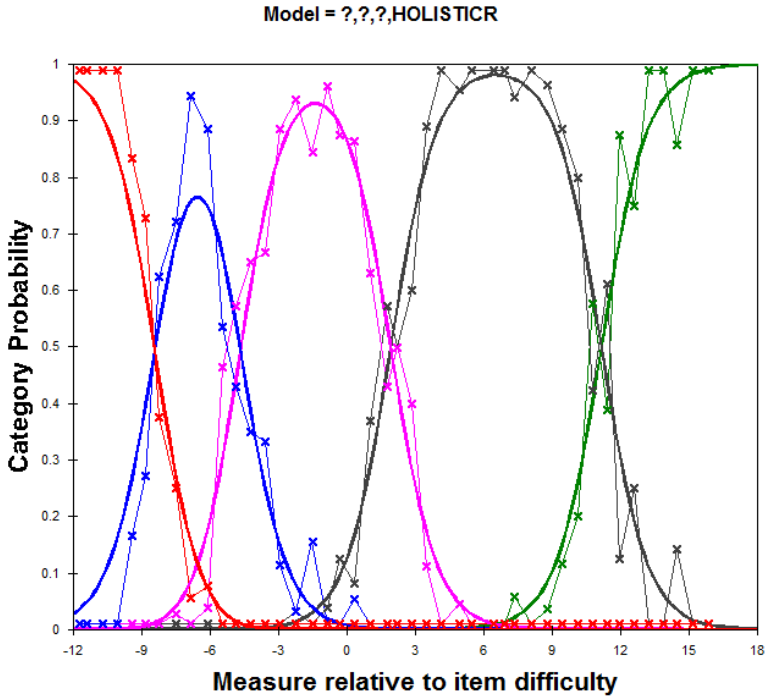


Figure (4.3) Holistic Rating Scale Empirical vs. Expected



Analytic Scoring Approach

Students' Facet:

The results of the MFRM analysis for the person facet show an overall good fit to the model. The students showed acceptable performance patterns in reciting the Quranic passages as evaluated by the raters. In other words, students' performances on the passages presented and evaluated by the raters matched model expectations.

The quality control indices for the student facet as evaluated by the analytic rubric suggest a good fit as the individual INFIT and OUTFIT statistics fall within the recommended range between 0.5 and 1.5. A similar conclusion can be drawn for the overall INFIT statistics as the mean was 0.98, and the respective standard deviation was 0.26. In the same way, OUTFIT statistics fall within the recommended range, and exhibit an overall mean of 0.97, and a standard deviation of 0.30.

Approximately, 6% of the data had INFIT statistics outside the recommended range of 0.5 and 1.5, suggesting that the data fit the model very well. Similarly, 9% of the data had OUTFIT statistics outside the boundaries of the recommended range of 0.5 and 1.5 but with only 7 observations resulting in statistical values exceeding 2 ZSD.

The calibrated students' measures span a logit scale of a minimum -8.08 to 7.97 with a mean of 1.57 and a standard deviation of 3.82. The student recitation proficiency can be distinctively separated into 6.8 strata with a separation index of 4.85 and a reliability of 0.96. Chi-square "fixed effect" was rejected χ^2 (92) = 5869.7; p <.0001. The data analysis suggest a normal distribution of the students' scores as the resulting Chi-square "random effect" value was not statistically significant, χ^2 (91) = 83.2; p =.71.

Rater Facet:

Table (4.6) Analytic Raters Measures and Quality Control Statistics

Observed raw score	Observed average	Rasch Logit Measure	Model SE	INFIT		OUTFIT		Rater ID
				Mean Square	Standardized	Mean Square	Standardized	
3442	2.4	.17	.05	.90	-2.4	.93	-1.2	Rater 2
3517	2.5	.05	.05	1.18	3.8	1.12	1.7	Rater 3
3660	2.6	-.22	.05	.91	-2.2	.88	-1.8	Rater 1

Figure (4.4) Analytic Variable Map

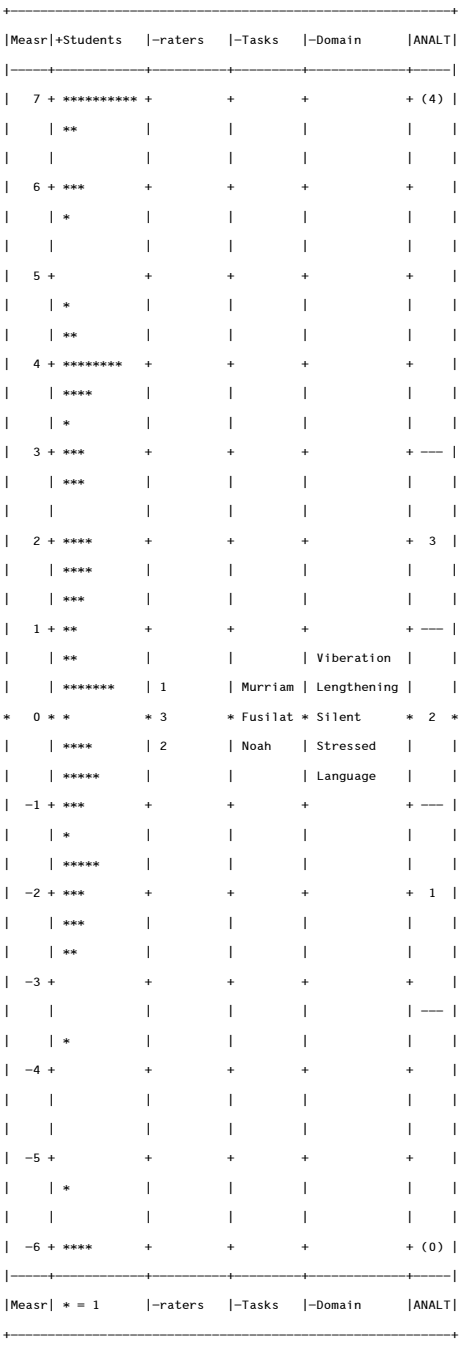


Table 4.6 shows MFRM analysis result for raters. The raters showed acceptable internal consistency in applying the analytic scoring for evaluating students' recitations. The INFIT statistics fall within the recommended range between 0.5 and 1.5. The overall INFIT statistics suggest a good fit to the model as the mean was 0.99, and the respective standard deviation was 0.16. In the same way, OUTFIT statistics fall within the recommended range, and exhibit an overall mean of 0.97 and a standard deviation of 0.13.

The calibrated measures of severity as seen from Table 4.6 show rater 1 as the most lenient (logit of -0.22) and rater 2 as the most severe rater (a logit of 0.17). The mean for this facet was, as anticipated, 0.00, and the Standard Deviation was 0.20.

To address the extent to which raters can be considered interchangeable, the separation index was found to be 4.18 and the reliability index was 0.95, and that raters can be reliably differentiated into 5.91 groups. The Chi-square "fixed effect" hypothesis that raters are the same was rejected χ^2 (2) = 36.8; $p < 0.00$. The indices suggest that the raters are dissimilar in their leniency and can be placed differently along the logit scale. The Chi-square random effect suggests a random distribution of the data χ^2 (1) = 1.9; p =.17. To address the extent to which raters' scores are like those assigned by scoring machines or independent raters (Linacre, 2011), MFRM outputs reveal that the raters had exact agreements of 1973 (55.6%) and the model expected agreements were 1752.9 (49.4%).

Task Facet:

Table (4.7) Analytic Passages Measures and Quality Control Statistics

Observed raw score	Observed average	Rasch Logit Measure	Model SE	INFIT		OUTFIT		Task ID
				Mean Square	Standardized	Mean Square	Standardized	
3378	2.3	. 35	.05	.97	-.6	.94	-1.00	Merriam
3539	2.5	.00	. 05	.93	-1.6	.89	-1.8	Fusilat
3702	2.6	-.35	. 05	1.08	1.8	1.09	1.2	Noah

Table 4.7 shows MFRM analysis result for passages. The INFIT statistics were all in the acceptable level between 0.5 and 1.5. These statistics examine the extent to which the passages were consistently evaluated. The overall INFIT statistics suggest a good fit to the model as the mean was 0.99, and associated with a standard deviation of 0.08, indicating a low spread of these statistics around their mean. In the same way, OUTFIT statistics fall within the recommended range, have an overall mean of 0.97 and a standard deviation of 0.11.

As seen from Table 4.7 calibrated measures of passages' difficulty were presented with passage 3 as the easiest (a logit of -0.35) and passage 1 as the most difficult (a logit of 0.35). The mean for this facet was, as anticipated, 0.00, and the standard deviation was 0.35.The separation index for the passage facet was 7.47 and the reliability index was 0.98. The indices suggest that the passages are of differing difficulties and hence different passages need different amount of Quranic skill to be recited properly. These passages can be reliably differentiated into 10.3 groups. The corresponding Chi-square "fixed effect" hypothesis that passages are the same was rejected χ^2 (2) = 113.6; $p < 0.00$. The Chi-square random effect value suggests a random distribution of the data χ^2 (1) = 2;

p = 0.16. The findings here again give support to the hierarchy of difficulty of passages with chapter Maryam as the most difficult and Nooh as the easiest.

Table (4.8) Analytic Sub-Domain Measures

Observed raw score	Observed average	Rasch Logit Measure	Model SE	INFIT		OUTFIT		Sub-Domain
				Mean Square	Standardized	Mean Square	Standardized	
1914	2.2	. 75	.06	1.36	6	1.23	3.1	5
2048	2.4	.28	.06	.96	-.7	1.10	1.3	4
2141	2.5	-.05	.06	.72	-5.4	.74	-3.3	2
2253	2.7	-.47	.06	.87	-2.3	.89	-1.1	3
2263	2.7	-.51	.06	1.04	.6	.90	-.9	1

Table 4.8 shows MFRM analysis result for the sub-domains. The sub-domains included in the study are correct language, ruling of Meem and Noon and Tanween, stressing, lengthening and vibration. The INFIT statistics were all in the acceptable level between 0.5 and 1.5. These statistics examine the extent to which the domains were consistently evaluated. The overall INFIT statistics suggest a good fit to the model as the mean was 0.99, and associated with a standard deviation of 0.24, indicating a low spread of these statistics around their mean. In the same way, OUTFIT statistics fall within the recommended range, have an overall mean of 0.97 and a standard deviation of 0.19.

Table 4.8 shows sub-domain measures of difficulty with “language correctness” 1 as the easiest sub-domain (a logit of -0.51) and “vibration” 5 as the most difficult sub-domain (a logit of 0.75). The mean for this facet was, as anticipated, 0.00, and the standard deviation was 0. 53. To address the extent to which these sub-domains are similar, the separation index was 8.69 and the reliability index was 0.99. The indices

suggest that the sub-domains can be reliably differentiated into 11.92 groups. The Chi-square "fixed effect" hypothesis that passages are the same was rejected χ^2= (4) = 309.3; p = 0.00. The Chi-square random effect suggests a random distribution of the data χ^2 (4) = 3.9; p = 0.27).

Such ordering of dimensions is in line with the belief that correct language should be mastered first. The vibration dimension is less emphasized in schools and is believed to be mastered by students as they become more proficient in recitations and hence to be found as the most severe dimension.

Figure 4.5 shows the probability characteristic curves for the five scoring categories (i.e., 0-4) of the analytic rating scale employed by the raters to appraise the recitations. The graph shows very smoothed and well-behaved curves. Category 2 was matched up with the center of the calibrated-measure distribution, and least able students are assigned category 0. Categories 1 and 3 were spread smoothly and functioned similarly as category 2. High performing students were assigned category 4 as it peaks high toward the high end of the distribution. A possible addition of categories between 0 and 1, and 3 and 4 might improve the scale functioning. Overall, the analytic rating rubric was endorsed very well.

As seen from Table 4.9 the rating scale was used in accordance with its intended goal. All response categories were sufficiently used and they had a range of logit estimates from -2.82 to 4.18. All reported OUTFIT statistics were around 1.00 and the expected measures by the model matched the observed average measures.

Table (4.9) Analytic Rating Scale

Data				Quality Control			Rasch-Andrich Thresholds		
Category									
score	Used	%	Cum. %	Average meas.	Exp. Meas.	Outfit Meas.	Measure	S. E.	Rubric labels
0	257	7%	7%	-2.82	-2.82	1.1			Very Weak
1	641	18%	25%	-1.34	-1.32	1.00	-2.92	.08	Fair
2	834	23%	49%	.04	.08	.9	-.92	.06	Good
3	782	22%	71%	2.07	1.97	.9	1.04	.06	V. good
4	1041	29%	100%	4.18	4.22	1.1	2.80	.06	Excellent

Figure (4.5) Analytic scale cateories

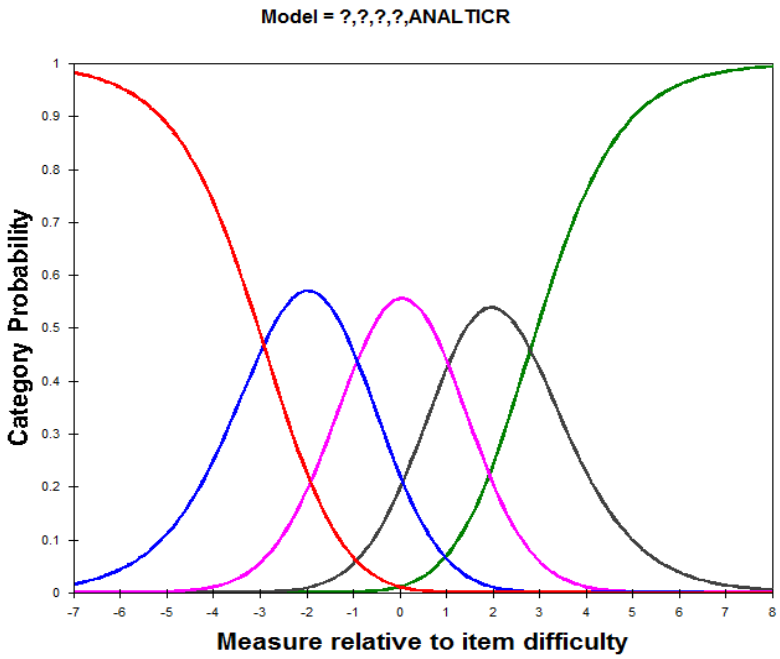


Figure (4.6) Analytic CI for the Measure Relative to Item Difficulty

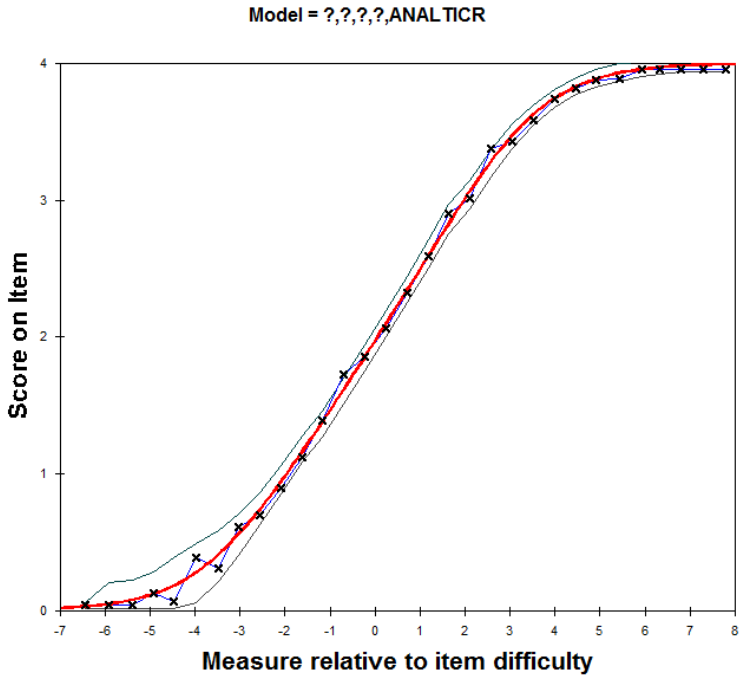


Figure (4.7) Analytic Scale Categories Empirical vs. Expected

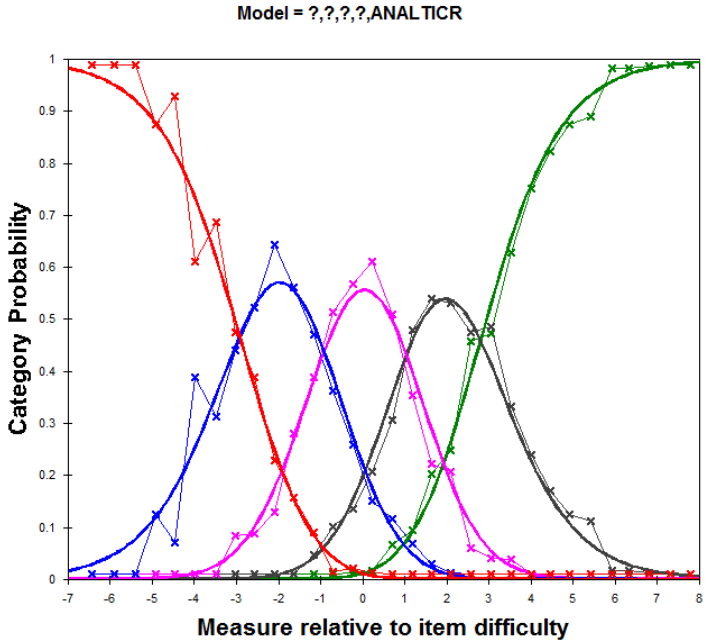


Table (4.10) The Correlation Matrix for Holistic and Analytic Raw Scores

Variables	Holistic Raw Score	Analytic Raw Score	Holistic Measure Score	Analytic Measure Score
Holistic Raw Score Sig. (2-tailed)	1			
Analytic Raw Score Sig. (2-tailed)	.942** .000	1		
Holistic Measure Score Sig. (2-tailed)	.988** .000	.928** .000	1	
Analytic Measure Score Sig. (2-tailed)	.933** .000	.956** .000	.933** .000	1

** Correlations are statistically significant at 0.001

Relationship between Scores

This section assesses the relationship of both raw and calibrated scores of holistic and analytic rubrics. Correlation of holistic raw scores with their calibrated measures produced a high Person's ρ value ($\rho = 0.988$ and $p < 0.0001$), suggesting that both scores have a shared variance of larger than 97%. Correspondingly, the correlation between analytic raw scores with their measure estimates resulted in a very high value ($\rho = 0.956$ and $p < 0.0001$), but slightly lower than that of holistic. This value shows that 91% of the variance in raw scores is shared with the calibrated measures counterparts. It might also, suggest that comparably a larger adjustment has been made to the analytic raw scores to objectively evaluate students' recitation.

Evaluation of the relationship of raw scores of holistic and analytic reveals very high estimates ($\rho = 0.942$ and $p < 0.0001$), indicating that, on average, rank ordering of students is consistent across the scoring rubrics. Similarly, the Pearson's correlation estimate between holistic calibrated measures and analytic calibrated measures is high as well ($\rho = 0.933$ and $p < 0.0001$). Notably, such high correlational estimates between analytic and holistic scores might suggest either scoring approach would yield very similar ratings if order is all that matter.

To test for differences in mean scores of analytic and holistic rubrics, a paired sample t-test was performed. Table 4.11 gives the means and standard deviations of both scoring rubrics, and shows that analytic raw and calibrated scores exhibited higher means than for holistic scores. However, only raw scores showed statistically significant differences ($p < 0.01$). Finding higher means for the analytic approach agrees with previous research results (e.g., Alharby, 2006; Goulden, 1994) and contrast with Chi's finding (2004), which showed statistically higher mean scores for the holistic.

Table (4.11) Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Holistic Mean Score	2.1789	93	1.07184	.11114
	Analytic Mean Score	2.5371	93	1.23898	.12848
Pair 2	Holistic Measure Score	1.0762	93	7.45424	.77297
	Analytic Measure Score	1.5667	93	3.81874	.39598

Table (4.12) Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	Holistic Raw Score and Analytic Raw Score	93	.942	.000
Pair 2	Holistic Measure Score and Analytic Measure Score	93	.933	.000

Table (4.13) Paired Samples Test

		Paired Differences		
		Mean	Std. Deviation	Std. Error Mean
Pair 1	Holistic Raw Score and Analytic Raw Score	-.35817	.42787	.04437
Pair 2	Holistic Measure Score and Analytic Measure Score	-.49043	4.12757	.42801

Table (4.14) Paired Samples Test

		df	Sig. (2-tailed)
Pair 1	Holistic Raw Score and Analytic Raw Score	92	.000
Pair 2	Holistic Measure Score and Analytic Measure Score	92	.255

Holistic Versus Analytic Individual Indices Comparison

This section compares the Quranic scoring rubrics on their individual facets.

Table (4.15) Students' Facet

Rubrics	Separation index	Reliability	Infit Statistics		Outfit Statistics	
			Mean	S.D	Mean	S.D
Holistic	6.84	0.98	0.81	0.59	0.88	0.90
Analytic	4.85	0.96	0.98	0.28	0.97	0.30

For the students' facet, the analytic scoring rubric produces better overall Infit and Outfit indices (c.f. Table 4.15) in that they are closer to the expected value of one (0.98 and 0.97 vs. 0.81 and 0.88, respectively). These Infit and Outfit indices show less deviation from their means (0.28 and 0.30 vs. 0.59 and 0.90, respectively). However, the holistic scoring rubric reveals better separation and reliability indices. Such a result contrasts the findings of previously cited work on analytic vs. holistic (c.f., Alharby, 2006; Chi, 2004).

For the rater facet, the analytic scoring approach outperforms its counterpart on all levels. As seen from Table 4.16, Infit and Outfit values give preference to analytic over holistic (0.99 and 0.97 vs. 0.90 and 0.88, respectively). Deviation of the Infit and Outfit figures from their expected value was higher for the holistic scoring rubric (0.16 and 0.13 vs. 0.19 and 0.26, consecutively). The analytic rubric shows lower separation index (4.85 vs. 6.84), this gives more support to the use of the analytic scoring rubric as it guides the raters to give more consistent ratings. The raters using the analytic rubric show to some extent a tendency to assign scores as used in computer-program machines which is desirable for school settings (observed agreement of 55.6% vs. an expected of 49.4%).

Lastly, Table 4.17 gives MSQ Infit and Outfit statistics for raters, and they show on average better fit indices for the analytic scoring rubric.

Table (4.16) Raters Facet

Rubrics	Separation index	Reliability	Rater Agreement		Infit Statistics		Outfit Statistics	
			Obs.	Exp.	Mean	S.D	Mean	S.D
Holistic	5.81	0.97	67.2%	65.7%	0.90	0.19	0.88	0.26
Analytic	4.18	0.95	55.6%	49.4%	0.99	0.16	0.97	0.13

Table (4.17) Analytic vs. Holistic Individual Raters' Fit Indices

Rater ID	Holistic		Analytic	
	Infit	Outfit	Infit	Outfit
1	0.89	0.98	0.91	0.88
2	0.71	0.58	0.90	0.93
3	1.00	1.08	1.18	1.12

For the passage facet, Table 4.18 shows a slightly higher separation index for the analytic scoring rubric (7.47 vs. 7.33) suggesting that the analytic scoring approach better detects differences in passages and differentiates them along the Quranic recitation continuum. Such a result can be useful if the intention of testing is to include passages with different difficulties to better evaluate students' Quranic performance along a continuum. Infit and Outfits indices of the analytic scoring rubric are closer to their expected values (0.99 and 0.97 vs. 0.90 and 0.88, respectively). However, the spread of the Infit-statistic was slightly higher for the analytic than for holistic (0.05 vs. 0.08). Table 4.19 gives MSQ Infit and Outfit statistics for each passage, and they show on average better fit indices for the analytic scoring rubric.

Table (4.18) Passage Facet

Rubrics	Separation index	Reliability	Infit Statistics		Outfit Statistics	
			Mean	S.D	Mean	S.D
Holistic	7.33	0.98	0.90	0.05	0.88	0.15
Analytic	7.47	0.98	0.99	0.08	0.97	0.11

Table (4.19) Analytic vs. Holistic Passage Fit Indices

Passage Label	Holistic		Analytic	
	Infit	Outfit	Infit	Outfit
Fusilat	0.92	0.75	0.97	0.94
Merriam	0.94	1.04	0.93	0.89
Noah	0.84	0.85	1.08	1.09

Although both rating scales show acceptable statistics (e.g., Outfit statistics between 0.5 and 1.5), the analytic scoring rubrics reveals better individual fit indices. Specifically, as seen from Table 4.20 the outfit statistics were higher and closer to the expected value (i.e., 1.00), and the observed average measures of each category match very well their expected values. These statistics taken together might give more preference for analytic scoring rubrics of Quranic recitations to the holistic rubrics.

Table (4.20) Rating Scales

Category	Holistic Quality Control Measures			Analytic Quality Control Measures		
	Ave.	Exp.	Outfit	Ave.	Exp.	Outfit
0	-9.06	-8.82	0.7	-2.82	-2.82	1.1
1	-5.81	-5.94	1.1	-1.34	-1.32	1.00
2	-1.73	-1.66	0.7	0.04	0.08	0.9
3	6.91	6.88	0.9	2.07	1.97	0.9
4	11.92	11.87	0.8	4.18	4.22	1.1

Holistic Versus Analytic Overall Indices Comparison

Another important way to compare scoring methods is on the overall fit indices. Misfitting observations of two absolute standardized residuals or larger for a model to be acceptable should be in the vicinity of 5%. Likewise, at most approximately 1% of the data should have Misfitting observations with three or larger absolute standardized residuals (Linacre, 2011). Two Chi-square Goodness of Fit tests were performed to compare the two scoring approaches on their overall fit to the MFRM model. One test uses the two or more absolute standardized residuals standard and the other adopts the three or larger absolute standardized residuals rule (Alharby, 2006).

First Chi-Square Test

In Table 4.21 the Fit-Misfit dimension represents the columns, and the rows are the observed vs. expected. Whereas the cells for observed fit and misfit were obtained from absolute standardized residuals of the analytic approach, the cells for expected fit and misfit were derived from the analytic approach yet by using the percentage of fitting and misfitting observations of the holistic. That was made to make the holistic approach as the base line for comparison (Alharby, 2006).

As seen from Table 4.21 the observed frequency of the observed fitting observations were very similar to that of the expected (4024 vs. 4026), and correspondingly, the observed frequency of the observed misfitting observations were similar to its counterpart expected ones (161 vs. 159). The resulting Chi-square value was χ^2 (1) = 0.026, and was not significant ($p > .05$). Knowing that holistic was used as the theoretical base level for analytic expected frequencies; such a result might suggest that

both scoring approaches were not statistically different when compared on their overall fit indices.

Table (4.21) Chi-Square Test 1 (2 Abs. St. Deviation)

	Fit	Misfit	Total
Observed	**4024**	**161**	**4185**
Expected	**4026**	**159**	**4185**
			$\chi 2$ **= 0.026**

Second Chi-Square Test

In the second chi-square test the three or larger absolute standardized residuals criterion was adopted as the criterion. Again, in Table 4.22 the fit-misfit dimension represents the columns, and the rows are the observed vs. expected. Whereas the cells for observed fit and misfit were obtained from absolute standardized residuals of the analytic approach, the cells for expected fit and misfit were derived from the analytic approach yet by using the percentage of fitting and misfitting observations of the holistic. As stated before, that was made to make the holistic approach as the base line for comparison.

As seen from Table 4.22 the observed frequency of the observed fitting observations were different from that of the expected (4159 vs. 4022), and correspondingly, the observed frequency of the observed misfitting observations were different from their counterpart expected ones (26 vs. 63). The resulting Chi-Square value was $\chi 2$ (1) = 22.062, and was significant ($p < .0001$). Knowing that holistic was used a theoretical base level to for analytic expected frequencies; such a result might suggest

that both scoring approaches were statistically different when compared on their overall fit indices, and that analytic approach showed a better overall fit to the model. This result was not surprising, since analytic approach individual indices were slightly better than their counterparts of the holistic.

Table (4.22) Chi-Square Test 2 (3 Abs. St. Deviation)

	Fit	**Misfit**	**Total**
Observed	**4159**	**26**	**4185**
Expected	**4122**	**63**	**4185**
			χ^2 **= 22.062**

Generalizability Theory Analyses

Holistic Scoring Approach

Data obtained through holistic scoring were analyzed by Genova application. Ordinary generalizability theory analyses are performed in two stages; generalizability and decision studies. The purpose of generalizability study, usually denoted G-study, is to obtain variance components for each source of variance. A decision study, on the other hand is to explore different estimates of reliability for different but pertinent scenarios given the amount and type of variables (or facets) introduced in the model.

The design adopted in this study is fully-crossed two-faceted (p x i x r) with students as the object of measurement and items and raters as random sources of errors. With such a design, seven sources of variance can be statistically identified; variance due to students (true variance), variance due to raters (error variance not pertinent in relative model), variance due to passages (error variance not pertinent in relative model), variance due to interaction between students and raters (error variance), variance due to interaction between students and passages (error variance) and interaction between passages and raters (not pertinent in relative model), variance due to interaction between students, passages and raters and residuals (error variance).

As seen in Table 4.23 the largest variance component was that of the students (1.10) accounting for approximately 82% of the observed variance. This is considered a true variance and desirable as the object of many testing situations is to differentiate reliably among participants' scores. This variance component is very large in magnitude, and suggests an evident variation between students on their Quranic recitations. On the other hand, raters and passages combined explained only no more 3.5% of the variance,

suggesting that rates grading of students were somewhat similar. In other words, consistent with cited works on rater effects (e.g., Lane et. al., 1996; Shavelson et. al, 1993), raters showed reasonably to some extent similar stringency. Likewise, tasks' main effect was slightly small and might suggest that students find passages slightly similar in difficulty. Clearly, any departure from zero suggests differences. Yet, some departure can be proportionally ignored. Of the interaction variance components, variance interaction between students and passages accounted for 3.8% of the observed variation. This estimate suggests that rank ordering of students depends on tasks. Clearly, the task by rater interaction variance component was very negligible .Yet, 4.5% of the variance was accounted for by the interaction between students and raters. In other words, different ranking ordering of students by raters accounts for 4.5% of the variance in the data. Likewise, 6% of the variance is accounted for by the triple interaction between students, raters and passages as well as residuals. That is rank ordering of students depends on tasks administered, raters employed, and other unexplained variables (i.e., not specified in the model).

Table (4.23)Variance Components of Holistic Scoring for Generalizability Study (pxixr)

Effect	Degrees of Freedom	Variance Component	Standard Error	% Variance Accounted for
Students (p)	92	1.10269	0.1677235	.8193383
Raters (r)	2	0.01746	0.0132589	.01297341
Passages (i)	2	0.02876	0.0211781	.02136971
Interaction (pr)	184	0.06080	0.0094051	.04517658
Interaction (pi)	184	0.05189	0.0085057	.03855613
Interaction (ir)	4	0.00099	0.0010923	.00073561
Interaction (pir)	368	0.08324	0.0061196	.06185031

Table (4.24) D-Study Scenarios, Pertinent True Variances, Relative $\sigma^2(\delta)$ ***and Absolute*** $\sigma^2(\Delta)$ ***Error Variances, and Estimates of Generalizability and Dependability Values (Holistic).***

	D-Studies						
n_r	1	1	1	1	2	3	5
n_p	1	2	3	5	1	1	1
$\sigma^2(\tau)$	1.10269	1.10269	1.10269	1.10269	1.10269	1.10269	1.10269
$\sigma^2(\delta)$	0.19592	0.12836	0.10584	0.08782	0.12390	0.09990	0.08069
$\sigma^2(\Delta)$	0.24313	0.16069	0.13321	0.11123	0.16189	0.13481	0.11314
ρ^2	0.84913	0.89573	0.91243	0.92623	0.89899	0.91693	0.93181
Φ	0.81934	0.87281	0.89221	0.90837	0.87198	0.89106	0.90694

D-Study Scenarios:

This section simulates different scenarios usually encountered in schools. In one configuration, the number of raters assigning students' scores given their recitation of one

passage was manipulated to see their effect on reliability estimates. In the other configuration, the number of passages was increased and the number of raters kept at one.

Using one passage and one rater yields a G-coefficient of 0.849 and a D-index of 0.819. When one rater rates students recitations of two passages the G-coefficient become and the D-index become 0.896 and 0.873, respectively. Using three passages result in a G-coefficient of 0.912 and a D-index of 0.892, and using five passages produces a G-coefficient of 0.926 and a D-index of 0.908.

On the other hand, rating students by 2 raters on one passage produces a G-coefficient of 0.899 and a D-index of 0.872. When three raters assign scores for students' recitation of one passage, the G-coefficient of 0.917 and a D-index of 0.891 were yielded, and using five passages produces a G-coefficient of 0.932 and a D-index of 0.907.

Analytic Scoring Analysis:

The design employed in this study was three-faceted (pxixrxd) with students as the object of measurement and items, raters and sub-domains as random sources of errors. With such a design, 15 sources of variance can be statistically estimated: 1) variance due to students (true variance); 2) variance due to raters (error variance not pertinent in relative model); 3) variance due to passages (error variance not pertinent in relative model); 4) variance due to sub-domains (error variance not pertinent in relative model); 5) variance due to interaction between students and raters (error variance); 6) variance due to interaction between students and passages (error variance); 7) variance due to interaction among between and sub-domains (error variance): 8) interaction between passages and raters (not pertinent in relative model); 9) interaction between passages and sub-domains (not pertinent in relative model); 10) interaction between

raters and sub-domains (not pertinent in relative model); 11) interaction among students, passages and raters (error variance); 12) interaction among students, passages and sub-domains (error variance); 13) interaction among students, raters and sub-domains (error variance); 14) interaction among passages, raters and sub-domains (error variance not pertinent in relative model); and, 15) variance due to interaction among students, passages and raters and residuals (error variance).

Table (4.25)Variance Components of Analytic Scoring for Generalizability Study (pxixrxd)

Effect	Degrees of Freedom	Variance Component	Standard Error	% Variance Accounted for
Students (p)	92	1.49385	0.22388	0.777642
Raters (r)	2	0.00202	0.00312	0.001052
Passages (i)	2	0.01152	0.00957	0.005997
Sub-domains (d)	4	0.02539	0.00358	0.013217
Interaction (pr)	184	0.03815	0.00719	0.019859
Interaction (pi)	184	0.01330	0.00459	0.006923
Interaction (pd)	368	0.05696	0.00158	0.029651
Interaction (ir)	4	0.00028	0.00099	0.000146
Interaction (rd)	8	0.00601	0.00074	0.003129
Interaction (id)	8	0.00601	0.00074	0.003129
Interaction (pir)	368	0.02926	0.00446	0.015232
Interaction (prd)	736	0.04677	0.00106	0.024347
Interaction (pid)	736	0.03990	0.00100	0.02077
Interaction (rid)	16	0.00349	0.00034	0.001817
Interaction (prid)	1472	0.14809	0.00109	0.07709

As shown in Table 4.25, the largest contributor to variation is still the students' scores. They are the object of measurement in the current study and the variance associated with them is a true variance and desirable. The variance for students' scores accounted for slightly less than 78% of the variation, though it was less than that found with holistic. That was anticipated since inclusion of more random variables will reduce the true variance unless the interaction component of that facet with the students' facet is zero. The second largest contributor to variation was the interaction between all main effects which cannot be separated from residuals which accounted for 0.77% of the total variation. Main effects of the raters and items in this design were found very small. In other words passages seem very similar in difficulty and the same can be said of raters for their stringency (0.6%and 0.1%, respectively). The inclusion of sub-domains in the model explains 0.13 of the variation in the observations indicating that sub-domains differ in difficulty. This inclusion seemed to help objectively evaluate students on their Quranic recitations.

Of all interaction variance components, the variance interaction of students and sub-domains was the second largest in magnitude after the quadruple-interaction of all main effects as well as unexplained variation (i.e. residuals). This accounts for approximately 0.3% of the variation. Item by passage interaction variance was very negligible (0.015%) and persons by raters interaction explained approximately 0.2% of the variation. Similarly, triple interactions of students by raters by sub-domain and person by passage by domain each explained approximately 0.2% of the variation. This means that students ordering from low to high depends on passages and sub-domains, and in a similar manner depends on raters and sub-domains.

D-Study Scenarios

The current explored two different designs. One configuration treated sub-domains as a fixed facet and the other as a random facet. This was adopted to allow for the possibility of adding and/or omitting specific sub-domains.

Sub-Domain as a Fixed Facet:

This design treats sub-domains as a fixed facet. Thus, it makes the interaction variance of sub-domains with students a part of the true variance (i.e., students' variance). Such a design is plausible if these sub-domains used in the current study are the only sub-domains on which recitations of high school and junior high schools students are assessed. Perhaps, another reason could be that these are the only dimensions as they span the whole set of criteria for recitation.

Different scenarios were explored to better aid the practice of Quranic assessment. Raters and passages were manipulated to investigate their impact on the assessment situations. Given that performance assessment is a heavy labor and time consuming project, schools in pursuit of fair and reliable scoring as well as practicality would either have students recite more than one passage of the Quran or have more than one rater appraise the students' performance. With these goals in mind, different D-studies were assessed for their derived score dependability for both norm-referenced and criterion referenced standards. On one hand, norm-referenced models (relative model) concern only rank ordering of students regardless of how similar or dissimilar raters, passages or

sub-domains are. Absolute models, on the other hand, are sensitive to main effects as well (i.e., how similar the raters or passages are in the universe of generalization).

Raters and passages were increased in number from 1 to 5 but with holding one variable constant. Using only one rater and one passage resulted in a G-coefficient of 0.92 and a D-index of 0.91. These estimates are quite high for performance assessment. Increasing passages to two while holing raters constant at one showed an increase in both the G-coefficient (0.945) and the D-index (0.939). Adding more passages yielded a G-coefficient of 0.953 and a D-index of 0.948. A Quranic recitation assessment of five passages by one rater results in a G-coefficient of 0.959 and a D-index of 0.956.

On the other hand, keeping one passage and increasing number of raters (as this is the other configuration schools usually adopt in rating their students) results in the following indices. Two raters and one passage produce a G-coefficient of 0.953 and a D-index of 0.944. An increase of one rater results in a G-coefficient of 0.964 and D-index of 0.955. Finally, employing fiver raters to rate students on a single passage produces a G-coefficient of 0.973 and a D-index of 0.964.

Clearly, the more the better; however, there are time and money limitations to be considered. Employing more raters would result in higher reliability estimates than employing more passages. However, these differences appear to be small when considering feasibility of raters and the large number of students. Either employing two raters and one passage or one passage and two raters would achieve scores of comparable dependability to those employing more rates or passages.

Table (4.26) D-Study Scenarios, Pertinent True Variances, Relative $\sigma^2(\delta)$ and Absolute $\sigma^2(\Delta)$ Error Variances, and Estimates of Generalizability and Dependability Values (Analytic for Fixed Sub-Domains).

	D-Studies						
N_r	1	1	1	1	2	3	5
N_p	1	2	3	5	1	1	1
N_d	5	5	5	5	5	5	5
$\sigma^2(\tau)$	1.50525	1.50525	1.50525	1.50525	1.50525	1.50525	1.50525
$\sigma^2(\delta)$	0.12767	0.08759	0.07423	0.06354	0.07448	0.05674	0.04256
$\sigma^2(\Delta)$	0.14459	0.09766	0.08202	0.06950	0.08930	0.07086	0.05612
ρ^2	0.92182	0.94501	0.95300	0.95950	0.95286	0.96367	0.97250
Φ	0.91236	0.93907	0.94833	0.95587	0.94400	0.95504	0.96406

Sub-Domain as a Random Facet:

This section deals with scenarios where sub-domains were treated as random. In other words, this design allow the possibility of considering these sub-domains used in this study to be a sample of comparable sub-domains. Using only on rater and one passage resulted in a G-coefficient of 0.915 and a D-index of 0.903. As stated before, these estimates are quite high for performance assessment. Raising the passage number to two and using one rater only showed an increase in both the G-coefficient (0.938) and the D-index (0.929). The addition of one more passage gives a G-coefficient of 0.946 and a D-index of 0.938. A Quranic recitation assessment of five passages by one rater results in a G-coefficient of 0.95 and a D-index of 0.946.

Alternatively, keeping one passage and increasing number of raters as this is the other configuration schools occasionally adopted in rating their students results in the following indices. A scenario of two raters and one passage produces a G-coefficient of 0.946 and a D-index of 0.934. The addition of one more rater results in a G-coefficient of 0.956 and D-index of 0.945. Lastly, employing fiver raters to rate students' recitation on a single passage produces a G-coefficient of 0.965 and a D-index of 0.954.

Clearly, the addition of more raters or slightly, equally more passages would increase dependability of the scores. Employing more raters would result in slightly higher reliability estimates than employing more passages. However, with limitations of time and labor considered, these differences appear to be small. Considering the feasibility of raters and the large number of students either employing two raters and one passage or one passage and two raters would achieve scores of comparable dependability to those employing more raters or passages.

Table (4.27) D-Study Scenarios, Pertinent True Variances, Relative $\sigma^2(\delta)$ and Absolute $\sigma^2(\Delta)$ Error Variances, and Estimates of Generalizability and Dependability Values (Analytic for Random Sub-Domains).

	D-Studies						
Nr Np Nd	1 1 5	1 2 5	1 3 5	1 5 5	2 1 5	3 1 5	5 1 5
$\sigma^2(\tau)$	1.49385	1.49385	1.49385	1.49385	1.49385	1.49385	1.49385
$\sigma^2(\delta)$	0.13906	0.09898	0.08562	0.07493	0.08587	0.06814	0.05395
$\sigma^2(\Delta)$	0.16106	0.11413	0.09848	0.08597	0.10576	0.08733	0.07259
ρ^2	0.91484	0.93786	0.94579	0.95224	0.94564	0.95638	0.96514
Φ	0.90268	0.92902	0.93815	0.94558	0.93388	0.94477	0.95366

Holistic Versus Analytic Results

This section compares holistic to analytic scoring approaches on individual estimates of variance components and estimates of generalizability and dependability. Analytic scoring approach of Quranic recitations showed appreciably lower variance components for all main effects and interaction components. With exception of student variance component, such a result might lend support to the use of analytic scoring approach in Quranic assessments. The addition of sub-domains seemed of value as it explains more of the variability in the scores. Appreciably, it reduces raters' effect and makes them to appear more comparable in stringency (0.1% vs. 1.3%). The same conclusion can be drawn for comparability of passages (0.6% vs. 2.1%).

Likewise, rating students' recitations on different dimensions helps reduce students by raters (2%vs. 4.5%), students by passages (0.7% vs. 3.9%), and raters by passages interactions (0.01% vs. 0.07%).

Contrasting analytic versus holistic on reliability estimates shows evident impact. Notably, it took the holistic five passages and one rater to achieve reliability estimates comparable in magnitude to those of analytic with fixed-sub-domains using only one rater and one passage (G-coefficient of 0.91 and D-index of 0.908 vs. 0.92 and 0.91). A similar conclusion can be drawn for Analytic scoring with random sub-domains (G-coefficient of 0.91 and D-index of 0.908 vs. 0.91 and 0.903). A holistic scoring approach with five raters and one passage yields reliability estimates less than an analytic approach of only two raters on passage (G-coefficient of 0.932 and D-index of 0.907 vs. 0.946 and 0.934). An analytic approach with fixed-sub-domains yielded even higher estimates (G-coefficient of 0.953 and D-index of 0.94).

To compare the observed estimates of generalizability and dependability, different intervals were built around the estimates. This approach was suggested by Hoi (cited in Alharby, 2006) where the lower confidence bands were built with minimum true variance and maximum error variance, and the upper confidence bands used maximum true variance and minimum error variance. These minimum and maximum variances use plus or minus one standard error to obtain the lower and upper limits.

Two approaches are used in the current study, one uses + or – one standard error and the other uses + or – two standard errors. For the CIs built with one standard error, four testing structures that can feasibly be adopted at schools were compared. As seen from Table 4.28 this approach shows non-overlapping confidence bands for estimates of

generalizability coefficients and dependability indices for all assessed situations. Written differently, all analytic estimates of reliability were higher than their holistic counterparts, and these indices might suggest a significant difference.

CIs with Two Standard Errors

In keeping with the conventional practice in statistics and measurement in using two standard errors for 95% CIs, four testing structures that can feasibly be adopted at schools were compared using CIs of two standard errors. As seen from Table 4.29 this approach shows overlapping confidence bands for estimates of generalizability coefficients and dependability indices for all assessed situations. Written differently, all analytic estimates of reliability were higher than their holistic counterparts, but however, the indices showed overlapping confidence intervals.

Table (4.28)One Std. Error Confidence Intervals for Estimates of Generalizability and Dependability Indices

	Analytic	Holistic	Difference
1r 1p			
G-coeff.	(0.90; 0.94)	(0.82; 0.87)	*
D-index	(0.90; 0.93)	(0.66; 0.86)	*
1r 2p			
G-coeff.	(0.93; 0.96)	(0. 87; 0.91)	*
D-index	(0.92; 0.95)	(0.84; 0. 90)	*
1r 3p			
G-coeff.	(0.94; 0.96)	(0.89;0.93)	*
D-index	(0.93; 0.96)	(0.86,0.92)	*
2r 1p			
G-coeff.	(0.94; 0.96)	(0.88; 0.92)	*
D-index	(0.93; 0.96)	(0.83; 0. 90)	*

*Indicates non-overlapping CIs.

Table (4.29)Two Std. Error Confidence Intervals for Estimates of Generalizability and Dependability Indices

	Analytic	Holistic
	1r 1p	
G-coeff.	(0.88; 0.95)	(0.78; 0.89)
D-index	(0.86; 0.94)	(0.72; 0.88)
	1r 2p	
G-coeff.	(0.91; 0.96)	(0. 84; 0.93)
D-index	(0.90; 0.96)	(0.80; 0. 92)
	1r 3p	
G-coeff.	(0.92; 0.97)	(0.86;0.94)
D-index	(0.92; 0.97)	(0.82,0.94)
	2r 1p	
G-coeff.	(0.93; 0.97)	(0.84; 0. 93)
D-index	(0.91; 0.97)	(0.79; 0. 93)

Table (4.30) Confidence Intervals for estimates of SEM (δ) and SEM (Δ) Indices

	Analytic	Holistic	Difference
	1r 1p		
δ	(0.334;0.379)	(0.414;0.469)	*
Δ	(0.345;0.412)	(0.434;0.848)	*
	1r 2p		
δ	(0.271;0.319)	(0.329;0.385)	*
Δ	(0.282;0.3399)	(0.349;0.446)	*
	1r 3p		
δ	(0.246;0.297)	(0.295;0.353)	
Δ	(0.256;0.3135)	(0.3137;0.410)	*
	2r 1p		
δ	(1.1E-06;4.5E-06)	(1.3E-05;5.9E-05)	*
Δ	(0.260;0.333)	(0.339;0.457)	*

*Indicates non-overlapping CI.

Two Liberal Approaches

The current study suggests alternatively two liberal but reasonable approaches to test for observed differences in generalizability and dependability estimates. Since reliability can be seen as absence of measurement error, this study suggests the comparison of the Quranic recitation scoring rubrics be made on the basis of their relative and absolute standard error of measurement (SEM) or alternatively on their measurement error variance (MEV). This might appear sensible; SEM is just another but equivalent

way of assessing dependability of scores. Second, it avoids undue maximization and minimization of true and error variances that make confidence intervals appear unduly large, and significantly less sensitive to the differences in generalizability and dependability estimates. Third, it can provide more reliability in the resulting estimates by using a 95% CI. In other words, this approach might provide a more sensitive but reliable approach to detect scoring rubric differences in the estimates of scores' dependability.

Alternatively, the two scoring rubrics can be compared twice; one assuming the best case scenario (i.e., both scoring rubrics showing their highest indices of reliability) and the other assuming the worst case scenario (i.e., both scoring rubrics showing their smallest indices of reliability). Again, these confidence intervals can be built with two standard errors in order to objectively detect such differences.

Comparing CIs of analytic best estimates of reliability indices to holistic best estimates of reliability indices, shows non-overlapping CIs for all four designs. Likewise, comparing CIs of analytic lowest estimates to holistic lowest estimates yields non-overlapping CIs for all four testing structures. The results of these comparisons lend support to the analytic scoring approach for Quranic recitations.

Similarly, comparing analytic scoring to holistic on the basis of their measurement error variance, yield very similar results except for the relative SEM of a design with one rater and three passages. Overall, such a result might give support to the analytic scoring approach as the most proper way of Quranic recitation assessment.

Table (4.31) CI for Analytic vs. Holistic Best Case Scenario

	Analytic	Holistic	Difference
	1r 1p		
G-coeff.	(0.95; 0.93)	(0.89; 0.87)	*
D-index	(0.94; 0.92)	(0.88; 0.83)	*
	1r 2p		
G-coeff.	(0.96; 0.95)	(0.93; 0.91)	*
D-index	(0.96; 0.94)	(0.92; 0.88)	*
	1r 3p		
G-coeff.	(0.97; 0.96)	(0.94; 0.92)	*
D-index	(0.97; 0.95)	(0.94; 0.89)	*
	2r 1p		
G-coeff.	(0.97; 0.96)	(0.93; 0.91)	*
D-index	(0.97; 0.95)	(0.93; 0.87)	*

*Indicates non-overlapping CI.

Table (4.32) CI for Analytic vs. Holistic Worst Case Scenario

	Analytic	Holistic	Difference
	1r 1p		
G-coeff.	(0.91; 0.88)	(0.82; 0.78)	*
D-index	(0.90; 0.86)	(0.80; 0.72)	*
	1r 2p		
G-coeff.	(0.94; 0.91)	(0.88; 0.84)	*
D-index	(0.93; 0.90)	(0.86; 0.80)	*
	1r 3p		
G-coeff.	(0.95; 0.92)	(0.90; 0.86)	*
D-index	(0.94; 0.92)	(0.89; 0.82)	*
	2r 1p		
G-coeff.	(0.94; 0.93)	(0.88; 0.84)	*
D-index	(0.94; 0.91)	(0.87; 0.79)	*

*Indicates non-overlapping CI.

Chapter Five

Discussion

The principle aim of the present study is to better guide the Quranic recitation assessment. It sheds light on the Quranic assessment in Saudi Arabia by exploring the general practice, suggesting tow different scoring rubrics and evaluating these scoring rubrics for their effectiveness. It, for the most part, examines the effect of scoring schemes on ratings of Quran recitations using two measurement models (i.e., G-theory and MFRM).

First, this study evaluates the relationship of raw scores of holistic and analytic ratings. Such estimates were very high ($\rho = 0.942$ and $p < 0.01$), suggesting that, on average, rank ordering of students is consistent across the scoring rubrics. Similarly, the Pearson's correlation estimate between holistic calibrated measures and analytic calibrated measures is high as well ($\rho = 0.933$ and $p < 0.01$). As anticipated, holistic raw scores correlate highly with their calibrated estimates ($\rho = 0.988$ and $p < 0.01$), and in the same manner do the analytic raw scores with their calibrated estimates ($\rho = 0.956$ and $p < 0.01$). Notably, such high correlational estimates between analytic scores and holistic scores might suggest little differences in preferring one scoring approach to another for relative scoring models.

Closer examination of analytic and holistic raw score data reveals higher arithmetic means of analytic scores than those of holistic. This study, in contrast to Chi's study (2001), finds that analytic raw scores have higher mean and show more variability. Finding higher means for analytic scoring rubrics is consistent with previously cited work (e.g., Alharby, 2006; Goulden, 1994), and a paired-sample t-test shows a statistically

significant difference between the means of analytic and holistic ($p < 0.001$). However, the calibrated mean difference between holistic and analytic is not statistically significant ($p > 0.05$).

Individual comparisons of holistic and analytic scoring rubrics of Quranic recitations using MFRM showed that analytic scoring rubric is associated with better overall Infit and Outfit statistics for all measurement facets than those of holistic. They generally produced quality control statistics that are closer to their expected values and showed less spread from those values. Moreover, closer evaluation of elements within facets exhibited similar results. Elements of analytic showed more internal consistency and were well modeled when compared to those of holistic. Likewise, analytic scoring rubrics of recitations were found to be associated with lesser rater separation than their holistic counterparts. And that might lend more support to the introduction of analytic scoring to the Quranic assessment practice, as it guides raters to rate consistently and similarly. In addition, analytic scoring rubrics of recitations revealed better gradation of passages along the logit scale of recitations. This is, specifically, of value for objectively and accurately placing students linearly on the recitation scale.

This study also compares both scales of recitation on their overall fit indices. Two tests of Goodness of Fit were conducted. The first test compared analytic to holistic regarding absolute standardized residuals with a magnitude of two or larger. No statistical difference was found, χ^2 (1) = 0.026, and p > .05. However, comparing both rubrics on their absolute standardized residual of three or larger (i.e., the second Goodness of Fit test) was statistically significant, χ^2 (1) = 22.062, and p < .0001. Again, this gives a

piece of support to the introduction of rubrics into the practice of Quran in general and to analytic rubrics in specific.

G-theory analysis showed similar results to those of MFRM. Analytic scoring rubrics were associated with lesser measurement errors and with higher coefficients of dependability (i.e., G-coefficients, and D-indices). The introduction of analytic rubrics might have helped guide raters to evaluate students more consistently and bring raters to a common understanding of the scoring scales.

Analytic scoring approach of Quranic recitations showed substantially lower variance components for all main effects and interaction components. The addition of sub-domains explained more of the variability in the scores. Considerably, it reduces the variance components associated with raters and makes them to appear more comparable in stringency (0.1% vs. 1.3%). Same conclusion can be drawn for comparability of passages (0.6% vs. 2.1%). Appreciably, evaluating students' recitations on different dimensions helps lower the inconsistencies of rank-ordering students by raters (2%vs. 4.5%), students by passages (0.7% vs. 3.9%), and raters by passages (0.01% vs. 0.07%).

Comparing analytic to holistic on reliability estimates shows evident impact. Remarkably, it required of the rater using holistic rubric to evaluate students on five passages to achieve reliability estimates comparable in magnitude to those of analytic with fixed-sub-domains using only one rater and one passage (G-coefficient of 0.91 and D-index of 0.908 vs. 0.92 and 0.91). Similar conclusion can be drawn for Analytic with random sub-domains (G-coefficient of 0.91 and D-index of 0.908 vs. 0.91 and 0.903).

To compare the observed estimates of generalizability and dependability, different intervals were built around the estimates. Three different approaches were used in the

current study, one uses + or – one standard error (Hoi cited in Alharby, 2006), and the other uses + or – tow standard errors. For the CIs built with one standard error, four testing structures that can feasibly be adopted at schools were compared and it showed non-overlapping confidence bands for estimates of Generalizability coefficients and Dependability indices for all assessed situations. Put differently, although all analytic estimates of reliability were higher than their holistic counterparts, none was found statistically different.

Consistent with the conventional practice in statistics and measurement in using tow standard error for 95% CIs, four testing structures that can feasibly be adopted at schools were contrasted using CIs of two standard errors. This approach shows overlapping confidence bands for estimates of generalizability coefficients and dependability indices for all assessed situations. Written differently, although all analytic estimates of reliability were higher than their holistic counterparts, they however, showed overlapping confidence intervals.

Since SEM is but another way of assessing dependability of scores, 95% CIs were built around absolute and relative the measurement error variances. These results reveal non-overlapping CIs, suggesting that analytic scoring rubrics within the Quranic recitation assessments are associated with lesser absolute and relative measurement errors. Comparing dependability estimates (i.e., G-coefficients and D-indices) on their best estimates lend support to the analytic as the most proper way to produce more consistent and similar ratings. A comparison on their worst estimates reveals a similar picture.

As Nakamura (2004) points out, students' demand for more meaningful feedback from teachers might lend more support to the use of analytic scoring rubrics (2004). Such a demand for more informative feedback seemed to be coupled with higher statistical evidences in the current study. Analytic scoring rubrics can be of great value for diagnostic purposes as they help find points of strength and weaknesses. Hence, directing students' attention to their strengths and weaknesses can help them improve. Also, such a rubric can help teachers tailor the instruction to accurately meet their students' needs. For instance, they can be of importance at the beginning of the academic year, where pretests are given to help assess students' performance and therefore guide the instruction.

Future research:

One possible study could explore more fine-grading rubrics (i.e., larger than five categories used in the current study). Possibly, a study manipulating passages difficulties and wealth of Tajweed rulings is suggested. Dimensions adopted for this study could be modified in order to improve the analytic rubric and a study to shoulder such a task is of ample importance. One way to improve the dimensions is by considering different weighing systems. Since this study assesses the consistency of ratings of different scoring schemes, a study assessing accuracy is needed. In other words, securing high reliability is not assurance of validity (e.g., Thompson, 2003). This would give more insight and better guide the assessment practice. A study evaluating the introduction of rubrics to the general Quranic assessment practice would be of ample importance and further research is needed to replicate the results of the current study.

Incorporating hierarchical modeling might be of great benefits as it incorporates educational settings effects, incorporates explanatory variables at different levels and handles non-response items (Raudenbush & Bryk, 2002). Educational and social Studies usually involve participants nested within classrooms or schools. Such nested designs suggest the use of multilevel IRT models (Raudenbush, Johnson & Sampson, 2003). Standard Rasch modeling treats ability and difficulty as fixed entities; however, with multilevel modeling specification of ability as random can be achieved and accurate conclusions can be drawn (Raudenbush & Bryk, 2002).

Limitations*:*

This study is not without certain limitations; first, the raters participating in this project were highly educated. This sophistication of raters cannot warrant any generalization of the findings to teachers of high or junior high schools. Second, rubrics used might be very coarse or crude and there is a possibility of modifying the rubrics to include more fine categories. Perhaps, another reason could be the use of rigid categories as it helps increase consistency but hurts validity. The break interval between the scoring sessions was short due to feasibility of raters, had a longer interval been given, a different picture might result. Another limitation is in the comparison design as the administration of the scoring rubrics was not counterbalanced to secure against carry-over effect. Although, the researcher reasoned that using holistic rubric first would not influence ratings from the analytic rubric, such reasoning might not prove true.

Another limitation lies in the sample of students included in the current study. Including junior high school students might exaggerate the differences between students and better more limited selection of students might give different results.

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.* Ph.D. thesis. Pennsylvania State University.

Alhussary, M. (1999). *The rulings of reciting the holy Quran* (In Arabic). Makah, K.S.A. Islamic Office.

Bartlett, C. J. (1983). What's the difference between valid and invalid halo? Forced-choice measurement without forcing a choice. *Journal of Applied Psychology, 68,* 218-226.

Becker, B. E, & Cardy, R. L. (1986). Influence of halo error on appraisal effectiveness: A conceptual and empirical reconsideration. *Journal of Applied Psychology, 71,* 662-671.

Bernardin, H. J., & Pence, E. C. (1980). Effects of rater training: Creating new response sets and decreasing accuracy. *Journal of Applied Psychology, 65,* 60-66.

Bingham, W. V. (1939). Halo, invalid and valid. *Journal of Applied Psychology, 23,* 221-228.

Blazer, W. K., & Sulsky, L. M. (1992). Halo and performance appraisal research: A critical examination. *Journal of Applied Psychology, 77,* 975-985.

Bonk, W. J. & Ockey, G. J. (2003). A many-facet Rasch analysis of the second language group oral discussion task. *Language Testing,* 20, 89–110.

Borman, W. C. (1979). Format and training effects on rating accuracy and rater errors. *Journal of Applied Psychology, 64,* 410-421.

Brennan, R. L. (1995). Generalizability of performance assessments. *Educational Measurement: Issues and Practic, 14*, 9-12.

Burton, N. W. (1981). Estimating scorer agreement for nominal categorization systems. *Educational and Psychological Measurement, 41*, 953-962.

Cardinet, J., Tourneur, Y., & Allal, L. (1976). The symmetry of generalizability theory: Applications to educational measurement. *Journal of Educational Measurement, 13*, 119-135.

Carr, N. T. (2000). A comparison of the effects of analytic and holistic Rating scale types in the context of composition tests. *Issues in Applied Linguistics, 11* (2), 207-241

Chase, C. I. (1999). *Contemporary assessment for educators*. New York: Longman.

Chi, E. (2001). Comparing holistic and analytic scoring for performance assessment with Many-Facet Rasch Model. *Journal of Applied Measurement, 2*, 379-388.

Cooper, W. H. (1981a). Ubiquitous halo. *Psychological Bulletin, 90*, 218-244.

Cooper, W. H. (1981b). Conceptual similarity as a source of illusory halo in job performance ratings. *Journal of Applied Psychology, 66*, 302-307.

Cornelius, E. T., & Lyness, K. S. (1980). A comparison of holistic and decomposed judgment strategies in job analyses by job incumbent. *Journal of Applied Psychology, 65*, 155-163.

Crick, J. E. & Brennan, R. L. (1993). *GENOVA: A generalized analysis of variance system, FORTRAN IV computer program and manual* Version 2.1. Iowa City, IA*:* American College Testing Program.

Dennis, I. (2007). Halo effects in grading student projects. *Journal of Applied Psychology, 92,* 1169-1176.

Eckes, T. (2009). Many-facet Rasch measurement. In S. Takala (Ed.), *Reference supplement to the manual for relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment* (Section H). Strasbourg, France: Council of Europe/Language Policy Division.

Embretson, S. E., & Reise, S. P. (2000). Item response theory for psychologists. Mahwah, NJ: Lawrence Erlbaum.

Facteau, J. D., & Craig, S. B. (2001). Are performance appraisal ratings from different rating sources comparable? *Journal of Applied Psychology, 86,* 215-227.

Feldman, J. M. (1986). A note on the statistical correction of halo error. *Journal of Applied Psychology, 71,* 173-176.

Fischer, G. (1995). Derivations of the Rasch model. In G. Fischer & I. Molenaar. (Eds.), *Rasch Model: Fundations Recent Developments, and Applications.* (pp. 15-38), NY: Springer.

Fisicaro, S. A. (1988). A reexamination of the relation between halo error and accuracy. *Journal of Applied Psychology, 73,* 239-244.

Fuqua, D. R., Newman, J. L., Scott, T. B., & Gade, E. M. (1986). Variability across sources of performance ratings: Further evidence. *Journal of Counseling Psychology, 33,* 353-356.

Green, S. B. (1981). A comparison of three indexes of agreement between observers: Proportion of agreement, G-index and kappa. *Educational and Psychological Measurement, 41,* 1069-1072.

Greguras, G. J., & Robie, C. (1998). A new look at within-source interrater reliability of 360-degree feedback ratings. *Journal of Applied Psychology, 83,* 960-968.

Haladyna, T. M. (1997). *Writing test items to evaluate higher order thinking*. Needham Heights, MA: Allyn and Bacon.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory*: *Principles and applications*. Norwell, MA: Kluwer

Hambleton, R. K., Swaminathan, H. & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, NY: Sage

Hamp-Lyons, L. (1995). Rating nonnative writing: The trouble with holistic scoring. *TEOSL Quarterly, 29,* 759-762.

Haris, D. J. (1997) Using reliability to make decisions. In Haris, D. J. (Eds.), *Reliability Issues with Performance Assessments: A Collection of Papers* (Vol. 3, pp. 1-12) ACT Research Reprot Series 97-3.

Harrison, J. A., McAfee, H. & Caldwell, A. (2002). *Examining, developing and validating the interview for the admission into the teacher education program*. Paper presented at the annual meeting of the Southeastern Region Association for Teacher Educators, Hot Springs, AR.

Heneman, R. L., Moore, M. l., & Wexley, K. N. (1987). Performance-rating accuracy: A critical review. *Journal of Business Research, 15,* 43-448.

Herman, J. L., Aschbacher, P. R., & Winters, L. (1992). *A practical guide to alternative assessment*. Alexandria, VA: Association of Supervision and Curriculum Development.

Kane, M. T. (2006). Validation. In Brennan, R. (Eds.). *Educational measurement.* (17-64). Westport, CT: Praeger.

Klein, S. P., Stecher, B. M., Shavelson, R. J., McCaffrey, D., Ormseth, T., Bell, R. M., Comfort, K., Othman, A. R. (1998). Analytic versus holistic scoring of science performance tasks. *Applied Measurement in Education, 11(2),* 121-137

Kraemer, H. C., Periyakoil, V. S., & Noda, A. (2002). Kappa coefficients in medical research. *Statistics in Medicine, 21,* 2109-2129.

Kvalseth, T. O. (1991). A coefficient of agreement for nominal scales: An asymmetric version of Kappa. *Educational and Psychological Measurement,* 91-95.

Lance, C. E. & Woehr, D. J. (1986). Statistical control of halo: Clarification from two cognitive models of the performance appraisal process. *Journal of Applied Psychology, 71,* 679-685.

Landy, F. J., Vance, R. J., Barnes-Farrell, J. L., & Steele, J. W. (1980). Statistical control of halo error in performance ratings. *Journal of Applied Psychology, 65,* 501-507.

Lane, S. & Stone, C. A. (2006). Performance assessment. In Brennan, R. (Eds.). *Educational measurement.* (17-64). Westport, CT: Praeger.

Linacre J. M. (2011). *What A User's Guide to FACETS Rasch Model Computer Program, version 3.68.1*. Beaverton, Oregon: Winsteps.com

Linacre J. M. (2002). What do Infit and Outfit, Mean-square and Standardized mean? *Rasch Measurement Transactions, 16,* 878.

Linacre, J. M. (1994). *Multi-facet Rasch measurement*. Chicago: MESA.

Linn. R. L., & Gronlund, N. E. (2000). *Measurement and assessment in Teaching.* Upper Saddle River, NJ: Prentice-Hall.

Lunz, M. E., Stahl, J. A., & Wright, B. D. (1994). Interjudge reliability and decision reproducibility. *Educational and Psychological Measurement, 54*, 913-925.

Mero, N. P., & Motowidlo, S. J. (1995). The effects of rater accountability on the accuracy and the favorability of performance ratings. *Journal of Applied Psychology, 80,* 517-524.

Ministry of Education. (n.d.). *Summary statistics on general education in K.S.A academic year 2009/2010*. Retrieved from http://www.moe.gov.sa/pages/stat30-31.aspx

Moskal, B.M., & Leydens, J. A. (2000). Scoring rubric development: Validity and reliability. *Practical Assessment, Research* & *Evaluation 7*, 10, 71-81. http://pareonline.net/getvn.asp?v=7&n=10

Murphy, K. R. (1982). Difficulties in the statistical control of halo. *Journal of Applied Psychology, 67*, 161-164.

Murphy, K. R., & Blazer, W. K. (1989). Rater errors and ratings accuracy. *Journal of Applied Psychology, 69,* 147-156.

Murphy, K. R., & Cleveland, J. N. (1995). *Understanding performance appraisal: Social, organizational, and goal based perspectives.* Thousand Oaks, CA: Sage.

Murphy, K. R., & DeShon, R. P. (2009). Interrater correlations do not estimate the reliability of job performance ratings. *Journal of Personnel Psychology, 53,* 873-900.

Musri, M. (2005). *The notebook of Tajweed* (In Arabic). Jeddah, K.S.A: House of Qiblah for Islamic Knowledge. .

Nakamura, Y. (2004). *A comparison of holistic and analytic scoring methods in the assessment of writing.* Paper presented at the Proceedings of the 3rd Annual JALT Pan-SIG Conference, Tokyo, Japan.

Nitko, A. (2004). *Educational assessment of students*. Upper Saddle River, NJ: Prentice Hall/Merrill Education.

Nunnally, Jr. J. C. (1959). Tests *and measurements: Assessment and prediction*. NY: McGraw-Hill.

Olson, D. (1988, February 25 - 26). *The reliability of analytic and holistic methods in rating students' computer programs*. Paper presented at the Proceedings of the nineteenth SIGCSE technical symposium on Computer science education, Atlanta, Georgia.

O'Neill, T.R. & Lunz, M.E. (2000) A method to study rater severity across several administrations. In M. Wilson & G. Engelhard, Jr. (Eds.), *Objective Measurement: Theory into Practice* (Vol. 5, pp. 135-146) Stamford, CT: Ablex.

Osterlind, S. J. (2006). Modern measurement: Theory, principles, and applications of mental appraisal. Upper Saddle River, NJ: Pearson.

Ostini, R., & Nering, M. L. (2006). *Polytomous item response theory models*. Thousand Oaks, CA: Sage.

Parkes, J. (2000). The Relationship between the Reliability and Cost of Performance and Cost Performance Assessments. *Education policy analysis archives,* 8(16*),* 1068-234.

Popham, W. J. (1993). *Educational evaluation*. Needham Heights, MA: Allyn and Bacon.

Rangel, M. (1990). Assessment methods and equal opportunities. In Riding, R. & Butterfield, S. (Eds.). *Assessments and examination in the secondary school* (200-205). NY: Chapman and Hall.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd Ed.). Thousand Oaks, CA: Sage

Raudenbush, S. W., Johnson, C., & Sampson, R. J. (2003). Multivariate multilevel Rasch model for self-reported criminal behavior. *Social methods, 33*, 169-211.

Reckase, M. D. (1995). The reliability of ratings versus the reliability of scores. *Educational Measurement: Issues and Practic, 14*, 31.

Sala, F., & Dwight, S. A. (2002). Predicting executive performance with multirater surveys: Whom you ask makes a difference. *Consulting Psychology Journal: Practice and Research, 54*, 166-172.

Sanchez, J. I., & Levine, E. L. (1994). The impact of raters' cognition on judgment accuracy: An extension to the job analysis domain. *Journal of Business and Psychology, 9*, 47-57.

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

Shrout, P. E., & Fliess, J. L. (1979). Interclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86,* 420-428.

Shultz, K. S. & Whitney, D. J. (2005). *Measurement theory in action*: *Case studies and exercises.* Thousand Oaks, California: Sage.

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

Spool, M. D. (1978). Training programs for observers of behavior: A review. *Personnel Psychology, 31,* 853-888.

Stone, M. H. (2002). Quality control in testing. *Popular Measurement, 4,* 15-23.

Suen, H. K. (1990). *Principles of test theories*. Hillsdale, NJ: Lawrence Erlbaum.

Suen, H. K., & Ary, D. (1989). *Analyzing quantitative behavioral observation data*. Hillsdale, NJ: Erlbaum.

Suen, H. K., Ary, D., & Covalt, W. C. (1990). A decision tree approach to selecting an appropriate observation reliability index. *Journal of Psychopathology and Behavioral Assessment, 12* (4), 359-363.

Suen H. K., Lee, P. S. C. & Prochnow-LaGrow, J. E, (1985). A critical review of the S/L reliability index. *Journal of Psychopathology and Behavioral Assessment, 7,* 277-287.

Suen, H. K., Logan, C. R., Neisworth, J. T., & Bagnato, S. (1995). Parent-professional congruence: Is it necessary? *Journal of Early Intervention, 19* (3), 243-252.

Sulsky, L. M., & Blazer, W. K. (1988). Meaning and measurement of performance rating accuracy: Some methodological and theoretical concerns. *Journal of Applied Psychology, 73*, 497-506.

Thompson, B. (2003). Understanding reliability and coefficient alpha, really. In Thompson, B. (Eds.). *Score reliability: Contemporary thinking on reliability issues.* (3-23). Thousand Oaks, California: Sage.

Thorndike, E. L. (1920). A constant error in psychological ratings. *Journal of Applied Psychology, 4,* 25-29.

Westgaard, O. (1999). *Tests that work*: *Designing and delivering fair and practical measurement tools in the workplace*. San Francisco, CA: Jossey-Bass/Pfeiffer.

Voskuijl, O. F., & Sliedregt, T. (2002). Determinants of Interrater reliability of job analysis: A meta-analysis. *Psychological Assessment, 18,* 52-62.

Wherry, R. J. (1931). A new formula for predicting the shrinkage of the multiple correlation coefficient. *Annals of Mathematical Statistics, 2,* 440-457.

White, E. M. (1984). Holisticism. *College Composition and Communication,* 35,400-409.

Wright, B. D. (1995). 3PL or Rasch? *Rasch Measurement Transactions, 9,* 408-409.

Wright, B. D. (1996). Reliability and Separation. *Rasch Measurement Transactions, 9,* 472.

Wright, B. D. & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions, 8,* 370

Wu, S. M., Whiteside, U. & Neighbor, C. (2007). Differences in inter-rater reliability and accuracy for a treatment adherence scale. *Cognitive Behaviour Therapy, 36*, 4, 230-239.

APENDIX A

DISCRIPTIVE STATISTICS

Descriptive Statistics

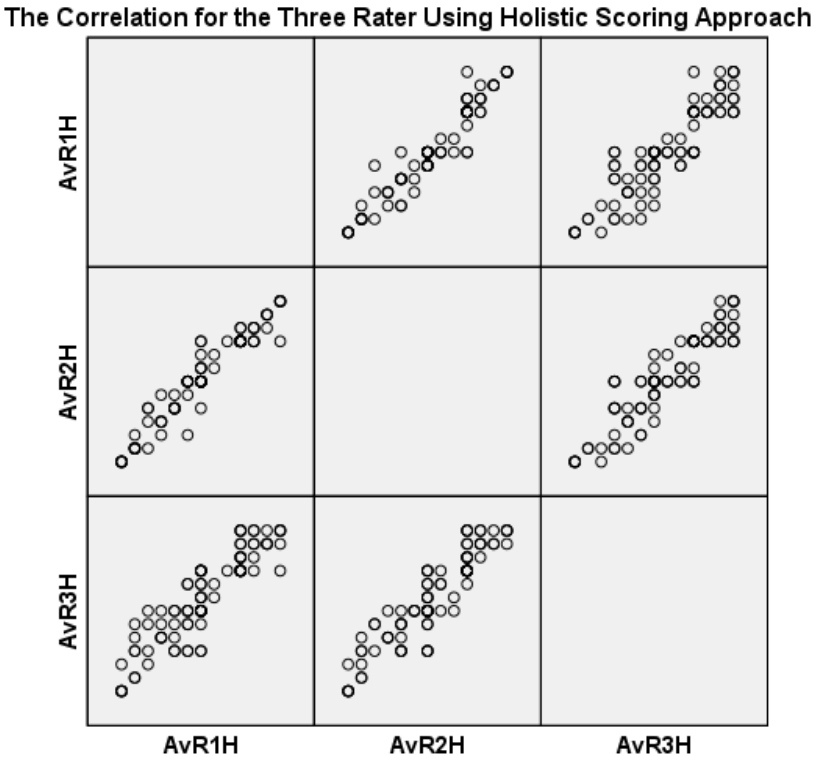
	N	Range	Minimum	Maximum	Mean	Std. Deviation
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic
AvAnR1	93	4.00	.00	4.00	2.6093	1.21483
AvAnR2	93	4.00	.00	4.00	2.4817	1.24486
AvAnR3	93	4.00	.00	4.00	2.5211	1.30879
Valid N (listwise)	93					

Descriptive Statistics

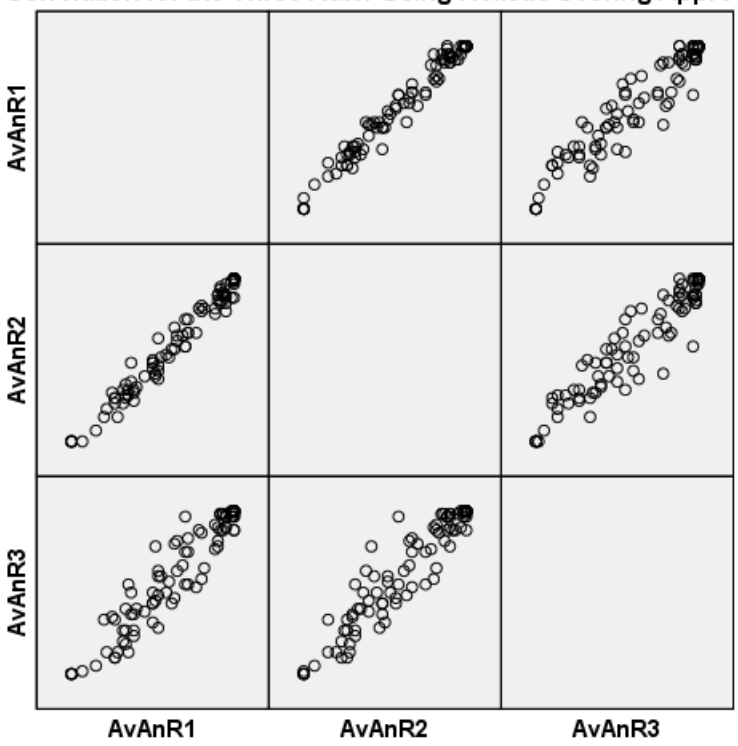
	Skewness		Kurtosis	
	Statistic	Std. Error	Statistic	Std. Error
AvAnR1	-.452	.250	-.952	.495
AvAnR2	-.369	.250	-1.091	.495
AvAnR3	-.422	.250	-1.143	.495
Valid N (listwise)				

APENDIX B

CORRELATION FOR THE THREE RATERS USING HOLISTIC SCORING APPROACHES



The Correlation for the Three Rater Using Holistic Scoring Approach



APENDIX C

EXPECTED MEAN SQUARE EQUATIONS

Effect	Variance	Expected mean square equation
Persons (*p*)	$\sigma^2 p$	$\sigma^2 ptr,e + nr\sigma^2 pt + nt\sigma^2 pr + ntnr\sigma^2 p$
Raters (*r*)	$\sigma^2 r$	$\sigma^2 ptr,e + np\sigma^2 tr + nt\sigma^2 pr + npnt\sigma^2 r$
Passages (*t*)	$\sigma^2 t$	$\sigma^2 ptr,e + np\sigma^2 tr + nr\sigma^2 pt + npnr\sigma^2 t$
Pr	$\sigma^2 pr$	$\sigma^2 ptr,e + nr\sigma^2 pr$
Pt	$\sigma^2 pt$	$\sigma^2 ptr,e + nr\sigma^2 pt$
Tr	$\sigma^2 tr$	$\sigma^2 ptr,e + np\sigma^2 tr$
Ptr,e	$\sigma^2 ptr,e$	$\sigma^2 ptr,e$

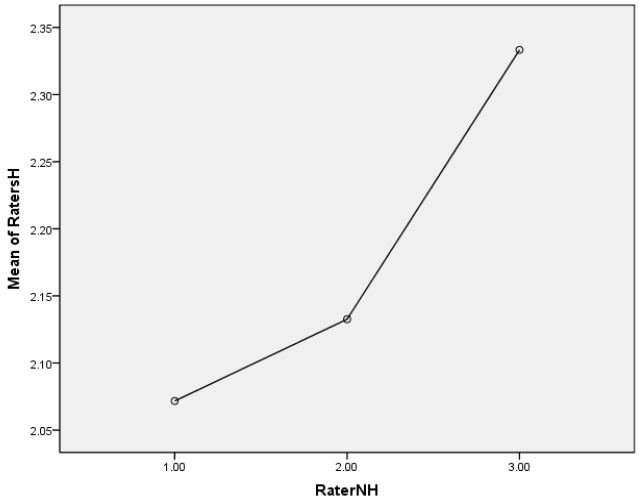
Effect	Variance	Expected mean square equation
Persons (*p*)	$\sigma^2 p$	$\sigma^2 ptr,e + nr\sigma^2 pt + nt\sigma^2 pr + ntnr\sigma^2 p$
Raters (*r*)	$\sigma^2 r$	$\sigma^2 ptr,e + np\sigma^2 tr + nt\sigma^2 pr + npnt\sigma^2 r$
Passages (*t*)	$\sigma^2 t$	$\sigma^2 ptr,e + np\sigma^2 tr + nr\sigma^2 pt + npnr\sigma^2 t$
Pr	$\sigma^2 pr$	$\sigma^2 ptr,e + nr\sigma^2 pr$
Pt	$\sigma^2 pt$	$\sigma^2 ptr,e + nr\sigma^2 pt$
Tr	$\sigma^2 tr$	$\sigma^2 ptr,e + np\sigma^2 tr$
Ptr,e	$\sigma^2 ptr,e$	$\sigma^2 ptr,e$

APENDIX D

DESCRIPTVES

Raters Holistic

	N	Mean	Std. Deviation	Std. Error
1.00	93	2.0717	1.10538	.11462
2.00	93	2.1326	1.08140	.11214
3.00	93	2.3333	1.11099	.11520



APENDIX E

ANALYTIC AND HOLISTIC SCORING RUBRICS

Analytic Scoring Rubric

التقييم	سلامة اللغة وإتقانها	أحكام النون الساكنة والتنوين وأحكام الميم الساكنة	أحكام النون والميم المشددتين	أحكام المدود	القلقلة
ضعيف جدا 0	فاحش الخطأ لايستطيع الاسترسال	لايطبق مطلقا	لايطبق مطلقا	لايطبق مطلقا	لايطبق مطلقا
سيء 1	كثير الخطأ لكنه يسترسل	نادرا مايطبق	نادرا	نادرا	نادرا
جيد 2	ثلاثة أخطاء أو بخطيء في الحركات فقط ويعاب عليه عدم إخراج الحروف من مخارجها بإتقان	يطبق قليلا من الإخفاء والإظهار والإدغام	يشدد الميم والنون على الصورة الصحيحة في قليل من المواضع	يطبق قليلا من أحكام المد ولايراعي الاتساق	يظهر القلقة في قليلا من المواضع
جيد جدا 3	خطأ واحد في اللغة *أو ماهر يغفل إخراج الحروف من مخارجها الصحيحة	*يطبق معظم الأحكام *أو قد يطبقها ولكن لا يستوفيها حقها	.*يشدد الميم والنون بالصورة الصحيحة في معظم المواضع *أو كلها مع عدم إعطائها حقها في التشديد والغنة	*يطبق معظم أحكام الم * أو قد يأتي بها جميعا لكنه لايراعي الاتساق أو عدد الحركات	يظهرالقلقلة في معظم أحرفها
ممتاز 4	يقرأ بلا أخطاء مطلقا ويخرج الحروف من مخارجها الصحيحة	يطبق كل الاحكام	يطبقها جميعا ويعطيها حقها	يطبق كل الاحكام مراعيا الاتساق وعدد الحركات	يظهر القلقة باتقان في جميع مواضعها

Holistic Scoring Rubric

المقياس الكلي:

فضلاً قم بتقييم تلاوة الطالب آخذا بالاعتبار اتقان الطالب لِ:
أحكام النون الساكنة والتنوين, أحكام الميم الساكنة,أحكام الوقف والابتداء,أحكام المدود, احكام النون والميم المشددتين, مخارج الحروف وصفاتها , وسلامة اللغة وإتقانها.

4. يظهر الطالب فهما وتطبيقا ممتازا لأحكام التجويد في التلاوة لاتشوب تلاوته أية شائبة نقص. (متقن)
3. يطبّق الطالب أكثر الأحكام التجويدية. (متمكن)
2. يطبّق الطالب بعض الأحكام التجويدية . (جيد)
1. يظهر الطالب مستوى ضعيفاً في فهم وتطبيق الاحكام التجويدية و يلحن لحنا جليا لايتجاوز الخمس مرات . (مقبول سيئ).
0. يلحن الطالب لحنا كثيرا فاحشا أو قد لايحسن القراءة. (ضعيف)

VITA
Saif Alkahtani

A teaching faculty at King Saud University
Department of Psychology
College of Education
Cell-Phone: + 966 5000 28023
Email: dr.alkahtani@hotmail.com

Educational Qualifications:

2012	Ph.D.	*Pennsylvania State University*, Educational Psychology; Measurement
2006	M.S.	*Pennsylvania State University*, Educational Psychology; Measurement
1998	B.A.	*Islamic University of I.M.I.S.*, Psychology

Career:

2000-present	*A Teaching Assistant at King Saud University*
1998-2000	*A Teaching Assistant at Teachers' College in Alahsa*

Publications:

2006	*The effectiveness of HSG and GAT in predicting FGPA in a Saudi university, a Master's Thesis*

Membership:

The National Council on Measurement in Education (NCME)
Psychometric Society