The Pennsylvania State University

The Graduate School

The Department of Educational Psychology, School Psychology, and Special Education

RATER AGREEMENT AND THE MEASUREMENT OF RELIABILITY IN EVALUATIONS OF ONLINE COURSE DESIGN USING THE QUALITY MATTERS RUBRIC^{TM}

A Thesis in

Educational Psychology

by

Whitney Alicia Zimmerman

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Master of Science

May 2011

The thesis of Whitney Alicia Zimmerman was reviewed and approved* by the following:

Jonna Kulikowich
Professor of Education
Thesis Advisor

Lawrence Ragan
Director of Faculty Development for Penn State World Campus

Rayne Sperling
Associate Professor of Education
Professor-in-Charge: Educational Psychology

*Signatures are on file in the Graduate School

ABSTRACT

Agreement between raters with diverse backgrounds in Quality Matters Program^{TM} online course reviews was examined. The Quality Matters Program uses a unique peer-review process that has both quantitative and qualitative components. Quantitative analysis compared agreement on 40 specific standards used to evaluate the quality of the design of online courses. Suggestions for measuring the reliability of the Quality Matters Program course review process in future studies are proposed.

TABLE OF CONTENTS

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to thank Jonna Kulikowich and Lawrence Ragan for the guidance that they provided as readers. I would also like to thank Kay Shattuck and Deb Adair with the Quality Matters Program for giving me the opportunity to work with them and for the assistance they provided as I was writing this thesis.

Chapter 1

Introduction

Online education is a relatively young discipline that is often thought of as a subcategory of distance education. Though online education is a somewhat new field, it is growing. In the fall of 2006 almost 3.5 million students participated in online courses and that number has been rapidly increasing with approximately 5.6 million students taking online courses in the fall of 2009 (Allen & Seaman, 2007, 2010). The Internet is now the most popular form of distance education. It has impacted community colleges greatly with the majority of online students enrolled at two-year schools, though many four-year colleges and universities also offer online courses (Cejda, 2010). With the growing number of online students it is important to have a way to assure the quality of online courses.

Research articles published as early as 1928 have provided evidence for what may be referred to as the "no significant difference phenomenon." In his book, *The No Significant Difference Phenomenon*, Russell (2001) compiled 355 research studies and found that an overwhelming proportion of those studies showed no significant difference between distance education and face-to-face education. It is important to note that online courses are thought to produce results similar to those of traditional face-to-face courses. It may be argued, however, that the courses typically studied are the most well-designed courses. With that in mind, we may still conclude that online education has the potential to produce equivalent results (Russell, 2001).

The quality of online courses has been measured by researchers using both qualitative and quantitative measures. Attrition rates and student grades are commonly used as quantitative measures of quality in online education (Garza Mitchell, 2010) as well as researcher-created surveys (Aman, 2009; Hathaway, 2009). The validity of these measures may be called into question as attrition rates and student grades may be affected by more than just course quality. Also, the reliability and validity of scores of many researcher-created surveys have not been fully examined.

The purpose of this Thesis is to examine one particular instrument intended to evaluate the design of online courses: the Quality Matters Rubric $^{\text{TM}}$. The Quality Matters Rubric reviewed in this Thesis was designed for use in peer-review evaluations of online and hybrid courses in post-secondary settings. The Rubric consists of 40 specific standards that fit into 8 general standard categories (see Appendix A and Appendix B, respectively). The standards are aimed at evaluating the design of the online components of courses. The Rubric is only meant to assess course design, not instruction and will be discussed in more detail in Chapter 2 (MarylandOnline, Inc.,2008, 2010).

The specific aspect of the Quality Matters Program $^{\text{TM}}$ that will be focused on is the issue of reliability of raters in the peer-review process. The course review process is unique in that every course is reviewed by three reviewers in the initial course review. I will examine the flaws behind using the correlations between raters or proportions of consistency as measures of reliability. In this particular case, I will argue that inter-rater agreement is not an appropriate measure of reliability.

It is important to examine the reliability of scores assigned using the Quality Matters Rubric and the process in which it is used so that it may be used in future research studies and also to validate its use by institutions around the world as a method of assuring the quality of the design of online courses.

I will begin by introducing the Quality Matters Rubric in greater detail and discussing its impact to this point in time. This will be followed by a review of literature concerning the measurement of reliability in similar cases. I will then discuss the rater consistency of the 2008-2010 Quality Matters Rubric. Finally I will suggest a method of measuring the reliability of scores of the Quality Matters Rubric for the future that takes into account the uniqueness of the process that uses multiple peer reviewers.

Chapter 2

Background

The Quality Matters Program ^{TM}

The Quality Matters Program was originally formed as a project of MarylandOnline funded by a grant from the United States Department of Education Funds for the Improvement of Postsecondary Education (FIPSE). The 3 year grant spanned from September 2003 to August 2006. Since the end of the grant period the Quality Matters Program has continued to grow as a subscription-based not-for-profit program of MarylandOnline (MarylandOnline, Inc., 2010).

The Program provides a peer-review process by online faculty for the evaluation of online and the online components of hybrid courses. Reviewers are online faculty members, online course designers, or others with experience delivering online courses. All reviewers are experienced online teachers who are trained in the use of the Quality Matters Rubric; once trained, reviewers are permitted to evaluate courses through the Quality Matters peer-review process (MarylandOnline, Inc., 2010; K. Shattuck, personal communication, January 12, 2011).

In addition to the rubric for use in higher education being reviewed here, Quality Matters has also created a rubric for assessing the quality of online and blended courses for grades 6 through 12 in a partnership with the Florida Virtual School (MarylandOnline, Inc., 2010).

The 2008-2010 Quality Matters Rubric consists of 40 specific standards (see Appendix A) that were developed by experts in the area and based on established best practices and research related to online education (Shattuck, 2007). Each specific standard is assigned a point value: 1, 2, or 3 points. Specific standards with point values of 3 are “essential standards” and must be met in order for a course to “meet standards.” In addition to meeting all 3 point standards, a course must earn at least 72 out of the possible 85 points in order to be considered meeting standards. A course that does not meet all essential standards or earns less than 72 total points does not meet standards and is given the opportunity to make changes and undergo an amended review (MarylandOnline, Inc., 2008).

Each course review is organized in one of three ways: Quality Matters managed, subscriber managed, and informally managed. The management types refer to the locus of control and the formality of the process. In Quality Matters managed and subscriber managed course reviews, a master reviewer and two other reviewers are assigned for a total of three raters per course review; one of the reviewers must be an expert in the subject matter covered by the course and at least one reviewer must be external to the institution submitting the course. The assignment of reviewers is at the discretion of the organization managing the review but is usually not completely random due to convenience and cost issues (D. Adair, personal communication, September 2010; K. Shattuck, personal communication, January 12, 2011).

Reviewers are given access to the course being reviewed and work independently as well as collaboratively to evaluate the design of the course using the Quality Matters Rubric. Communication with the course instructor during a pre-review conference call is encouraged (Kane, 2004; K. Shattuck, personal communication, January 12, 2011).

After the initial course review, each reviewer submits his or her ratings for each of the 40 specific standards. Each specific standard is rated dichotomously: it is either met or not met. Comments, including recommendations for improvements, are required for each standard that is not met. Master Reviewers are responsible for reviewing comments to make sure that comments are clear and that no comments are conflicting; if conflicts exist they may be discussed in a conference call. A majority-rule is used to determine whether or not the course has met each standard. For each standard, if at least 2 of the 3 raters said that the standard was met, then the course can be said to have met that standard. As stated previously, a course must meet all of the required 3 point standards and must earn a total of 72 points out of the possible 85 points in order to be said to be meeting standards and to be given permission to display the Quality Matters seal for the course. If a course does not meet standards after the initial review, the course may be revised and submitted for an amended review. In the amended review, the master reviewer from the initial review examines the changes made to the course and determines whether or not ratings may be changed and the course may be said to have met standards at that point (D. Adair, personal communication, January 12, 2011; MarylandOnline, Inc., 2008, 2010).

Review of Quality Matters Research

Research and published literature related to the Quality Matters Program and Rubric is relatively limited but growing rapidly. Here, a number of publications that have relied heavily on the Quality Matters Rubric and/or review process will be reviewed.

An overview of the earliest years of the Quality Matters Program was published by Shattuck (2007). According to Shattuck, “traditions of formal and informal collaboration are key considerations for understanding the development of the Quality Matters (QM) program” (¶ 3). The development of the Program is described as having four phases: the emergence of the idea; the evolution and formal development; the implementation and dissemination of the Program; and the transformation to a nationally recognized Program.

The Quality Matters Rubric standards (see Appendices A and B) have been compared to the benchmarks for the quality of online postsecondary courses developed by the National Education Association (NEA), the American Distance Education Consortium (ADEC), and the American Federation of Teachers. There are some similarities between the Quality Matters Rubric and the standards of these organizations but the Quality Matters Rubric is unique in that it is focused only on the design of online courses (Wang, 2008).

The Quality Matters Program has been used as an independent variable in empirical research studies. Dietz-Uhler, Fisher, and Han (2007) used the Quality Matters Rubric to develop and edit online courses in psychology and statistics. Their goal was to increase student retention rates which are known to be lower in online courses than in face-to-face courses. After undergoing the Quality Matters review, the average retention rates in their online psychology and statistics courses were 95.5% and 95% respectively. Unfortunately Dietz-Uhler, et al. do not report what the retention rates were in these courses before the implementation of the Quality Matters Program so there is no baseline with which to compare date. Although no clear conclusions can be drawn from their study, it contributes to the Quality Matters Program research base and may encourage others to conduct a similar but more rigorous study.

The Quality Matters Program was also used as an independent variable in the doctoral dissertation of Richard Aman (2009). Aman used a survey of student satisfaction that was created for use in his research as well as retention rates as dependent measures. The survey contained one item that measured overall student satisfaction as well as a collection of other items that measured five quality factors: (1) outcomes, (2) assessments, (3) resource materials, (4) interaction, and (5) technology.

Aman (2009) found significant results when he compared the mean of the five factors of student satisfaction scores for courses that were reviewed by the Quality Matters Program to courses from institutions that do not participate with the Quality Matters Program: *t* =1.54 (552), p = .06 (an alpha level of .10 was used). It should also be noticed that an alpha level of .10 was used as opposed to the standard .05. Students in the courses reviewed by the Quality Matters Program gave an average satisfaction rating of 4.12 (SD = 0.62) which students in the courses not reviewed gave an average satisfaction rating of 4.04 (SD = 0.59) on a 5 point scale.

An interesting finding of Aman (2009) was that student comfort with distance learning, student age, and student gender were all significant confounding variables with the mean of the five factors of student satisfaction. To take into account the confounding variables a one-way ANCOVA was conducted to compare the mean of the five factors of student satisfaction as measured by his survey for courses that had been reviewed by the Quality Matters Program to those that had not been reviewed while controlling for student comfort with distance learning, student age, student gender, and prior experience with distance learning (measured by number of distance courses taken). It is unclear why prior experience with distance learning was used as a covariate as it was found to not be significant in the previous analyses. Results were significant: F(1, 492) = 4.62, p = .03, one-tail test. Students in courses that had been reviewed by the Quality Matters Program were more satisfied than those in courses that were not reviewed when student comfort level, age, gender, and prior experience were taken into account.

To summarize the findings of Aman (2009), although the retention rates were not different for courses reviewed by the Quality Matters Program and courses not reviewed, significant differences were found in the mean student satisfaction rates when taking into account the confounding variables of student comfort with distance learning, student age, student gender, and number of previous distance learning courses taken.

Quality Matters Program grants for research.

In 2009 the Quality Matters Program offered the opportunity to those associated with the organization to apply for grants to support research related to the Program. Five grants were awarded and the recipients of the grants presented their findings at the Second Annual Quality Matters Conference in June of 2010. Those five studies that were funded by the Quality Matters Program will be reviewed here.

Swan, Matthews, Boles, and Bogle (2010) examined the link between course design and student learning processes. They used the Quality Matters Rubric while applying the Community of Inquiry (CoI) framework in their study. Though many of their findings were not statistically significant, they did suggest that student learning improved as a result of the implementation of the Quality Matters process. The implementation of the Rubric may lead instructors to be aware of some of the issues related to the design of their course such as the use of course objectives.

Hall (2010) also applied the CoI framework in her study in which she paralleled the Quality Matters Rubric with CoI's teaching presence factor of design and organization. She found that the implementation of the Quality Matters Rubric increases teacher and teaching presence. It reduces the necessary amount of management required by the instructor as well as students' self-management by improving the quality of the course design. The implementation of the Quality Matters Rubric was also found to have a positive effect on students' course satisfaction ratings and discussion board grades.

Knowles and Kalata (2010) examined the influence that the Quality Matters Rubric has on students' perceptions of online courses. Two courses from different disciplines were examined. Initially both courses did not meet Quality Matters standards Some sections of the courses were taught before the courses were amended to meet standards and some were taught after the courses were designated as meeting all 40 specific standards. Students were asked to review the course that they were taking. Knowles and Kalata compared the ratings given by the students to those given by the certified master reviewer. While the reviewer found the initial sections of the courses to not meet standards and the latter sections to meet standards, students in all sections of the course rated it as meeting standards. Students tended to rate all courses highly, even the earlier sections that did not meet Quality Matters standards per the master reviewer. Response set may be a threat to validity in this study as it was stated that the majority of responses on all questions was yes. Other possible explanations are that students have lower expectations than the master reviewer or that students did not take the time to access the descriptions of all of the specific standards that they were to be using when rating the courses.

The fourth grant project by Alarcon, Strickland, and Rodrigo (2010) was extended past the date of the conference and is a review of the use of the Quality Matters process at Maricopa County Community College District located in the state of Arizona. Alarcon, et al. examined the organizational factors related to implementing the Quality Matters Program at a large community college system that is made up of multiple colleges. The fifth grant project was a pilot study of Quality Matters' general standard 8 (accessibility) discussed by Bowen (2010).

Quality Matters has awarded four research grants for the 2011 year. Findings from those studies will be presented at the 2011 Quality Matters Conference. Topics to be covered include the impact of the Quality Matters process on attrition, preparing faculty to participate with Quality Matters, the development of technology pedagogical content knowledge, and the evaluation of the Quality Matters process on student and faculty perception of course quality (MarylandOnline, Inc., 2010).

Current state of research.

Research on the Quality Matters Program as well as research that uses the Quality Matters Rubric and Program as a measure of online course design quality is growing. Grants provided by the Program should continue to encourage research on this topic though possible bias must be acknowledged in these studies as they are receiving some funding from the Program. Research conducted by those not receiving funding from the Program should continue as well such as what we have seen in the publications of Dietz-Uhler, Fisher, and Han (2007), Wang (2008), and Aman (2009).

The perspective of those publishing works related to the Quality Matters Program may also be of interest. Many of those receiving grants from the Program to conduct their research are primarily practitioners and many are primarily working in community college settings. The setting that the research is conducted in may bear heavily on its findings. Therefore, it is important to examine the differences between different demographics. As found by Aman (2009), factors such as student age, gender, and comfort with online learning may be related to student perception of quality.

Reliability

Reliability is a term used to describe the consistency of collected data. Statistical measures of reliability include test-retest, parallel forms, internal consistency, and interrater reliability. Some of these measures are not appropriate for use with the Quality Matters process due to their assumptions, others may be appropriate (Colton & Covert, 2007; Merriam-Webster, Inc., 2011).

The test-retest method would measure reliability by finding the correlation between the ratings given by the same course reviewer at two different periods in time. Given that the course being reviewed does not change, it is appropriate to assume that the ratings given by the rater should stay consistent (Colton & Covert, 2007).

The parallel forms method of measuring reliability cannot be applied to the Quality Matters Program because only one appropriate and current form of the Rubric exists. It may be possible to compare ratings from two different editions of the form, but that would not be appropriate as changes have been made to improve the Rubric from one year to the next.

Next is the internal consistency method of measuring reliability. “An important caveat about tests of internal consistency reliability is that they *apply only to multi-item scale* (because this approach assumes that all the items are measures of the same construct) and not to items that function as independent measures" (Colton & Covert, 2007, p. 79). This approach would not be appropriate because even within subscales unidimensionality is cannot be assumed, it does not make sense to use measures of internal consistency with the Quality Matters Rubric. Instead, the reliability of raters is the primary method by which the consistency of scores can be evaluated.

Inter-rater, or inter-scorer, reliability is often defined as "the degree of agreement about individuals' performance among different scorers of a test" (Hogan, 2007, p. 656). This definition says nothing about the relationship between the raters. The term "inter-rater reliability" could refer to a number of different instances. A more specific definition is necessary because there are some assumptions that come with the traditional view of inter-rater reliability that are often overlooked. One is that raters are interchangeable; it is expected that all raters should be giving the same scores. Traditional views also assume that all raters are looking at the same things. Reliability is often viewed as a prerequisite for validity. If the previously mentioned assumptions are met, then inter-rater reliability is a prerequisite for validity. If these assumptions are not met, as *may* be the case with the Quality Matters Rubric, where the raters have different points of view, then this is not truly reliability and is therefore not a prerequisite for validity. It may even be possible that high inter-rater correlations are not desired if we expect raters to be looking at courses from different perspectives (Aiken &Groth-Marnat, 2006; Suen, et al., 1995).

According to Suen et al. (1995), there is a second perspective of conventional inter-rater reliability that appears to be similar to the aforementioned perspective: raters must be randomly selected. The ratings of any randomly selected raters should be consistent. In many cases this is not applicable because raters may intentionally be drawn from populations with different backgrounds.

There are different types of inter-rater correlations: that in which raters are expected to be the same, that in which raters are expected to be different, and that in which one rater is an expert and the others are novices. The first, when raters are considered to be equivalent, is similar to parallel-forms reliability; the raters are comparable to the different forms of the assessment. The second, when raters are different, is a type of concurrent validity. The third possibility, expert-novice pairings, assumes that the expert is always correct and is therefore only measuring the accuracy of the novice (Hoi Suen, Personal Communication, April 2009).

I would consider these three possible combinations of raters to be three completely different concepts. For the sake of communicating clearly I would not call any of these three examples inter-rater reliability. The purpose of this is to demonstrate the need for more precise language when discussing what is currently labeled “inter-rater reliability.” It is not enough for one to say that a test lacks inter-rater reliability; more information is needed as to what kind of inter-rater reliability was performed.

I will clarify the three types of inter-rater reliability. The first type will be parallel raters; this is when the raters are considered to be equivalent. The second would be based on information about the raters, such as parent-teacher correlational validity or teacher-peer correlational validity. The third would be the novice reliability; this is the correlation between the expert, who is assumed to always be correct, and a novice.

Because inter-rater reliability does not exist as a unitary concept, it should not be communicated as a unitary concept. It is necessary to have multiple terms to describe these different cases. The correlation between two nonequivalent raters is not a form of reliability and therefore should not be described as one.

As presented in this Thesis, rater agreement is different from inter-rater reliability in that it is calculated as the proportion of time that all raters agree on the score to be given. This is in contrast to an inter-rater reliability coefficient which is typically a correlation coefficient.

It is arguable which type of inter-rater correlation is most applicable in the case of the Quality Matters' Rubric. This depends on the view of the individual. The Rubric was developed to be used by a team of raters. The fact that raters may communicate with one another before submitting their ratings makes it difficult to validly measure the reliability of the instrument itself. What is more important is the reliability and validity of the entire process. "The principles of QM [Quality Matters] include collegiality, inter-institutional, faculty-driven, and … continuous quality improvement" (K. Shattuck, personal communication, January 12, 2011). The overall goal is to improve the quality of online courses via these principles. The validity of using the numeric scores as a measure of reliability does not take into account the values of the Quality Matters Program (D. Adair, personal communication, January 12, 2011).

Suggested techniques for measuring the reliability of the Quality Matters Rubric will be discussed in greater detail in Chapter 4.

Chapter 3

Analyses of The Quality Matters Rubric^{TM} : Methods and Results

Data Collection

The data used in the following analyses were collected by Quality Matters and were received with all identifying information removed. Information concerning the type of course review (Quality Matters, subscriber, or informally managed), score after initial course review, score after amended course review, and course review status (standards met after initial review, standards met after amended course review, or standards not yet met) were obtained from Quality Matters. Also, for every course review and for each specific standard the number of raters who rated the course as meeting the standard and the number of raters who did not rate the course as meeting the standard was received from Quality Matters. The ratings from individual raters were not received.

Initial Analyses

Before analyses of rater agreement were performed, more general descriptive statistics were examined as well as a comparison of the three review types. Table 1 presents the proportion of courses meeting each specific standard in the initial course review. The mean proportion of courses meeting a specific standard in initial course review is .821 with a standard deviation of 0.103; proportions range from .54 to .96.

Table 1: Standards' Rates of Being Met in Initial Reviews

Specific Standard	Courses Meeting in Initial Review	Specific Standard	Courses Meeting in Initial Review
1.1	.78	4.4	.87
1.2	.88	5.1	.85
1.3	.83	5.2	.91
1.4	.75	5.3	.63
1.5	.86	5.4	.84
1.6	.80	6.1	.96
1.7	.67	6.2	.91
2.1	.87	6.3	.77
2.2	.71	6.4	.93
2.3	.85	6.5	.96
2.4	.67	6.6	.89
2.5	.91	6.7	.81
3.1	.84	7.1	.89
3.2	.81	7.2	.82
3.3	.63	7.3	.76
3.4	.94	7.4	.82
3.5	.60	8.1	.80
4.1	.89	8.2	.54
4.2	.85	8.3	.91
4.3	.95	8.4	.86

The three review types were compared in terms of the total number of points achieved in initial review. The point values assigned to each specific standard are listed in Appendix A. This analysis was performed prior to the analysis of rater agreement in order to assure that there were not differences between review types. It was determined a priori that informally managed reviews would not be included in the analysis of rater agreement therefore in addition to a one-way ANOVA comparing the three review types two contrasts were also planned; the first to compare formally managed (Quality Matters and subscriber managed) to informally managed and the second to compare Quality Matters and subscriber managed course reviews.

Table 2: Total Points Earned in Initial Course Reviews by Review Type

Managed by:	N	Mean	Standard deviation
Quality Matters	138	70.36	12.585
Subscriber	136	71.60	12.128
Internal	180	70.16	14.387
Total	454	70.65	13.188

Table 2 contains the descriptive statistics concerning the total number of points earned in initial course reviews by review type. A one-way ANOVA was conducted and results were not significant, F(2, 451) = 0.505, *p* = .604; the mean number of total points earned in initial course reviews did not differ for the three different review types. Pre-planned post-hoc tests using contrasts were performed. The first compared formally managed (Quality Matters and subscriber managed) to informally managed course reviews and was not significant, *t*(451) = 0.646, *p* = .519; total points earned in initial course reviews did not differ for formally and informally managed course reviews. Finally, Quality Matters managed and subscriber managed course reviews were compared and again results were not significant, *t*(451) = 0.773, *p* = .440; total points earned in initial course reviews did not differ for Quality Matters and subscriber managed course reviews. Quality Matters managed and subscriber managed course reviews will be combined in the analyses of rater agreement.

Agreement by Specific Standard

Rater agreement was calculated using the initial reviews of 282 courses. These are the ratings submitted at the end of the initial course review process. Only subscriber managed and Quality Matters managed reviews were included and only those with at least 110 submitted values (out of a possible 120; 40 items x 3 raters) were included.

Any specific standard not reviewed by all 3 raters was considered to be “missing”; missing values were excluded piecewise and therefore sample sizes vary for each of the specific standards. Valid sample sizes vary from 261 to 282. The proportion of course reviews in which all 3 raters gave a course the same rating on a specific standard is presented in Table 3.

Table 3: Rater Agreement by Specific Standard

Specific Standard	Proportion Rater Agreement
1.1	.85
1.2	.90
1.3	.81
1.4	.85
1.5	.90
1.6	.78
1.7	.71
2.1	.86
2.2	.81
2.3	.85
2.4	.75
2.5	.86
3.1	.85
3.2	.80
3.3	.75
3.4	.87
3.5	.70
4.1	.88
4.2	.80
4.3	.93
4.4	.75
5.1	.86
5.2	.84
5.3	.75
5.4	.80
6.1	.91
6.2	.88
6.3	.79
6.4	.91
6.5	.94
6.6	.83
6.7	.70
7.1	.82
7.2	.80
7.3	.74
7.4	.73
8.1	.85
8.2	.57
8.3	.82
8.4	.75

Proportion of rater agreement ranges from .57 to .94. The mean rater agreement proportion is .814 with a standard deviation of 0.075 and the median is .82. It is important to stress that these numbers are not inter-rater reliability coefficients. These are the proportions of course reviews in which all three course reviewers submitted the same score for that specific standard. Inter-rater reliability coefficients were not calculated with the given dataset because the ratings of specific course reviewers were not available.

It may be meaningful to compare the proportions of rater agreements for different standards but it must be done with several considerations. There may be situations in which course reviewers would be expected to disagree. For example, course reviewers coming from diverse backgrounds may consider the course from different perspectives and thus disagree on the ratings to be given. It is for that reason that rater agreement may not always be expected or even desired. On the other hand, a very low rater agreement rate may mean that the standard is unclear or too subjective which is also not desirable as it affects the overall reliability and validity of the Rubric.

There are 6 specific standards with rater agreement rates greater than or equal to .90 as seen in Table 4. Half of those specific standards come from general standard 6: course technology. The 11 specific standards with rater agreement rates at or below .75 are presented in Table 5. Interestingly, there is a moderately high correlation between the proportion of courses meeting each standard presented in Table 1 and the proportion of rater agreement presented in Table 3. The Pearson's product-moment correlation coefficient between the two is .710 (N = 40), r^2 = .5041. Specific standards that are met in more initial reviews are agreed upon in more initial reviews.

Table 4: Specific Standards with Rater Agreement ≥ .90

Specific Standard	**Text**	**Rater Agreement Rate**
6.5	The course components are compatible with current standards for delivery modes.	.94
4.3	The instructional materials have sufficient breadth, depth, and currency for the student to learn the subject.	.93
6.1	The tools and media support the learning objectives, and are appropriately chosen to deliver the content of the course.	.91
6.4	Students have ready access to the technologies required in the course.	.91
1.2	A statement introduces the student to the purpose of the course and to its components; in the case of a hybrid course, the statement clarifies the relationship between the face-to-face and online components.	.90
1.5	Students are asked to introduce themselves to the class.	.90

Table 5: Specific Standards with Rater Agreement ≤ .75

Specific Standard	Text	Rater Agreement Rate
8.2	Course pages and course materials provide equivalent alternatives to auditory and visual content.	.57
3.5	Self-check or practice assignments are provided, with timely feedback to students.	.70
6.7	The course design takes full advantage of available tools and media.	.70
1.7	Minimum technical skills expected of the student are clearly stated.	.71
7.4	Course instructions answer basic questions related to research, writing, technology, etc., or link to tutorials or other resources that provide the information.	.73
7.3	Course instructions articulate or link to an explanation of how the institutions student support services can help students reach their educational goals.	.74
2.4	Instructions to students on how to meet the learning objectives are adequate and stated clearly.	.75
3.3	Specific and descriptive criteria are provided for the evaluation of students work and participation.	.75
4.4	All resources and materials used in the course are appropriately cited.	.75
5.3	Clear standards are set for instructor responsiveness and availability (turn-around time for email, grade posting, etc.)	.75
8.4	The course ensures screen readability.	.75

Agreement by Course Review

Rater agreement was also calculated for each course review. The distribution of rater agreement rates is presented in Table 6. The rater agreement rates for course reviews ranged from .45 agreement to 1.00 agreement with a median rater agreement rate of .85 (N=174).

Table 6:Proportion of Rater Agreement by Course Review

Proportion Rater Agreement	Frequency
0.45	3
0.48	2
0.55	2
0.58	3
0.60	2
0.63	2
0.65	6
0.68	4
0.70	6
0.73	6
0.75	4

Proportion Rater Agreement	Frequency
0.78	13
0.80	14
0.83	11
0.85	13
0.88	17
0.90	17
0.93	13
0.95	11
0.98	11
1.00	14

Possible Problems with Data

As with the previous analyses, the data in this analysis were collected at the end of the initial course review. This is after the reviewers have had the opportunity to discuss the course. The fact that there were numerous course reviews that had perfect (1.00) rater agreement rates calls into question the validity of using these scores as a measure of the raters true opinion. It is possible that raters changed their initial ratings after discussions with their two fellow raters. The rater agreement rates presented in Tables 3 through 5 may be inflated if some disagreeing raters changed their initial ratings after collaborating with their fellow raters and before submitting their ratings to Quality Matters' online system.

Chapter 4

Discussion

The Quality Matters Program course review process is unique in that it is a peer-review process that involves both quantitative and qualitative evaluations. The qualitative comments are highly valuable to course instructors and designers trying to improve the quality of their courses. To some, this is the primary purpose of the Quality Matters Program and the quantitative values are secondary. However, the numerical scores are still important as they are the determining factor as to whether a course meets standards or does not meet standards. There is some conflict between these values.

Because the Quality Matters Program seal is awarded solely on the basis of quantitative measures, it is important to examine some of the psychometric properties of the Quality Matters Rubric. The purpose of this Thesis was to examine the reliability of the Quality Matters Rubric. The analyses that could be performed were limited due to the nature of the data that were received. The data contained in these analyses were collected before the author was in contact with the organization.

The reliability of the Quality Matters Program course review process is of interest here as opposed to the Rubric itself due to how the evaluations are used. While the Rubric is used to evaluate courses on a scale of 0 to 85 with comments, what the public sees is simply whether or not the course has met standards or not which is dichotomous in nature. It must be taken into account that groups of diverse raters are used in each course review. In this case, the reliability of the process may be determined by examining the extent to which the final results are the same for different groups of course reviewers.

Any method for measuring the reliability of the Quality Matters Rubric and process should take into account the value of both the numerical scores and the qualitative comments. The reliability of individual reviewers' scores, final results (met/did not meet standards), and comments should be examined.

Reliability of Numeric Scores

The reliability of the numeric scores could be measured using traditional inter-rater reliability techniques. The correlations between course reviewers could be calculated and used as a measure. Another option would be calculate the correlation between numeric scores given by reviewers with the numeric scores given by a set of experts on practice course reviews; the assumption here is that the experts' ratings would be the correct ratings. As stated in an earlier chapter, this may not be the most appropriate measure of reliability because the diversity of course reviewers is a strength of the Quality Matters process. Perfect agreement may not be expected or desirable.

The reliability of scores given by individual raters could be measured using test-retest techniques. Raters could be asked to evaluate a course at two different time points. Assuming that no changes were made to the sample course, the ratings made at both time points should be the same. This method of measuring reliability would not interfere with the desire for diverse course reviewers and thus should be taken as a suggestion.

Reliability of Final Results

In the end, a course either meets standards or does not meet standards. Only courses that meet standards may bear the Quality Matters Program seal. The reliability of the final results could be examined by having multiple sets of raters reviewing the same courses. Regardless of what set of raters a course is assigned, the final results should be the same. There is a problem if the final results of a course review are dependent on the group of reviewers.

The reliability of the final results is important as it is related to the validity of the Quality Matters Program seal. The value of the seal will be fortified if it can be found that the awarding of the seal is not dependent on the raters chosen for that particular review. This is particularly of interest because course reviewers are not always randomly selected. This should not interfere with the desire to have diverse individual raters because we are looking at sets of raters. Therefore inter-team reliability is a suggested method for evaluating the reliability of the final results of the Quality Matters process.

Reliability of Comments

In addition to the dichotomous ratings given for each specific standard, course reviewers may leave related comments. While these comments are not directly taken into account when determining whether or not a course has met standards or not, they are received by the course designers who may take them into account when revising the course for an amended review if necessary.

There was some discussion at the 2010 Quality Matters Conference concerning the reliability of these comments. Some do not see the reliability of comments as being necessary because diversity of course reviewers is desirable; everyone should have unique opinions and comments. On the other hand, every course should be receiving high quality, valuable comments regardless of who the three reviewers of the course are. Because course reviewers are not randomly selected, the reliability of the quality of the comments should be addressed. A mixed-methods study in which experts examine the comments submitted through the review process may be appropriate.

Conclusions

In this Thesis is presented some of the analyses that have been performed on the 2008-10 edition of the Quality Matters Rubric. In the presented analyses it is important to take into account the data collection methods. The rater agreement rates may be inaccurate as course reviewers were able to communicate with one another before submitting their evaluations. A future study that examines the opinions of course reviewers before they collaborate with others may be of interest. It is possible that some reviewers opinions are influenced by others and their initial opinions may be different.

Even if raters were swayed by one another, it is still evident that rater agreement is higher on some standards than on others. Standard 8.2 is of particular interest as its rater agreement rate is much lower than the agreement rates of all other standards; it was also the most frequently missed standard.

Several suggestions for examining the reliability and validity of the Quality Matters Rubric were made. A test-retest measure of the ratings given by individual course reviewers may be of interest to assure that individuals are consistent. The reliability of the final results (meeting standards or not meeting standards) should be examined by having multiple sets of reviewers rating the same courses to assure that the final results are consistent regardless of what set of reviewers a course is assigned. The reliability of the quality of comments should be assessed by experts using a combination of qualitative and quantitative methods.

The Quality Matters Program has brought attention to the need for quality assurance in the area of online education. As seen in Chapter 2, the Program has also influenced the research being done in the field. The Program should continue to encourage research in the area and should continue to be used as a basis for studies independent of the organization.

REFERENCES

Aiken, L. R., & Groth-Marnat, G. (2006).*Psychological Testing and Assessment*. Boston: Pearson Education Group, Inc.

Allen, I. E., & Seaman, J. (2007). Online nation: Five years of growth in online learning.*The Sloan Consortium.* Retrieved from http://www.sloanconsortium.org/sites/default/files/online_nation.pdf

Allen, I. E., & Seaman, J. (2010). Class differences: Online education in the United States, 2010. *The Sloan Consortium.* Retrieved from http://sloanconsortium.org/sites/default/files/class_differences.pdf

Aman. R. R. (2009). *Improving student satisfaction and retention with online instruction through systematic faculty peer review of courses.* (Unpublished doctoral dissertation). Oregon State University, Corvallis, OR.

Bowen, E. (2010). *QM Standard 8 Pilot Project.*Presentation at the Second Annual Quality Matters Conference, Chicago, IL.

Cejda, B. (2010). Online education in community colleges. In R. L. Garza Mitchell (Ed), *Online Education* (pp. 7-16). Hoboken, NJ: John Wiley & Sons, Inc

Colton, D., & Covert, R. W. (2007). *Designing and constructing instruments for social research and evaluations.* San Francisco: Jossey-Bass.

Dietz-Uhler, B., Fisher, A., & Han, A. (2007-2008). Designing online courses to promote student retention. *Journal of Educational Technology Systems, 36*(1), 105-112.

Garza Mitchell, R. L. (2010). Approaching common ground: Defining quality in online education. *New Directions for Community College, 150,* 89-94.

Hall, A. (2010, June). *Quality Matters Rubric as "Teaching Presence": Application of Community of Inquiry Framework to Analysis of the QM Rubric's Effects on Student Learning*. PowerPoint presented by Kay Shattuck at the Second Annual Quality Matters Conference, Chicago, IL.

Hathaway, D. M. (2009). *Assessing quality dimensions and elements of online learning enacted in a higher education setting* (Doctoral dissertation, George Mason University). Available from ProQuest Dissertations and Theses database. (UMI No. 3367067)

Hogan, T. P. (2007). *Psychological testing: A practical introduction.* Hoboken, NJ: John Wiley & Sons, Inc.

Kane, K. (2004). Quality Matters: Inter-institutional quality assurance in online learning. *Sloan-C View, 3*(11). Retrieved from http://sloanconsortium.org/publications/view/v3n11/pdf/v3n11.pdf

Knowles, E., & Kalata, K. (2010, June).*The Impact of Quality Matters Standards on Student Perception of Online Courses.*Presentation at the Second Annual Quality Matters Conference, Chicago, IL.

Little, B. (2009). The use of standards for peer review of online nursing courses: A pilot study. *Journal of Nursing Education, 48*(7), 411-416.

MarylandOnline, Inc. (2008) *Quality Matters™ rubric for online and hybrid courses: 2008-2010 edition.*

MarylandOnline. (2010). Quality Matters Program. Retrieved from http://www.qmprogram.org

Merriam-Webster, Inc. (2011). Reliability. Retrieved from http://www.merriam-webster.com/dictionary/reliability

Pollacia, L., & McCallister, T. (2009). Using web 2.0 technologies to meet quality matters[TM] (QM) requirements. *Journal of Information Systems Education, 20*(2), 155-164.

Russell, T. L. (2001). *The No Significant Difference Phenomenon.*International Distance Education Certification Center (IDEC).

Shattuck, K. (2007). Quality Matters: Collaborative program planning at a state level. *Online Journal of Distance Learning Administration, X*(III). Retrieved from http://www.westga.edu/~distance/ojdla/fall103/shattuck103.htm

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.

Swan, K., Matthews, D., Boles, E., & Bogle, L. (2010, June).*Linking Course Design to Learning Processes Using the Quality Matters and Community of Inquiry Frameworks*.Presentation at the Second Annual Quality Matters Conference, Chicago, IL.

Wang, H. (2008). Benchmarks and quality assurance for online course development in higher education. *US-China Education Review, 5*(3), 31-34.

Appendix A: Quality Matters Rubric- 40 Specific Standards

Specific Standard	Point Value	Text
1.1	3	Instructions make clear how to get started and where to find various course components.
1.2	3	A statement introduces the student to the purpose of the course and to its components; in the case of a hybrid course, the statement clarifies the relationship between the face-to-face and online components.
1.3	1	Etiquette expectations (sometimes called netiquette) for online discussions, email, and other forms of communication are stated clearly.
1.4	1	The self-introduction by the instructor is appropriate and available online.
1.5	1	Students are asked to introduce themselves to the class.
1.6	1	Minimum student preparation, and, if applicable, prerequisite knowledge in the discipline are clearly stated.
1.7	1	Minimum technical skills expected of the student are clearly stated.
2.1	3	The course learning objectives describe outcomes that are measurable.
2.2	3	The module/unit learning objectives describe outcomes that are measurable and consistent with the course-level objectives.
2.3	3	All learning objectives are stated clearly and written from the students perspective.
2.4	3	Instructions to students on how to meet the learning objectives are adequate and stated clearly.
2.5	2	The learning objectives are appropriately designed for the level of the course.
3.1	3	The types of assessments selected measure the stated learning objectives and are consistent with course activities and resources.
3.2	3	The course grading policy is stated clearly.
3.3	3	Specific and descriptive criteria are provided for the evaluation of students work and participation.
3.4	2	The assessment instruments selected are sequenced, varied, and appropriate to the content being assessed.
3.5	2	Self-check or practice assignments are provided, with timely feedback to students.
4.1	3	The instructional materials contribute to the achievement of the stated course and module/unit learning objectives.
4.2	3	The relationship between the instructional materials and the learning activities is clearly explained to the student.
4.3	2	The instructional materials have sufficient breadth, depth, and currency for the student to learn the subject.
4.4	1	All resources and materials used in the course are appropriately cited.

5.1	3	The learning activities promote the achievement of the stated learning objectives. (Note: in some institutions learning objectives may be called learning outcomes.)
5.2	3	Learning activities foster instructor-student, content-student, and if appropriate to the course, student-student interaction.
5.3	2	Clear standards are set for instructor responsiveness and availability (turn-around time for email, grade posting, etc.)
5.4	2	The requirements for student interaction are clearly articulated.
6.1	3	The tools and media support the learning objectives, and are appropriately chosen to deliver the content of the course.
6.2	3	The tools and media support student engagement and guide the student to become an active learner.
6.3	3	Navigation throughout the online components of the course is logical, consistent, and efficient.
6.4	2	Students have ready access to the technologies required in the course.
6.5	1	The course components are compatible with current standards for delivery modes.
6.6	1	Instructions on how to access resources at a distance are sufficient and easy to understand.
6.7	1	The course design takes full advantage of available tools and media.
7.1	2	The course instructions articulate or link to a clear description of the technical support offered.
7.2	2	Course instructions articulate or link to an explanation of how the institutions academic support system can assist the student in effectively using the resources provided.
7.3	1	Course instructions articulate or link to an explanation of how the institutions student support services can help students reach their educational goals.
7.4	1	Course instructions answer basic questions related to research, writing, technology, etc., or link to tutorials or other resources that provide the information.
8.1	3	The course incorporates ADA standards and reflects conformance with institutional policy regarding accessibility in online and hybrid courses.
8.2	2	Course pages and course materials provide equivalent alternatives to auditory and visual content.
8.3	2	Course pages have links that are self-describing and meaningful.
8.4	1	The course ensures screen readability.

(From MarylandOnline, Inc., 2008)

Appendix B: Quality Matters Rubric- 8 General Standards

1. Course Overview and Introduction
2. Learning Objectives (Competencies)
3. Assessment and Measurement
4. Resources and Measurement
5. Learner Engagement
6. Course Technology
7. Learner Support
8. Accessibility

From MarylandOnline, Inc., 2008