Spring 2010

The Effects of Training in Peer Assessment on University Students' Writing Performance and Peer Assessment Quality in an Online Environment

Yun Xiao
Old Dominion University

THE EFFECTS OF TRAINING IN PEER ASSESSMENT ON UNIVERSITY STUDENTS'

WRITING PERFORMANCE AND PEER ASSESSMENT QUALITY IN AN ONLINE

ENVIRONMENT

by

Yun Xiao
B.S February 1982, Yunnan Normal University, P. R. China
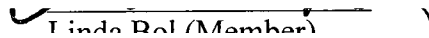M.S. May 2004, Old Dominion University

A Dissertation Submitted to the Faculty of
Old Dominion University in Partial Fulfillment of
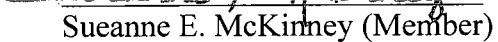Requirement for the Degree of

DOCTOR OF PHILOSOPHY
CURRICULUM AND INSTRUCTION

OLD DOMINION UNIVERSITY
May 2010

Approved by:

Linda Bol (Member)

Sueanne E. McKinney (Member)

ABSTRACT

THE EFFECTS OF TRAINING IN PEER ASSESSMENT ON UNIVERSITY STUDENTS' WRITING PERFORMANCE AND PEER ASSESSMENT QUALITY IN AN ONLINE ENVIRONMENT

Yun Xiao
Old Dominion University, 2010
Director: Dr. Robert Lucking

This study was designed to examine the effects of peer-assessment skill training on students' writing performance, the quality of students' feedback, the quality (validity and reliability) of student-generated scores, and the students' satisfaction with the peer assessment method in an online environment. A quasi-experimental design was employed to test group differences on the dependent variables. Four hundred and seventy-three sophomore and junior undergraduate students who were enrolled in a Foundations of Education course were selected by convenience sampling at a Large East-Coast Urban University. Students enrolled in Spring and Fall semesters of 2008 were assigned to the two experimental groups that received principle-based peer-assessment skill training or target-criteria-based peer-assessment skill training, while students enrolled in Fall semester of 2007 were assigned to the comparison group and did not receive structured peer-assessment skill training.

The results of the study indicated that students who had peer assessment skill training in the experimental groups outperformed their counterparts in the comparison group on writing performance, and provided higher quality written feedback to their peers than those in the comparison group. The findings revealed that students in experimental groups generated more reliable assessment scores than those in the

comparison group in the second round of peer assessment. The findings also revealed that students in the target-criteria-based training groups exhibited a higher level of satisfaction with peer feedback than those in the other groups. In addition, the results indicated that use of the target-criteria-based training method had no apparent superiority to use of principle-based training method on students' writing performance, and peer-assessment skill training had no apparent positive impacts on the validity of student-generated assessment scores during peer assessment.

To my parents

and

Kun, Kangming, Yin, Xuejie, Changfeng, Yunfan

For supporting me through the process, you have my love, always

ACKNOWLEGEMENTS

I would like to take this opportunity to thank Dr. Robert Lucking, my advisor and committee chair. Without your guidance and support throughout the years, I would never have reached this point. Your philosophy, insights, kindness, and encouragement were essential for me to complete this journey. I feel honored and privileged to have worked with you, and look forward to continuing our friendship.

I would also like to thank Dr. Linda Bol. Your feedback and support have made this research much more meaningful and significant. Without your guidance, expertise, helpful critiques and support along the way, this research would have never reached this high quality.

Dr. Sueanne E. McKinney provided tremendous insight and guidance for this research. Without your support and encouragement, it would be impossible for me to complete this research.

I am very grateful to Dr. Dwight Allen for guiding me in my academic journey. Your philosophy and insights in education has inspired me tremendously. Without your strong support and encouragement, I would never be able to reach this point.

I would also like to thank Dr. Kidd for providing me tremendous support and help for my dissertation project. Without your help and cooperation, this research would have never been conducted and completed.

Finally, I would like to thank my friends, Mr. Peter Baker and Dr. Kangming Ma for spending much time to help me polish my dissertation and data collection. Without your support and help, this dissertation would have never been readable.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER 1: INTRODUCTION

Introduction to the Chapter

Writing is a very important skill in university students' academic studies. Traditionally, essay writing and written examinations are the two main methods whereby university students are assessed (Smith, Campbell, & Brooker, 1999). One of the most important reasons essay writing is standard practice in many undergraduate courses is that it is believed to promote higher-order thinking (Smith et al., 1999) and is associated with deep learning. Learning how to write at a proper academic level is an integral part of university education and is a key part of the particular disciplinary content; accordingly, mastery of academic writing skills is a long-term process (Berg, Admiraal, & Pilot, 2006). In an attempt to discover more effective ways to help students become better writers, teachers and researchers have explored and experimented with various instructional strategies in many curricula. Peer assessment is one of the effective methods that has been used in universities to improve students' academic paper writing skills for many years, and its benefits to both students and instructors has been well documented (Berg et al., 2006; Boud, Cohen, & Sampson,1999; Davies, 2006; Falchikov, 1995a; Stefani, 1994; Topping, 1998).

The present study investigates the effects of training in peer-assessment on students' writing performance, the quality of peer assessment, and the student satisfaction with peer assessment. It compares the differences of student writing performance, quality of student qualitative feedback, the quality of student-generated rating scores, and their satisfaction with peer assessment among three groups of students each of which received different forms of structured peer-assessment skill training by using a quasi-experimental

design. This first chapter of the study consists of the background of the study, description of its significance, an overview of the methodology that is used, and delimitations of the study.

Background of the Study

For more than fifty years, peer assessment aimed at assisting student learning has been widely used in many institutions (Sluijsmans, Brand-Gruwel & Van Merrienboer, 2002). Peer assessment has been defined as an arrangement in which individuals consider the amount, level, value, worth, quality, or success of the learning products or outcomes of their peers with similar status (Topping, 1998; Topping, Smith, Swanson, & Elliot, 2000). In contrast, peer review is defined as "the evaluation of creative work or performance by other people in the same field in order to maintain or enhance the quality of the work or performance in that field" (Linfo, 2005, p 1). In applying peer assessment to university students' writing, the forms and the types of peer assessment vary. For example, peer assessment can be implemented in face to face classrooms, or it can take place via the Internet outside the traditional classroom; it can be qualitative and/or quantitative. The functions of peer assessment are also different, which can be characterized as either formative or summative or both. Summative peer assessment focuses on the end of an event to determine whether predetermined objectives have been achieved. For example, when students have completed an academic writing assignment, in a summative peer assessment model, the instructor asks students working together to assess students' writing according to the criteria for the assignment by providing qualitative feedback and quantitative grades to decide the success of students' writing assignments. Students likely benefit from playing the role of the assessor during the

assessment process. In contrast, formative peer assessment occurs during the process of the learning activities. It focuses on maximizing students' learning by providing rich and detailed qualitative feedback information about strengths and weaknesses, and not merely a quantitative mark or grade during the learning process (Topping et al., 2000). Students also likely benefit from being the assessee during this process.

Studies on the use of peer assessment in writing tasks among college students began about 35 years ago (Ford, 1973). Since then, many studies have demonstrated the importance of using peer assessment in student learning (Berg et al., 2006; Boud, Cohen, & Sampson, 1999; Topping, 1998) and illustrated how peer assessment practices can be applied in curricula in both formative and summative ways (Sluijsmans, Dochy & Moerkerke, 1999; Topping et al., 2000). In an overview of the literature on peer assessment for university student writing, most studies focused on evaluating students' perceptions and feelings of students regarding the process of peer assessment, and the effects of peer assessment on students' writing performance (Eisenberg, 1993; Liu, Lin, Chiu, & Yuan, 2001; Li & Steckelberg 2004; Richer, 1992; Saito & Fujita, 2004; Xiao & Lucking; 2008; Zhao, 1998). These studies provide a diverse picture. On one hand, students regard peer assessment as a useful means for improving their own learning. McIsasc and Sepe (1996), for example, pointed out that peer assessment benefits students in developing their writing skills because the major activities of writing such as editing and reviewing are very similar to the processes involved in some forms of peer assessment. Li and Steckelberg (2004) found that students exhibit a high level of satisfaction toward the peer assessment process in which they are actively involved, and Xiao and Lucking (2008) reported that over 80% of students held positive attitudes

toward peer assessment and that these students valued peer assessment as a worthwhile activity and acknowledged benefit from providing and receiving peer feedback. Similarly, Liu, Lin, Chiu, and Yuan (2001) found that nearly 70% of participants claimed that they preferred peer review for their writing assignment over traditional instructor-provided comments and that most participants viewed the feedback generated during the peer assessment process as equal to those comments submitted by instructor.

The literature also indicates that students benefit from reading peers' essays, providing feedback to peers, and obtaining critical insight from others during the review process (Liu et al., 2001). Additionally, during the peer assessment process students compare their own work with their peers in order to become more aware of their strengths and weaknesses than they would in conventional teacher evaluation situations. Robinson (1999) also affirmed that there are many potential benefits to learning because in assessing peer's work, each student must read, compare, or question ideas, suggest modifications, or even reflect how well one's own work compares with that of others. When students play the assessor's role, they review, summarize, clarify, give feedback, diagnose misconceived knowledge, identify missing knowledge, and consider deviations from the ideal (Van Lehn, Chi, Baggett, & Murray, 1995). These are all potentially cognitively and metacognitively demanding activities that can help to consolidate, reinforce, and deepen understanding of the assessor (Topping, 1998). In addition, this process of providing and receiving peer feedback can help students articulate the attributes of good and poor performance and promote their thinking and learning. Many studies have demonstrated that peers' feedback could deepen students' understanding of

learning tasks and improve their writing performance (Chaudron, 1983; Lin, Liu, & Yuan 2001; Lu & Bol 2007; Plutsky & Wilson, 2004; Richer, 1992; Xiao & Lucking, 2008).

Some studies of peer assessment in university student writing have examined the validity and reliability of student peer assessment (Chen & Warren, 1999; Cho, Schunn, & Wilson, 2006; Falchikov, 1986; Haaga, 1993; Marcoulides & Simkin, 1995; Mowl & Pain, 1995; Saito & Fujita, 2004; Stefani, 1994; Xiao & Lucking, 2008). Overall, most of the analyses of validity and reliability of peer assessment revealed moderate or relatively high validity and reliability scores for students' assessment. One exception was Chen and Warren's (1999) report of a very low index of validity.

On the other hand, in spite of the many potential benefits of students' participation in peer assessment there are some disadvantages or problems in the implementation of peer assessment practice (Brindley & Scoffield, 1998; Cheng & Warren, 1997; Falchikov, 1995b; McDowell, 1995; Mowl & Pain, 1995). For example, students found that criticizing their friends was difficult, which impacts the quality of peer assessment. The literature shows that the common problem with peer assessment is that students are easily biased or not honest in giving feedback and rating scores because of friendship, gender, race, interpersonal relationships, or personal likes or dislikes (Carson & Nelson, 1996; Zhao, 1998; Macleod, 1999; Ghorpade & Lackriz, 2001; Nilson, 2003). In face-to-face peer assessment, students frequently express anxiety in sharing their feedback for fear of being wrong or rejected by peers (Zhao, 1998). Students find it extremely challenging to give negative feedback to their classmates, especially their friends, to avoid damaging personal relationships (Schaffer, 1996; Macleod, 1999). Topping et al., (2000) found that most students considered the peer assessment process as

time-consuming, intellectually challenging, and socially uncomfortable although it was effective in improving their learning.

Furthermore, Liu (2005) found that about 63% of 1740 surveyed university students had never or had rarely been involved in peer assessment activities. This finding indicated that most students are naïve and lack assessment skills and knowledge before engaging in peer assessment. Additionally, students doubt the objectivity of peer assessment and claim to have no training in such assessment practices (Cheng & Warren, 1997; Sluijsmans, Moerkerke, Dochy, & Merrienboer, 2001). Therefore, the lack of training in peer assessment is the central issue in this study.

According to Sluijsmans et al., (2002), peer assessment consists of a set of complex skills, which include: (a) defining assessment criteria; (b) judging the performance of a peer; and (c) providing feedback for future learning. Students lacking these skills in implementation of peer assessment could lower the quality of peer assessment, which decreases the value of peer assessment in students' learning, and impacts students' attitude toward peer assessment. Therefore, students should be trained to develop their assessment skills and develop their capability of giving meaningful feedback before they engage in peer assessment (Sluijsmans et al., 2002).

Context for the Study

ECI 301, Social and Cultural Foundations of American Education course, is a compulsory course for all students in a large, East-Coast Unban University, who will be teaching in secondary education. The course is offered to students both face to face on the campus and online through multiple delivery approaches for the distance. The online peer assessment method has been implemented as one of major teaching and learning

strategies in this course from Fall semester of 2006. For the past three years, students enrolled this educational foundations course have used the Wikibooks environment to participate in the collaborative creation of a series of student-authored online textbooks. Wikibooks (www.wikibooks.org) is one form of online interactive collaborative knowledge information resources, which has potential for classroom teaching and learning (Ferris & Wilder, 2006). Richardson (2006) has defined "Wikibook" as an online textbook for a curriculum to which both teachers and students contribute; that is, under teachers' direction, students create or edit entries to books that can be used within the classroom. As part of the project, students have been required to engage in peer assessment. The aim of the peer assessment is threefold: to improve students' writing skills, to improve the overall quality of the textbook the students generate, and to teach students to provide effective feedback in preparation for their future careers as teachers. The peer assessment component consists of two parts: one formative and one summative. In preparation for the submission of their articles to the class text, students complete peer-reviews, providing formative written feedback and rating scores from which students are expected to make improvements in an online Wiki environment. After the articles are submitted to the online formal course WikiText, students are organized to rate their peers' articles using the Wiki Article Evaluation Rubric, a rubric developed by instructors specifically for this assignment; Articles that receive highest rating scores were selected to become a part of the official online class textbook which served as the summative assessment. In the present study, three groups of students were compared across semesters. Students enrolled in the fall 2007 semester received a brief logical introduction to the peer review and rating process, while students enrolled in the spring

2008 semester received more substantive structured peer-assessment skills training, and students enrolled in the fall 2008 semester received different structured peer-assessment skills training. The training, for both students in the spring and fall 2008 semesters, focused on the rationale for peer assessment, defining assessment criteria, using a rubric to judge performance, and providing quality feedback before they engaged in peer assessment. The details of training methods and the differences between the two structured training are presented in the methodology chapter.

The Problem Statement

Very few studies have investigated the effects of peer-assessment training on student writing performance, the quality of peer assessment and the student satisfaction with the implementation of peer assessment. Even though a few studies suggest that students should be trained before engaging in peer assessment, the findings from previous studies do not provide strong evidence that training overcomes the problems associated with peer assessment. Examinations of the current understanding of the effects of peer-assessment training on student writing performance, student satisfaction, and the quality of peer assessment indicate a paucity of studies on peer-assessment training and the need for a richer understanding of the effects training has on peer assessment. Therefore, the present study attempts to begin filling this void. The purpose of this study is to examine the effects of peer-assessment training on students' writing performance, the quality of students' feedback, the quality (validity and reliability) of student-generated assessment scores, and the students' satisfaction with the peer assessment method by using a quasi-experimental design with three groups of university pre-service teachers. Specifically, this study compares the differences on students' writing performance, the validity and

reliability of the student-generated assessment scores, the quality of students' provided written feedback, and the students' satisfaction with the peer assessment method among three groups of student using different training methods in peer assessment. It also examines whether a particular form of training in peer assessment results in better students' writing performance, higher levels of students' satisfaction with a peer-assessment method, and higher quality of peer assessment. The following research questions are addressed:

1. Is there a difference in student's writing performance among students receiving target-criteria-based peer-assessment training, students receiving principle-based peer-assessment training, and students receiving no peer-assessment training?

2. Will the target-criteria-based training group students receive higher writing performance scores than the principle-based training group students?

3. Is there a difference in the quality of students' written feedback among students receiving target-criteria-based peer-assessment training, principle-based peer-assessment training, or no peer-assessment training?

4. Will training in peer assessment lead to valid and reliable student-generated peer-assessment scores?

5. Is there a difference in students' satisfaction with the implementation of this peer assessment method among students receiving target-criteria-based peer-assessment training, principle-based peer-assessment training, and no peer-assessment training?

The Professional Significance of the Study

The prevalence of peer assessment in higher education has demonstrated the value of peer assessment in certain contexts. The present study attempts to enrich the research on peer assessment and seeks to better inform peer assessment practices.

First, the potential impact of peer assessment training on student academic writing achievement and on the quality of their peer assessment has not received adequate attention. In spite of many previous studies on university student peer assessment, very few studies focus on the effectiveness of peer assessment training. Thus, the study attempts to fill the void of research on the effects of peer assessment training on students' writing performance, the quality of peer assessment and students' satisfaction with the implementation of the peer assessment method.

Second, although peer assessment yields many potential benefits when employed in the classroom, many teachers and students are concerned with its validity and reliability, which may restrict its implementation and deprive students of its potential learning benefits (Falckikov & Goldfinch, 2000). Therefore, this study is designed to make contributions to the knowledge of validity and reliability of peer assessment in university students' writing. Should the research of this nature reveal that peer assessment training can lead to improved validity and reliability of peer assessment, it will increase practitioners' confidence in peer assessment as a means to improve students' learning.

Third, the findings of this study attempts to provide empirical evidence about how the amount of training time, selection of training contents, and method of training impacts the effectiveness of peer assessment. Additionally, the findings of this study will

help researchers and instructors to understand the effects of using peer assessment training and provide first-hand evidence of how peer-assessment skill training impacts student academic writing skills and students' qualitative feedback in peer-assessment. Thus, findings of this study will also benefit those who will use peer assessment in teacher education programs.

Fourth, the emergence of information technology and the rapid increase of online capacity have provided a new arena for education. As online peer assessment becomes more prevalent in the university setting, it is important to apply new technology or teaching methods to support the implementation of peer assessment. As one example of new, innovative interactive software, Wiki uses online peer assessment. Thus the findings of this study, the results of student satisfaction with the peer assessment method, will contribute to knowledge and first hand empirical data about the application of educational online interactive software. This study focused on the effectiveness of peer assessment, peer feedback or interaction in on-line instructional environment, thus it will make practical contributions to university students' academic writing instruction.

Overview of Research Design

This study involved a nonrandomized control-group design. Three intact classes of university pre-service teachers who were enrolled in a Foundations of Education course were selected by convenience sampling. The pre-service teachers in Fall semester of 2007 were assigned to the comparison group (Group A). The pre-service teachers who enrolled in Spring and Fall semesters of 2008 were assigned to the two experimental groups (Group B and Group C) respectively. Two experimental groups of pre-service teachers had structured training before engaging in peer assessment and the comparison

group of pre-service teachers had no structured training before peer assessment. The peer-assessment skill trainings were based on peer-assessment skills identified by Sluijsmans et al., (2002), and designed integrating in course teaching and learning content. However, the training methods were different between the two experimental groups. One group received the structured training, known hereafter as Principle-Based Training Group. This group (students in Spring semester of 2008) received principle-based training and follow-up exercises. In contrast, another experimental group, hereafter known as the Target-Criteria-Based Group (students in Fall semester of 2008), received target-criteria-based training and follow-up exercises, and the training was tightly tied to the target assessment criteria applied by students in the peer assessment assignment. Thus the independent variables in this study are peer review-plus-rating peer assessment method with two levels of structured training (Principle-Based Training and Target-Criteria-Based Training) vs. peer review-plus-rating peer assessment method without training. The dependent variables are student writing performance, the quality of student written feedback, the quality (validity and reliability) of student-generated assessment scores, and student satisfaction with the peer assessment method. In this study, peer review-plus-rating peer assessment method refers to one form of peer assessment applied in this study. This peer assessment method consisted of two parts: one formative and one summative. In preparation for the submission of their articles to the class text, students assessed peers' articles by providing formative written feedback and rating scores from which students were expected to make improvements, hereafter referred to as peer review. After the articles were submitted to online WikiText, students were required to rate their peers' articles. To minimize group differences, several variables were controlled: the

same instructor, the same textbooks, the same assignments, the same tests and quizzes, the same number of assessors assessing each draft product and final products, and the same online technology format were used.

The pre-service teachers' post writing performance were measured on students' final Wiki articles' scores with the use of a specific tailored rubric, which was called the Wiki Article Evaluation Rubric. The pre-service teachers' quality of written feedback was measured on written feedback scores by using the Rubric of Quality of Feedback developed for this study. The pre-service teachers' satisfaction with peer assessment was measured by a questionnaire – Student Satisfaction with Peer Assessment Questionnaire – specifically developed for this study. The validity and reliability of the pre-service teacher-generated assessment scores were assessed by correlation scores on instructors' assessment scores and pre-service teacher-generated assessment sores, and consistency of pre-service teacher-generated assessment sores respectively.

To avoid the possible "unfair peer assessment" effects caused by "friend grading" or "under grading" and to increase students' sense of responsibility and accountability in the peer assessment processes in both the formative and the summative peer assessment, students were asked to submit their assessment reports to be graded by the instructor; the student summative peer assessment scores were contribute to their part of peer assessment assignments grade. In addition, if individual students were not satisfied with the summative peers' assessment, they could appeal and require instructor to re-assess his/her writing product.

Although this study had been designed carefully, there were some limitations and threats inherent in the design. For example, selection could be considered a problem since

there was no attempt made to randomize the groups that participate in this research, which limited the findings of the study. Another potential threat, instrumentation, could influence the quality of the study. Additionally, because the three groups' subjects were drawn from different semesters and the experimental period lasted throughout three semesters respectively, the conditions might be slightly different from semester to semester. Some threats may impact the external and internal validity, which discussed in the Chapter 5. The strategies used to decrease these threats were discussed in Chapter 3 Methodology.

The Delimitations of the Study

This study is delimited to selected issues. First, the peer assessment refers to university level students' peer assessment. This peer assessment consists of two parts: the first part is peer review that serves as formative assessment in the students' preparation for the submission of their article to the class text. The second part is peer rating that serves as summative assessment after students submitted their revised article to online formal course WikiText. Second, students' peer assessment activities refer to those taking place through the Internet that are asynchronous in nature. No other environments are considered. Third, the method of peer assessment used in this study is called by the author Peer Review-Plus-Peer-Rating throughout the following chapters. Likewise, in preparation for the submission of their articles to the class text, students provide the formative written feedback and rating scores to their peers' article for the purpose of making improvement; the author will call this assessment 'Peer Review'. After the articles are submitted to online formal course WikiText, students are organized to provide summative ratings to their peers' articles; the author will call this evaluation 'Peer Rating'

throughout the following chapters. Fourth, in this study, students who provide written feedback or quantitative feedback (rating score) in peer assessment, the author will call them as 'assessors' (assessor). In contrast, students who receive written feedback or quantitative feedback (rating score), the author will call them 'assessees' (assessee). Fifth, students' major writing assignment is a part of students' contribution to the online WikiText for students use in this course, so in this study, the author calls this writing assignment as Wiki article (or article). Sixth, the subjects are pre-service teachers; in other words they are students in the teacher education program. Students of no other disciplines will be included. Finally, students' writing performance refers to the text generated for students' Wikibook articles that is a major assignment required in the class.

CHAPTER 2: REVIEW OF LITERATURE

Peer assessment as a method to assist student learning has been widely used in

many institutions for more than fifty years (Sluijsmans et al., 2002). Although students

benefit from peer assessment, there are some problems associated with peer assessment,

which hinder the practice of peer assessment in higher education settings. The essential

assumption underlying this study is that students' peer-assessment skills training might

improve students' writing performance and the quality of peer assessment. Students'

assessment skills may be the main factor for effective peer assessment, which might

influence the successful application of peer assessment in the classroom as a teaching and

learning strategy helping students refine cognitive and meta-cognitive skills and greatly

improving student learning.

In this present study, peer assessment refers to the process during which students

(pre-service teachers) in an educational foundation course assess and provide written

feedback and rating to their peers' generated academic articles for the web-based

WikiText. The aim of the peer assessment in this course is threefold: to improve students'

writing skills, to improve the overall quality of students' articles that constitute the

WikiText used as the course textbook, and improve students' assessment skills in

preparation for their future careers as teachers.

This chapter reviews the literature relevant to this study's research questions and

it is categorized into five sections: theoretical foundation for peer assessment; effects of

peer assessment of writing; reliability and validity of peer assessment; problems with

peer assessment, and peer assessment training. At the end of this chapter, a summary of

the literature and the hypotheses are presented.

Theoretical Foundation of Peer Assessment

The forms of peer assessment that have been used in higher education settings vary, and the variables involved are disparate. Theoretically, peer assessment's effects for assessors and/or assessees might be created by the careful introduction of some specific variables. For example, the impact of peer assessment could vary depending on the type of peer assessment, the organization, and types of operational contexts; these variables could also include the amount of time used on task, engagement, and practice, in combination with a varying degree of accountability and responsibility (Topping et al., 2000). As such, it is very difficult to articulate the rationale of peer assessment through a single overarching theory or model for each of the many different types of peer assessment. In this section, the author attempts to articulate the theoretical underpinnings of peer assessment and define the role that students play in the process of peer assessment. Therefore, understanding the role of the learners in the process of peer assessment requires a review of theories that provide a theoretical foundation for peer assessment. These learning theories and perspectives on the learner's role in building knowledge and skills can be viewed from perspectives of social constructivism, collaborativism, and self-regulation and feedback that are related to a learner's role both as assessor and assessee in the process of peer assessment.

Constructivism and collabortivism perspectives

Educational theorists have paid much attention to the learner's role in building knowledge and skills. From social constructivism and collaborativism perspectives, social and cultural context enables learners to participate in a learning process "by which a learner internalizes knowledge, whether 'discovered,' transmitted from others, or

'experienced in interaction' with others" (Lave & Wenger, 1991, p.47). Situated cognition theory suggests that a learner can acquire knowledge and skills through his/her practices within a cultural system or participation in communities of practice (Lemke, 1997; Lave & Wenger, 1991). Piaget (1969) argues that a learner's social knowledge is culturally specific and can be learned only from others within the same cultural group. In terms of creating an authentic social environment for learning, the learner's self-efficacy beliefs, motivations for learning, self-regulation, and ability to collaborate with others are involved (Bandura, 1997; Keller, 1987; Schrunk & Zimmerman, 1994).

In the peer assessment, social constructivism and collaborativism could theoretically account for the learner's role in building knowledge and skills. According to a social constructivist perspective, student learning requires exchanging, sharing, and negotiating, as well as occasionally drawing on the expertise of more knowledgeable individuals; student learning also involves a personal, internal process and a social aspect (Liu, Lin, Chiu, & Yuan, 2001). Vygotsky (1971) suggests that learning is not an individual, secluded activity, but rather a cognitive activity that occurs in, and is mediated by, social interaction. Thus, peer interaction is vital to the improvement of students' learning, because it allows students to construct knowledge through social sharing and interaction (Liu et al., 2001).

The view that students acquire knowledge and skills within peer assessment also involves collaborativism. Collaborativism is similar to constructivism in assuming that knowledge is constructed rather than existing as a separate entity. The central idea of collaborativism is that learning emerges through shared understandings of multiple learners, whose goals are active participation and communication (Leidner & Jarvenpaa,

1995). Collaborativism assumes that much of students' learning occurs in peer groups, and learning is knowledge sharing among learners. Based on the learning theory of collaborativism, different types of collaborative activities consist of basic ingredients of peer assessment, especially formative peer assessment. For example, during the development of assessment criteria, students become involved in discussion, negotiation, and idea sharing.

Assessors

As the assessor, the learner is involved in social relationships between him/her and others in prescribed activities. In the process of peer assessment, the main role of the assessor is to assess peers' work and give feedback. Constructivism provides a theoretical framework for students' role as assessor. When students play the assessor's role, they review, summarize, clarify, give feedback, diagnose misconceived knowledge, identify missing knowledge, and consider deviations from the ideal (Van Lehn, et al., 1995). These are all cognitively and meta-cognitively demanding activities that could help to consolidate, reinforce, and deepen the assessor's understanding (Topping, 1998). The constructivist perspective claims that people construct their own understanding and knowledge of the world by experiencing reality and reflecting on this kind of experience. According to the constructivist's perspective, people learn best as active participants in designing their own activities (Papert, 1993). Learners are constantly encouraged to make decisions about how activities help them understand what they are experiencing. Thus, appropriate student-centered-learning conditions are required for learning. Kafai and Resnick (1996) suggested that students are particularly likely to construct new ideas when they are actively engaged in making some type of external artifact that they can

reflect upon and share with others. This process, in constructivist terms, is called "learning by design." It is an approach to learning in which students learn by collaboratively engaging in design activities and reflecting appropriately on their experiences. In such learning conditions, participants learn concepts by experiencing how those concepts work; they learn applicability of concepts by applying them to the solution of real-world problems; they learn problem-solving, decision-making, and collaborative skills by engaging in activities that require them to develop those skills.

The perspective of learning by design has particular relevance to the student's role as assessor. In peer assessment, assessors can benefit from being actively involved in assessment activities. During the assessment process, assessors actively use their prior knowledge to assess a peer's work and construct new knowledge based on the interaction with other peers and peer assessment experiences.

Assessees

In contrast, as the assessee, the learner's main responsibility is to receive peer feedback, which leads the learner to be focused on important factors in his/her learning. In dealing with peer feedback, this process also involves the assessee's analyzing, comparing, exchanging, negotiating and revising activities. Topping (1998) regards peer assessment as a reflexive act, and in the context of peer assessment, learning – an enterprise traditionally achieved by teaching is now accomplished by assessing.

From the social constructivist and collaborativist points of view, scaffolding is a learning approach related to the student role as assessee. Scaffolded instruction as a teaching strategy originates from Vygotsky's sociocultural theory and his concept of the zone of proximal development (ZPD). The zone of proximal development is the space

between what learners can do by their own and the next level of learning they can attain with the help of qualified instructors (Vygtotsky, 1971). Using Vygotsky's perspective, Raymond (2000) defines scaffolded instruction as that employed when others support the learner's development by providing support structures to get to the next stage or level. An important aspect of scaffolding instruction is that the scaffolds are temporary. As the learner's abilities increase, the instructor's scaffolding is removed, step by step. Finally, the learner is able to complete the task or master the concepts independently (Chang, Chen & Sung, 2002). By providing scaffolding, the instructor or peers support learners of lower competency as they develop more sophisticated understands of the world and construct knowledge. Therefore peer assessment involves training students to ask intelligent questions at both macro and micro levels while assuming assessment tasks such as thinking, comparing, contrasting, and communicating (Graesser, Peason & Magliano, 1995; Van Lehn, et al., 1995). In dealing with feedback, the assessor should be trained to question, prompt, and scaffold instead of only providing a rationally correct answer (Chi, 1996). Through peer assessment, the assessee benefits from his/her peer scaffolded feedback and improves his/her learning.

In peer assessment, the assessee can be supported with scaffolding to facilitate their learning, but they are not always supported by advanced peers (assessors). For this reason, many studies have reported the importance of development of assessment criteria for assessors because well-developed assessment criteria can increase the quality of peer feedback (Miller, 2003; Orsmond, Merry & Reiling, 2000, 2002; Woolf, 2004). Besides, students should be trained to master necessary assessment skills before they engage in peer assessment (Sluijsmans et al., 2002). Only when an assessor provides meaningful

feedback (corrective, specific, direct, accurate, achievable, practicable and comprehensible feedback), can he/she help the assessee to substantively acquire the scaffolding necessary to advance their learning.

Self-regulated learning and feedback perspectives

In formative peer assessment, self-regulated learning and feedback could be related to the learner's role in building knowledge and skills. According to Schunk & Zimmerman (2008), "Self-regulated learning (or self-regulation) refers to the process by which learners personally activate and sustain cognitions, affects, and behaviors that are systematically oriented toward the attainment of learning goal" (p. vii). Self-regulated learning mainly refer to, in the learning process, how learners monitor their learning progress, make an evaluative judgment on their learning direction, and take an action to improve their learning; specifically, learners focus on observing their performance, comparing the performance to the criteria and the goals that have been established, and reacting and responding to the perceived differences between the criteria and gaps in their progress toward the goal. On the other hand, feedback is information about how the learner's present state of learning and performance relates to these goals and standards. Learners generate internal feedback as they monitor their engagement with learning activities and tasks, and they assess their progress toward goals. Meanwhile, self-regulated learners also actively interpret external feedback from peers in relation to their internal goals (Nicol and Macfarlane-Dick, 2006). Feedback is considered central to learning (Black & Wiliam, 1998); as Topping et al., (2000) have argued, "Formative assessment seems likely to be more helpful if it yields rich and detailed qualitative feedback information about strengths and weakness, not merely a mark or a grade"

(p.150). Clearly, feedback plays an important role in building knowledge and skills; numerous previous studies that focus on feedback's effects on students' learning achievement have demonstrated that peer feedback can help students improve their learning within peer assessment arrangements (Lin et al., 2001; Lu & Bol, 2007; Richer, 1992; Xiao & Lucking, 2008). While generating and providing feedback as part of formative assessment, learners must know what good performance is, how current performance relates to good performance, and how to act to close the gap between current and good performance (Sadler, 1989).

Formative assessment intends to help students identify their strengths and weaknesses, and guide students toward the achievement of learning goals during the learning process (Boud, 1995; Dierick & Dochy, 2001). In the process of formative peer assessment, learners have to meet the criteria for assessment that have been discussed, negotiated, and employed. They have to select goals within the planning stage of peer assessment, articulate the attributes of good and poor performance, develop a vocabulary for thinking about and discussing quality, and compare their performance to good performance; additionally, students then provide and receive feedback. Finally, learners evaluate weaknesses and strengths of their own product with respect to a given assessment task (Oldfield & MacAlpine, 1995); they must then revise their product based on reflection and peer feedback according to their beliefs and understandings regarding the task's goals (Butler & Winne, 1995). All of these activities involve self-regulated learning by focusing on learners setting goals for their learning and monitoring, regulating, and controlling their cognition, motivation, and behavior, guided and constrained by their goals (Pintrich and Zusho, 2002). Previous studies on self-regulated

learning reveal that self-regulated learners are more effective learners; they are more persistent, resourceful, confident and successful (Pintrich, 1995; Zimmerman & Schunk, 2001). Therefore, well-designed peer assessment can help learners to develop their skills of self-regulated learning and improve their learning performance.

Assessors

The role of an assessor in peer assessment is to assess the peers' products and provide useful feedback to peers. Self-regulated learning and feedback perspectives may relate to the student's role as the assessor. In providing feedback, the assessor intends to help assessees identify their strengths and weaknesses, give them suggestions, and guide them toward the achievement of learning goals. To provide useful feedback, assessors have to meet the criteria of effective feedback, and also to meet the criteria for assessment that they have discussed, negotiated, and employed during the peer assessment process. In assessing peers' products, on one hand, assessors assess their peers' products and provide feedback according to assessment criteria and learning goals. On the other hand, assessors often compare assessees' good products with their own, check whether their own learning goals are appropriate and what strengths and weaknesses that their own products have, and then take action to improve their own products. All these activities involve self-regulated learning by focusing on learners setting goals for their learning and monitoring, regulating, and controlling their cognition, motivation, and behavior, under the guidance and constraint of their goals.

Assessees

In contrast to the assessor's role, the major responsibility of the assessee is to receive peer feedback. Self-regulated learning and feedback perspectives could serve as

the theoretical frameworks related to the student as assessee. In the peer assessment process, the assessees' learning can be improved if strengths or merits of assessees' products are rewarded or positively reinforced and if assessees' weaknesses or shortcomings are negatively reinforced. The positive (or negative) reinforcement serves to stimulate the assessees to take another step in learning. In this process, assessees can be guided by an assessor's correct feedback. Many studies have demonstrated that the formative purpose of peer assessment can greatly improve student learning through feedback (Liu et al., 2001; Lu & Bol, 2007; Richer, 1992; Topping et al., 2000; Xiao & Lucking, 2008). Such assessment is intended to help students plan their own learning, identify their own strengths and weaknesses, and develop meta-cognitive and other personal and professional transferable skills (Boud, 1990; Brown & Knight, 1994). In addition, when peer feedback uses more, specific criteria, it could be more accurate and further improve students' learning (Bloxham & West, 2004; Miller, 2003).

Additionally, peer feedback can be a factor influencing affective as well as cognitive dimensions of learning. For example, positive peer feedback can support the assessee's intrinsic motivation (Csikszentmihalyi, 1978) or self-attribution (Dweck, 1975). Also, peer feedback can assist self-regulated learning by cueing self-monitoring and engaging learners in other meta-cognitive processes (Toping & Ehly, 1998).

Effects of Peer Assessment of Writing

Many previous studies on a wide range of topics have investigated the effects of peer assessment (Lin et al., 2001; Lu & Bol, 2007; Xiao & Lucking, 2008). The major effects of peer assessment fall into categories of effects. The first is the learning outcome which is focused either on cognitive or affective aspects and, second, on the learning

process. The most common learning outcomes are students' learning performance and attitude. Performance is assessed in terms of subject matter-related skills or general abilities such as writing skills. Attitude relates to students' feeling and perceptions of peer assessment interventions and is frequently assessed according to levels of motivation, self-efficacy and/or satisfaction. The effects on the learning process are frequently examined from the perspective of students' awareness of their learning process. Requisite skills include analytical and critical skills, problem solving, and intellectual flexibility (Johnson, 1999). Although research in this area of peer assessment started early in the 1970s and is now extensive, much of the relevant literature is descriptive and focuses on introducing a particular peer assessment method and summarizes use of the method experiences.

Effects on Students' Awareness of Their Own Learning Process

The value of peer assessment in helping students improving their learning awareness is reported in many studies. In a study examining the learning effect, and students' perceptions of, peer assessment on the third-year computer science majors, Liu, Lin, Chiu, and Yuan (2001) reported that participants ($n = 143$) viewed the peer assessment method as effective and reported benefiting from reading peers' essays, digesting feedback provided by peers, and obtaining critical insight from others' work during the review process. Additionally, many participants mentioned that they compared their own work with that of their peers in order to become more aware of their strengths and weaknesses than in conventional teacher evaluation situations. Similar study findings indicate that when students assess their peers' work, both formatively and summatively, there are many potential benefits to learning because in assessing others' work, each

student must read, compare, or question ideas, suggest modifications, or even reflect how well one's own work compares with others' (Robinson, 1999).

McIsasc & Sepe (1996) found that students benefited from using peer assessment in their writing because the major activities of writing such as editing and reviewing are very similar to the process of peer assessment. In addition, students had deepened their understanding of the objectives of course learning and the criteria of learning task through the activities of negotiating about performance criteria in peer assessment (Falchikov, 1995b; Orsmond, Merry, & Reiling, 1996; Orsmond, Merry, & Reiling, 2000). After using a peer assessment method in teaching the process of scientific writing to 39 undergraduate students, Guiford (2001) found that participants learned more course content knowledge through the peer review and technical skill-writing for publication than through traditional term paper approaches. Venables and Summit (2003) also reported that the overwhelming majority of student responses indicated that they learned a great deal in scientific essay writing through peer assessment (n = 63) and increased their learning. The students and staff noticed that the written peer assessment was generally more detailed than that provided by staff, and the majority of students reported that use of a peer assessment method did enhance their understanding of the material and also provided them with practice in the skill of scientific writing.

Effects on Students' Performance

Peer feedback seems an important factor in improving student writing skills and learning achievement. Richer (1992) compared the effects two kinds of feedback, peer directed and teacher based, on first year college students' writing proficiency in an experimental study with 87 participants. The study results showed that there was a

significant difference in writing proficiency in favor of the peer-feedback-only group. The finding indicated that using peer feedback provides a feasible method enabling college students to enhance their writing skills and improve their learning achievement. In their quasi-experimental study comparing three methods for teaching student writing, Plutsky and Wilson (2004) found that peer feedback helped students become proficient writers. Similar results were found from empirical studies on English as a Second Language (ESL) among college students and found that peer feedback was as effective as teacher's feedback in assisting revising and improving students' writings (Chaudron, 1983; Paulus, 1999), but Chaudron (1983) reported that peer feedback was more cost-effective than teacher feedback.

Xiao & Lucking (2008) compared the effects of two peer assessment methods on university students' academic writing performance and their satisfaction with peer assessment by using a quasi-experimental study with 232 pre-service teachers in an online assessment environment. They found that participants in the group using a rating-plus-qualitative-feedback method demonstrated greater improvement in their academic writing than those in the rating-only group.

Working with college computer science major students, Lin et al., (2001) compared the effects of two kinds of peer feedback -- specific and holistic -- on students' writing according to students' different thinking styles in a web-based online peer assessment environment by using a factorial experiment design. Fifty-six participants arranged in four groups participated in this study. The finding of the study indicate that students with high executive thinking styles who received holistic and specific peer feedback, and students with low executive thinking styles who received specific peer

feedback group significantly improved their writing. However, students with low executive thinking styles who received holistic peer feedback did not improve their writing. Additionally, high executive thinking students contributed substantially better feedback than their low executive counterparts.

However, peer assessment practiced in the classroom showed that not all students who received peer feedback outperformed those who did not receive peer feedback. For examples, Biekeland (1986) compared the effects of three kinds of feedback -- self feedback, peer feedback and teacher feedback -- on students' writing skills with 76 adult technician-students. The results showed no significant differences existed between gain scores of those in the teacher feedback group and the self-evaluation group, between those of the self-evaluation group and the peer feedback group, and between those of teacher feedback group and peer feedback group. A similar finding was derived in a study by Li and Steckelberg (2004). Li and Steckelberg compared the effects of two kinds of feedback -- peer feedback and self-feedback -- on students' writing project with 47 undergraduate students in a web-based online peer assessment environment and found no significant difference.

Effects on Students' Attitude

According to a survey, Li and Steckelberg (2004) reported that students expressed a high level of satisfaction toward computer-meditated peer assessment process in which they were actively involved. Liu et al., (2001) reported that nearly 70% of participants claimed that they preferred using peer review for their writing assignments, and most participants viewed it as effective as the instructor's, which were all positive responses to the peer assessment. Xiao and Lucking (2008) also found that participants in the group of

using a rating-plus-qualitative-feedback method exhibited higher levels of satisfaction with peer assessment than those in the rating-only group, and over 80% of all participants not only showed positive attitude toward peer assessment, but valued peer assessment as a worthwhile activity and benefited from providing and receiving peer feedback in a experimental study with 232 pre-service teachers in a online peer assessment environment. A similar result was also found by Venables and Summit (2003).

Saito and Fujita (2004) investigated the effects of college students' attitude toward peer assessment among 61 business major freshmen in Tokyo. The students' responses to the two questionnaires implied an overall acceptance of the peer assessment. Using a regression analysis, the results also showed that peer rating was not a statistically significant predictor of student attitude. In other words, peer feedback ratings do not seem to influence student attitudes toward peer assessment.

Different reactions to peer assessment were also found in the literature. For example, in a qualitative study, Topping et al., (2000) found that most students considered the peer assessment process as time consuming, intellectually challenging, and socially uncomfortable although it was effective in improving their learning. Zhao (1998) found that in face-to-face peer assessment, students frequently expressed anxiety in sharing their feedback for fear of being wrong or rejected by peers. Macleod (1999) reported that some students doing face-to-face peer assessment were caused to be dishonest in giving feedback because of interpersonal relationships, and about half students reported that the computer peer reviews helped them to be more honest when giving negative comments to their friends.

In short, the studies of the effects of peer assessment on student writing have greatly increased in the last two decades, and they are found in a wide range of subjects. Many studies show various benefits that students received in the writing learning process by using peer assessment, but few research studies were conducted on the effects of peer assessment on learning outcome (achievement). Also, many of these studies on the effects of peer assessment on student learning process were descriptive, summarizing classroom practice. Therefore, more experimental studies are needed to explore the effects of peer assessment on university students' academic writing in order to isolate the effects of this form of pedagogy.

Reliability and Validity of Peer Assessment in Writing

Although peer assessment yields many benefits when employed in the classroom, many teachers and students are concerned with its validity and reliability. Falckikov and Goldfinch (2000) point out, "Fears of teachers about the lack of reliability or validity of peer assessment may act to restrict its use and, thus, deprive many students of its learning benefits" (p. 288). The validity of peer assessments refers to how strongly the students' assessments correlate with assessments made by professionals, while reliability of peer assessments refers to the consistency of assessments made between peers or the same peers over time (Topping, 1998). The literature review shows that, except for one study (Topping, *et al.*, 2000) that used qualitative methods, almost all studies (Chen & Warren, 1999; Cho, Schunn, & Wilson 2006; Falchikov, 1986; Haaga, 1993; Marcoulides & Simkin, 1995; Mowl & Pain, 1995; Saito & Fujita, 2004; Stefani, 1994; Xiao & Lucking, 2008) on the effects of peer assessment on student academic writing used quantitative method, and only one study investigated both validity and reliability of peer assessment.

Finally, most of these quantitative studies that intended to examine the reliability of peer assessment actually seem to be studies of validity (Topping, 1998). That is, they compared peer assessments with assessments made by instructors rather than with those of other peers or the same peers over time.

Of the quantitative studies, five studies investigated validity (Falchikov, 1986; Stefani, 1994; Mowl & Pain, 1995; Chen & Warren, 1999; Saito & Fujita, 2004). The sample size in these studies was between 48 and 67, and the participants were all undergraduate students from different disciplines. Mowl and Pain (1995) and Cheng and Warren (1999) reported relative low index of validity (r = .22, in geography and r = .29, in electrical engineering) but Stefani (1994) and Saito and Fujita (2004) found high validity in their studies (r = .89 in biology and r = .72, in business). These four studies computed validity by using average means of peer generated scores against instructor's grading scores. An exception is that practice Falchikov (1986) used percentage agreement between a single peer rating and a single faculty member's rating method to determine the validity, which was inconsistent with the method that other studies used. Two studies examined reliability but employed very different metrics. Haaga (1993) reported relatively high reliability (r = .55) by using Pearson product-moment correlation between pairs of students assessing common papers in a sample size of 45 graduate students who majored in Psychology. Marcoulides and Simkin (1995) used a percentage-of-variance approach to investigate the reliability in a sample size of 60 undergraduate students majoring in computer science and found that peer reviewers seemed to be consistent evaluators.

Cho et al. (2006) investigated the validity and reliability of peer assessment of writing from instructor and student perspectives with 708 students across 16 courses over three years in a web-base peer assessment environment. Of the 16 courses from four different universities, 12 were undergraduate level and 4 were graduate level, covering disciplines that include Cognitive Psychology, Psychological Methods, Health Psychology, Cognitive Science, Education Rehabilitation Sciences, Leisure Studies, and History. The results showed the duality of quality of peer assessment. From an instructor perspective, the findings of their study indicate that validities of peer assessment were relatively high and similar ($r \approx .5$ to .7). No evidence showed that graduate student ratings were more valid than undergraduate student rating and the reliability of peer assessment was high ($r = .78$). In addition, the results suggested that three or four peer raters produced middling effective reliabilities whereas six peer reviewers produced high effective reliabilities. Similarly, in an experimental study that investigated the effects of peer assessment on university student writing performance and the quality of student-generated assessment scores with 232 participants, Xiao & Lucking (2008) found that the validity of students generated rating scores were relatively high ($r = .83$) and the reliability of students generated rating scores were moderately high (3 raters: $r = .62$ and 20 raters: $r = .75$), which indicated that twenty raters yielded higher reliability than three raters in this peer assessment study.

Topping *et al.* (2000) explored the reliability and validity of qualitative formative peer assessment in the area of academic writing with fifteen graduate students. This case study compared the frequency of students' feedback with instructors' that focused on positive, negative, and neutral feedback and checked students' and instructors' inter-rater

reliability. The results indicate that students' feedback was more valid and reliable than instructor's.

In short, the studies on validity and reliability of peer assessment of writing show that validity was relatively high in all but two studies and that most researchers used Pearson product-moment correlation to compare peer assessments with assessments made by instructors. In four studies (including one qualitative study) on reliability, the researchers used very different methods and metrics. The results showed that by using assessment criteria six peer raters assessing per student essay might produce higher reliabilities than three or four peer raters assessing. Besides, the results of these studies indicated that the methods that researchers used were not consistent, especially in calculating reliability. Therefore, the validity and reliability of peer assessment of writing still need further investigation, as do the development of effective methods and a common metric.

Problems with Peer Assessment and the Factors Influencing Quality of Peer Assessment

In spite of the many potential benefits of peer assessment, there are some disadvantages or problems in the implementation of peer assessment. The most common problems associated with peer assessment are quality of peer assessment, students' attitude toward the peer assessment, and students' understanding of the significance of peer assessment. These problems are addressed in the following sections.

Many teachers and participants (students) are concerned with the quality of the peer assessment. In the formative function of peer assessment, they worry about the usefulness of assessor's feedback, while in the summative function of peer assessment they express concern about accuracy of assessor's grade – specifically, the reliability and

validity of peer assessment. Many previous studies have been carried out since 1980 to investigate the quality of summative function of peer assessment, and most of these studies compare students' generated grade to instructors' grade by employing correlation method to decide the degree to which students' generated grades are accurate. Another line of inquiry focuses on the effectiveness of assessor's feedback for improving students' learning as part of the formative function of peer assessment.

Assessment Criteria and the Quality of Peer Assessment

Assessment criteria are the essential factors that impact the quality of peer assessment. Many previous studies indicated that students' (assessors') "over-marking" is more frequent than "under-marking" in peer assessment (Cheng & Warren, 1999; Falchikov, 1995a; Falckikov & Goldfinch, 2000; Miller, 2003; Orsmond et al., 1996). The problem of students "over-marking" is frequently related to the criteria students used in assessing their peers' products and students' leniency (Falckikov & Goldfinch, 2000; Miller, 2003), and it is also frequently related to issues about students' lack of ability to discriminate levels of performance and their reluctance to judge their peers (Falckikov, 1995a; Li, 2001; Orsmond et al., 1996; Sluijsmans et al., 2001).

Falckikov and Goldfinch (2000), after analyzing 48 quantitative experimental studies focusing on assessors' grades' accuracy comparing instructors' grades in peer assessment, found that students' generated average grades are moderately positively correlated to instructors' grades ($r = .69$); they also found that peer assessments that require grading of several individual dimensions appear to be less valid than peer assessment that requires a global judgment based on well understood criteria after compared three different categories criteria that students used in peer assessment, (overall

global judgment, global judgment plus several dimensions or criteria, and judgment for each dimension separately). They suggest assessment using many individual dimensions seems more difficult than assessment using global judgments or with few dimensions.

Miller (2003) conducted a quasi-experimental study to investigate the effect of scoring criteria specificity on peer and self-assessment by comparing two groups of students using different assessing criteria (general vs. specific), assessing students' presentation in a five-year Master of Physical Therapy course. The study results revealed that when increasing the number of criteria (from 5 criteria plus 5 point rating scale to 25 criteria plus 5 rating scale) decreased the mean scores of students' marks and increased the standard deviations of the peer and self-assessment, which improved students "over-marking" in peer and self-assessment. The results also showed that the correlation between peer and self-assessment was improved with more specific criteria and increased criteria also increased the number of critical feedback items. Miller (2003) suggested that the more specific criteria may have allowed the assessors to reflect on more aspects of the quality of the performance and, since there were more rating criteria, the peer assessors may have felt more comfortable about downgrading certain areas of performance as it would not have a large impact on the overall score. In contrast, other studies have argued that peer assessment instruments (refers to assessing criteria) should be kept as simple as possible (Lopez-Real & Chan, 1999; Oldfield & MacAlpine, 1995); however, Miller's study indicated that a more complex instrument produces better quantitative discrimination of performance, which is necessary if the instrument is to have a high degree of validity.

Therefore, if the assessing criteria are too vague or difficult to understand or if the rating scale offers too few choices for scoring, then an accurate, fair judgment can be difficult to make, possibly causing assessors to grade too leniently and rendering the assessment ineffective. Additionally, if the instruments (assessing criteria) do not allow for the discrimination of performance, they have little formative or summative value for the students being assessed.

The Number of Student Ratings and the Quality of Peer Assessment

The number of student assessors is also an important factor, which impacts on the quality of peer assessment. According to their analysis of forty-eight experiment studies about peer assessment, Falckikov and Goldfinch (2000) found that single raters performed as well as larger groups of students and when the number of assessors in a group increased beyond 20 the accuracy of rating scores starts to decrease markedly. So they suggested that 2 to 7 assessors in a group would increase the accuracy of rating scores. Kerr and Bruun (1983) found when group size increased, motivation of individual student decreased, which they called "free-rider" effect. Similar the earlier study of Latane, Williams, and Harkins (1979) found that when group size increased, individual effort decreased. They called this phenomenon "social loafing."

However, it has been argued that that the use of multiple raters tends to improve reliability by increasing the ratio of true score variance to error variance (Ferguson, 1966). Fagot (1991) has argued that multiple ratings are superior to single ones, and Magin (1993) found that when individual students were poor judges, the reliability of averaged scores were increased by increasing the number of raters. Additionally, Cho et al., (2006) investigated the validity and reliability of peer assessment of writing with 708 students

over three years in a web-based peer assessment environment, and the results show that

six peer reviewers produced higher effective reliabilities than three or four peer raters. In

their quasi experimental study, Xiao & Lucking (2008) found that twenty raters yielded

higher reliability of assessment scores than three raters' assessment scores in a quasi-

experimental study with two hundred and thirty-two students in an online assessment

environment.

The Nature of the Assessment of Task and the Validity of Peer Assessment

The nature of the assessment task is another important influence over the quality

of peer assessment. According to a meta-analysis study of Falckikov and Goldfinch

(2000), after analyzing 48 quantitative experimental studies focusing on assessors' rating

scores accuracy comparing instructors' grades in peer assessment, the results of the study

indicated that peer assessment focusing on academic products (students' essays and

examinations, etc.) and processes (oral presentation, etc) seems to correspond more

closely to faculty ratings than peer assessment in the context of professional practice

(clinical skills, teacher performance, etc). When peer assessment focused on professional

practice, the quality of assessment seemed lower.

In contrast, many previous studies did not report the lower quality of assessment

in peer assessment of professional practice (Calado, 1994; Hunter & Russ, 1995; Laster,

1994; Nilan, 1983; Ramsey, Wenrich, Carline, Tnui, Larson, & Logerfo 1996). For

example, Ramsey, et al., (1996) studied peer assessment of the professional performance

of 187 medical interns; the results showed that the process was acceptable to the subjects,

and reliability was adequate despite the use of self-chosen raters. Topping's meta-analytic

results (1998) also suggested that peer assessment of professional skills showed adequate

reliability even in high-stakes areas such as medicine. But he also noticed that outcome data are limited in peer assessment of professional skills, often representing only participant perceptions.

Quality of Peer Feedback and the Quality of Peer Assessment

Student's lower feedback quality is a major problem that impacts the quality of formative peer assessment. Assessor's lower feedback quality greatly decreases the value of peer assessment. The literature indicated that lower quality of peer feedback is mainly associated with the students' ability and skills, the formats of peer feedback and students attitudes towards peer assessment. Students' ability and skills impact on the quality of peer assessment. For example, in evaluating a 2-year program involving peer grading of essays in a microeconomics course, Kerr, Park, and Domazlicky (1995) found that students with better writing ability were better at the task of grading the essays of their peers (in better agreement with the instructor grading) than those with lower writing ability. The students with better writing ability provided better quality feedback than those with lower writing ability. Similarly, in an experimental study, Lin et al., (2001) also found that the quality of feedback provided by assessors impacted peer assessment quality. They found that the high executive thinking students provided better quality feedback than their low executive counterparts and the assessees benefit more from better quality feedback.

Likewise, the types of peer feedback also impact on the quality of peer assessment. For example, Lin et al., (2001) compared the effects of different peer feedback formats (holistic vs. specific) and different thinking styles (high executive thinking styles vs. low executive thinking styles) on students learning outcomes by using an experimental design.

The findings of this study revealed that thinking style and feedback format interactively affected student learning. Low executive students receiving specific feedback significantly outperformed those receiving holistic feedback. Similarly, high executive students receiving specific feedback did better slightly than those receiving holistic feedback, but not significantly. The authors argued that while high executive thinkers could overcome the disadvantages of holistic feedback, the low executive thinkers could not.

According to the literature, the biggest problem with peer review is that students are easily biased or not honest in giving feedback because of friendship, gender, race, interpersonal relationships, or personal preferences (Carson & Nelson, 1996; Ghorpade & Lackritz, 2001; MacLeod, 1999; Nilson, 2003; Zhao, 1998). Students' attitudes and understanding the meaning of peer assessment directly influence the quality of peer assessment. For example, Hanrahan and Isaacs (2001) found that some students showed hostility towards peer assessment in their university courses. Students also showed discomfort in evaluating other students' paper. Other studies also reported that students find it extremely difficult to give negative feedback to classmates, especially friends, because they hate to hurt others' feelings or damage personal relationships (MacLeod, 1999; Schaffer, 1996; Topping, 1998). In addition, some studies indicate that students lack confidence in doing peer assessment, which also impact on the feedback quality (Cheng & Warren, 1997; Sluijsmans et al., 2001; Toping, 1998).

Critical Peer Feedback

Previous studies have demonstrated that constructively critical feedback is more useful in helping students improve their work (Falckikov, 1995b; Lu & Bol, 2007; Zhu,

1995). Lu & Bol (2007) compared the effects of anonymous and identifiable electronic peer review on college student writing performance and the extent of critical peer feedback by using a quasi-experimental study with 92 freshmen college students for two semesters. The results (both semesters) of the study showed that there was a significant difference in students' writing performance task in favor of students participating in anonymous e-peer review, and students in anonymous e-peer review group providing significantly more critical feedback to their peers than did students participating in the identifiable e-peer review. Zhu (1995) argues that if students do not approach their peers' writing critically, they will fail to provide meaningful and useful feedback. A related benefit observed by Kerr et al. (1995) is that students who take a critical approach when reading and scoring peers' work are likely to be more critical of their own work, and thus create improved products. Unfortunately, most researchers agree that the problem with peer assessment is the lack of critical feedback in peer feedback, which greatly reduces the objectivity of peer feedback (Bhalerao & Ward, 2000; Carson & Nelson, 1996; Ghorpade & Lackritz, 2001; Kerr, et al., 1995; Liu et al., 2001; MacLeod, 1999; Mangelsdorf, 1992; Nilson, 2003; Quible, 1997; Zhao, 1998). In peer assessment, students tend to over-rate their peers and give positive feedback and always feel reluctant to grade down their peers (Falckikov, 1995b; Toping, 1998), which is called the "halo error" phenomenon by Farh, Cannella, and Bedeian (1991).

Various reasons could contribute to students' reluctance to provide critical feedback and several factors need to be considered. According to the results of previous research, the most important reason that prevents students from being critical in providing feedback to their peers could be related to interpersonal factors. Many

assessors in peer assessment are unwilling to offer negative comments for fear of damaging personal relationships, being wrong or rejected by peers because of different opinions, or hurting their peers' feelings, especially those who are their friends (Zariski, 1996; Zhao, 1998). For example, Campbell and Zhao (1996) conducted a study to investigate the feedback formats that assessors provided to their peers in a pre-service teachers' program by asking pre-service teachers to make comments on their peers' journals. The findings of the study indicated that there were very few comments in critical nature in the all pre-service teachers' comments. Most of them were superficial. That is to say, the pre-service teachers were willing to compliment rather than to challenge their peers because they might be afraid of hurting someone else's feelings. Bump (1990) and Mangelsdorf (1992) also had similar findings. Mangelsdorf's study showed that most of the time the feedback provided by their peers were not critical enough to be of very much help.

Furthermore, students' cultural background may play a key role in impacting students' critical feedback. In a qualitative study, Carson and Nelson (1996) found that in an advanced ESL composition class, Chinese students' primary goal to participate in group activities was social. In group interaction, they paid more attention to the maintenance of the relationship constituting the group and group harmony among group members. The data analysis of the study indicated that the most salient characteristic of the Chinese speakers' interactions was their reluctance to both speak and make negative comments. The authors contend that Chinese students did not want to hurt anyone's feeling, they did not want to generate conflict by disagreeing with their peers, and felt vulnerable as readers and writers. They often withheld comments or tried to soften their

critical comments by "under-specifying the writer's problems" or indirection," which did not always "have the desired effect of helping the writer recognize a problem in his or her writing" (p.16).

In short, the literature has presented the important information and different points of view on the problems and the factors that influence the quality of peer assessment. To solve the problems related to the quality of peer assessment and the quality of the peer feedback, many researchers advocate specific instructional approaches believed to increase the quality of peer feedback. First, electronic online communication can be used to avoid the possible embarrassment students may experience in face-to-face interaction (Eisenberg, 1993; Lin et al., 2001; Liu & Bol, 2007; Mabrito, 1991). Second, multiple assessors can be used to balance the uneven quality of peer feedback for any single piece of writing (Cho et al., Fagot (1991); 2006; Magin (1993); Topping, 1998; Xiao & Lucking, 2008). Third, anonymous peer review can be used to minimize opportunities for students to reward friends or otherwise game the system during peer review process (Lu & Bol, 2007; Zhao, 1998).

Accordingly, the problems related to peer assessment may be solved by using a peer assessment training approach. If students can fully understand the rationale of peer assessment and feedback, they may be more likely to change their attitudes and actively participate in the peer assessment process. Likewise, if students can acquire peer-assessment knowledge and improve their assessment skills, they will likely feel more confident. If the results of students' assessment are similar to those of their instructors', students should develop an objective perspective of peer assessment. If students recognize that critical feedback can help improve their own learning as well as improving

their peers' learning, they should provide more critical feedback without unduly worrying about inter-personal relationships.

<center>Effects of Training in Peer Assessment</center>

Many previous studies covered a wide scope of topics on college student peer assessment with implications for improving the quality of peer assessment and student learning. Some of these studies suggest that students need to be trained before they engage in peer assessment (Falckikov, 1995b; Mike & Tim, 1997; Orsmond, 1996; Topping, 1998; Xiao & Lucking, 2008). However, there are only three studies related to effectiveness of peer-assessment training on university students' academic learning and their attitudes toward peer assessment.

Definitions of Peer-Assessment Training

The literature review does not find any specific definition of peer-assessment training or any available training programs that provide instructors with explicit guidelines about how to train their pre-service teachers to gain peer-assessment skills. Such a problem was also found by Tillerma, Kessels, and Meijers, (2000). To define peer-assessment training, it is important to understand the concept of peer-assessment skills and the method used to improve these skills. The literature on peer assessment particularly focuses on the importance of negotiating about performance criteria (Falckikov, 1995a; Mehrents, Popham, & Pyan, 1998; Orsmond, 1996; Orsmond, 2000), but that is only one of the skills required for conducting reliable assessments. Sluijsmans et al. (2002) identified three main peer-assessment skills: (1) defining assessment criteria (thinking about what is required and referring to the product or process); (2) judging the performance of a peer (reflecting upon and identifying the strengths and weaknesses in a

peer's product and writing an assessment report), and (3) providing feedback for future learning (giving constructive feedback about the product of a peer). The identification of peer-assessment skills provided a theoretical foundation for peer-assessment training. Accordingly, the peer-assessment skills training can be defined as an institution-based systematic, structured formal learning, where learners spend a considerable amount of time learning the rationale of the peer assessment, different peer assessment structures and improve their assessment skills through practices either in or outside of classroom. The training can be face-to-face in a classroom or can use a distance learning method. The amount of training time depends on the degree of difficulty that students confront in using the type of peer assessment in the classroom. The training content selection depends on students' prior peer-assessment knowledge and skills. Likewise, the training time and content are dependent upon the nature of the peer feedback task.

The Effects of Peer-Assessment Training on Students' Task Performance and Attitude

Peer-assessment skill training seems an important factor in improving students' task performance and enhancing their attitudes toward peer assessment. In an experimental study investigating the effects of training on peer revision in a university-level freshman composition course with 169 participants, Zhu (1994) reported that students in the experimental group, comprised of four sections, were trained through special teacher-student conferences in which the teacher met students in small groups of three to develop and practice strategies to conduct peer revision. The results of the study showed that training students for peer revision led to more and higher quality peer feedback, enhanced students' attitudes toward peer revision, and helped students engage in peer revision tasks more actively. However, the writing quality of students who were

in the experimental group with training did not differ significantly from those in the comparison group who received no training.

In an experimental study to investigate the effects of peer assessment training on pre-service teachers' learning performance and perceptions with 93 second-year pre-service teachers, Sluijsmans et al., (2002) reported that peer assessment training had the expected positive effects on developing students' peer assessment skills and task performance. The results of this study showed that the pre-service teachers who were in the experimental group and received a four-hour training period of instruction focused on defining performance criteria, giving feedback and writing assessment reports were more likely to use the criteria and give more constructive comments than those in the comparison group which received no peer assessment training. Also, the pre-service teachers in the experimental group scored higher on structure and used fewer naïve words than those in the comparison group. In addition, the results of the study showed that pre-service teachers in the experimental group received higher scores on their products (designing creative lessons) than those in the comparison group. Finally, the findings of this study indicated that pre-service teachers in the experimental group exhibited higher levels of satisfaction with peer assessment than those in the comparison group. After training, however, when comparing students' assessment skills with instructor's skills, the researchers found that the students could not be regarded as expert assessors. So the researchers suggested that the training period should be extended for future implementations.

In another experimental study, with 93 second year pre-service teachers, Sluijsmans, Brand-Gruwel, Van Merriembore and Martens (2004) investigated the effects

of peer-assessment training (focusing on one important peer-assessment skill – 'defining criteria') on the development of the peer-assessment skill and the effects on the students' learning performance. The results revealed that pre-service teachers who received peer-assessment training in the experimental group were more capable in using the set criteria determined during the peer-assessment tasks than those in the comparison group who did not receive peer-assessment training; pre-service teachers in the experimental group showed higher positive attitudes toward peer assessment than those in the comparison group. However, the results of the study showed that there was no significant difference on pre-service teachers' task performance scores between the two groups. The authors explained that the peer-assessment training only addressed one of the important peer-assessment skills and, therefore, students made limited progress in improving their peer-assessment skills. It is possible that further training would eventually lead to an effect on the level of performance.

Summary of Literature and Hypotheses

This literature review provides the theoretical foundation of peer assessment, the effects of peer assessment on university students writing, the validity and reliability of student generated assessment scores in peer assessment, the factors affecting the quality of peer assessment, and the effects of training upon peer assessment. This literature review also shows that there are many strengths and weaknesses inherent in implementation of peer assessment in higher education settings. Stefani (1998) and Topping (1998) have suggested that peer assessment can have an extremely positive effect on a learning process; many studies show various benefits that students have received in the writing process by using peer assessment (Berg et al., 2006; Boud et al.,

1999; Davies, 2006; Richer, 1992; Stefani, 1994; Xiao and Lucking, 2008). Formative assessment "seems likely to be most helpful if it yields rich and detailed qualitative feedback information about strengths and weaknesses, not merely a mark or a grade" (Topping et al., 2000). Additionally, providing an online peer-assessment environment can avoid the possible embarrassment that students may experience in face-to-face interaction (Eisenberg, 1993; Lin et al., 2001). Specific and critical peer feedback may greatly improve students' writing skills and their learning outcomes (Lin, et al., 2001). Multiple raters assessing each student's written work with assessment criteria may produce high validity and reliability (Cho et al., 2006; Xiao & Lucking, 2008). In an anonymous peer review condition, assessors may provide more objective rating scores (Zhao, 1998) and produce more critical feedback (Lu & Bol, 2007).

However, both Sluijsmans et al., (2001) and Cheng & Warren (1997) noted that students doubted the objectivity of peer assessment in the peer assessment process and claimed to have no training in such assessment practices. According to a survey study with 1740 university students (Liu, 2005), about 63% of students responded that they had never or had rarely been involved in peer assessment activities. This finding indicated that most students are naïve and lacking assessment skills and knowledge before engaging in peer assessment. Students who are novices in assessing their peers' performance are insecure about their ability to assess and indicate that they need more guidance on the grading criteria (Cheng & Warren, 1997). Results also indicated that students lacked confidence in doing peer assessment because they lacked the peer-assessment skills and doubted the quality of peer assessment (Falckikov & Goldfinch, 2000). Traditionally, teachers are viewed as the custodians of standards because they are

thought to possess the necessary knowledge and experience to conduct reliable

assessments (Nicol & Macfarlane-Dick, 2006). Peer assessment can be considered to

incorporate complex skills for which students often need to be supported or trained

(Sluijsmans et al., 2002). Accordingly, students' peer-assessment skills play an important

role in the effective implementation of the peer assessment process and students should

be trained to master such skills before they engage in peer assessment (Sluijsmans et al.,

2002). Additionally, studies have demonstrated that peer-assessment training has positive

effects on students learning performance (Sluijsmans et al., 2002), positive effects on

students' assessment skills, and positive effects on their attitude toward peer assessment

(Sluijsmans et al., 2002; Sluijsmans et al., 2004; Zhu, 1994).

Although these studies provide valuable evidence about the effects of peer-

assessment training on students' learning, examinations of the current understanding of

the effects of peer-assessment training on university students' writing performance, the

quality of peer assessment, and the satisfaction with peer assessment indicate a paucity of

studies on many of details of peer-assessment training. In addition, there is no sound

peer-assessment skill training method used in university writing course that is supported

by empirical evidence. Therefore, the effects of peer-assessment skill training in peer

assessment on university students' writing performance, the quality of peer assessment

(students' qualitative feedback and rating scores), and students' satisfaction with peer

assessment warrants further exploration. Thus for the current study, it is hypothesized

that:

1. Students who receive target-criteria-based peer-assessment training and principle-based peer-assessment training will have higher writing scores than those who receive no structured peer-assessment training,

2. The target-criteria-based training group students will receive higher writing performance scores than the principle-based training group students,

3. Students who receive target-criteria-based peer-assessment training and principle-based peer-assessment training will have higher written feedback scores than those who receive no structured peer-assessment training,

4. Students who receive target-criteria-based peer-assessment training and principle-based peer-assessment training will have higher level of validity of their generated assessment scores than those who receive no structured peer-assessment training,

5. Students who receive target-criteria-based peer-assessment training and principle-based peer-assessment training will have higher level of reliability of their generated assessment scores than those who receive no structured peer-assessment training,

6. Students who receive target-criteria-based peer-assessment training and principle-based peer-assessment training will have higher levels of satisfaction with the assessment system than those who receive no structured peer-assessment training.

7. Students who receive target-criteria-based peer-assessment training and principle-based peer-assessment training will have higher levels of satisfaction with the peer feedback than those who receive no structured peer-assessment training.

CHAPTER 3: METHODOLOGY

This chapter describes the methodology that was employed in sampling, collecting, and analyzing data for this study. Methodological concerns include the study design, population and sample, instruments, procedure, and data analysis procedures.

Research Design

This study employed a quasi-experimental posttest only comparison group design. Three intact undergraduate student groups ($n = 473$) were conveniently sampled from students enrolled in the Foundations of Education course at a Large East-Coast, urban university. One group enrolled in Fall semester of 2007 ($n = 114$) was a comparison group with no structured peer-assessment skill training. The other two groups that students enrolled in Spring 2008 ($n = 177$) and Fall 2008 ($n = 182$) were experimental groups, which received different structured peer-assessment skill training methods.

The independent variables in this study were peer review-plus-rating with two levels of structured training (principle-based training and target-criteria-based training) vs. peer review-plus-rating without structured training. The dependent variables were the final scores of students' writing performance, scores of students' written feedback in assessing their peers' writings, the correlation value (r) of validity and reliability of student-generated rating scores, and student satisfaction scores based on the implementation of this peer assessment. These variables and their measurement instruments are depicted in Table 1.

Table 1

The Variables and Measures of the Study

Independent Variable (with three levels)	Dependent Variables	Measures
Peer Review-Plus-Rating without structured training (Group A)	Final scores on student Wiki articles	Wiki Article Evaluation Rubric
	Final scores on student written feedback	Quality of Student Written Feedback Rubric
Peer Review-Plus-Rating with Target-Criteria-Based peer-assessment skill training (Group B)	Correlation value (r) of validity of student generated rating scores	Correlation of scores from students' evaluation and instructor's grading
Peer Review-Plus-Rating with Principle-Based peer-assessment skill training (Group C)	Correlation value (r) of reliability of student generated rating scores	Correlation of scores from students' evaluation
	Scores of student satisfaction with the peer assessment	Peer Assessment Satisfaction Questionnaire

To address the effects of peer-assessment skill training in peer assessment on university student writing performance and student peer assessment quality in an online environment, this study examined whether the structured peer-assessment skill training would result in better scores in evaluation of student writing performance, better scores in evaluation of the quality of student written feedback in assessing their peers' writings, and greater validity and reliability in student generated peer-assessment scores. This study also examined whether structured training in peer assessment would result in differences in student satisfaction with implementation of this peer assessment. Finally, this study examined whether the target-criteria-based training in peer assessment would result in better scores in the evaluation of student writing performance over principle-

based training in peer assessment. The questions concerned the effects of structured peer-assessment skill training were as follows:

1. Is there a difference in students' writing performance among students receiving target-criteria-based peer-assessment training, students receiving principle-based peer-assessment training, and students receiving no peer-assessment training?

2. Will the target-criteria-based training group students receive higher writing performance scores than the principle-based training group students?

3. Is there a difference in the quality of students' written feedback among students receiving target-criteria-based peer-assessment training, principle-based peer-assessment training, or no peer-assessment training?

4. Will training in peer assessment lead to valid and reliable student-generated peer-assessment scores?

5. Is there a difference in students' satisfaction with the implementation of this peer assessment method among students receiving target-criteria-based peer-assessment training, principle-based peer-assessment training, and no peer-assessment training?

Subjects

To address the research questions in this quasi-experimental study, three educational foundations undergraduate classes were selected at a large East-Coast, urban university. The target population for this study was undergraduate students enrolled in a teacher education foundations course during the fall semester of 2007, the spring semester of 2008, and the fall semester of 2008. Four hundred and seventy-three sophomore and junior students were selected through convenience sampling in this case.

A total of 114 students (59 online students) enrolled in the course in Fall 2007 were

assigned to the comparison group (Group A), while 177 students (87 online students)

who enrolled in the course in Spring 2008 were assigned to the experimental group

(Group B); and182 students (96 online students) in the course in Fall 2008 were assigned

to the experimental group (Group C). The majority of the subjects were Caucasian

(70.6%) and female (81.2%). Table 2 presents the students' gender and ethnicity

information in detail. Students' ages ranged from 19 to 61, accounting in large part for

the large standard deviations in this demographic category obvious in Table 3. Three

group students' GPA and age averages were checked when the students entered the study.

A one-way ANOVA was used to test difference between means of students' GPAs and

the results showed that there was no significant difference among the three groups, F

$(2,420) = .32, p > .05$. A one-way ANOVA was also used to assess difference between

means of students' age and the results showed that there was no significant difference

between the three groups, $F(2, 470) = .003, p > .05$. The students' final course grades

were checked after the course completed, and final course grade averages of the three

groups were similar (see Table 3).

Because the sample included a large number of online students, an Independent-

Samples t test was conducted to assess students' GPAs between on campus students and

online students by groups (Group A, Group B and Group C). The results showed there

was no significant difference between on campus students' and online students' average

GPAs in each group [$t (100.51) = 4.43, p > .05$ (Group A), $t (153.88) = 4.43, p > .05$

(Group B), and $t (150.36) = 1.84, p > .05$ (Group C)]. Table 5 shows the descriptive

statistics of students' age, GPA and final total course grade averages between campus

students and online students in each group. The descriptive statistics showed that online students' average age in Groups A and B was higher than that of students enrolled in on campus sections. It also showed there were similar final course grade averages between online students and on campus students in the three groups. However, there was no significant difference in the GPA averages between online students and on campus students in each group (see Table 5).

Additionally, the conditions might differ slightly among subjects drawn from different semesters during the process of the study. Some extraneous variables could not be completely controlled, such as unexpected events that occurred from semester to semester. These factors might weaken the design validity. To minimize these threats, various controls were in place. Across the three semesters (Fall 2007, Spring 2008, and Fall 2008), all students had (a) the same instructor teaching class; (b) the same course learning materials (students' authored online Text) and the same syllabus; (c) the same assignments and tests; (d) the same peer assessment rubrics and instruments; and (e) the same online environment for implementation of peer assessment.

Table 2

Students' Gender and Ethnicity Information by Group

		Group A		Group B		Group C	
Variables		N	%	N	%	N	%
	Female	93	81.6%	146	82.5%	145	79.7%
Gender	Male	21	18.4%	31	17.5%	37	20.3%
	White	86	75.4%	118	66.7%	130	71.4%
Ethnicity	Black	14	12.3%	36	20.3%	27	14.8%
	Hispanic	5	4.4%	9	5.1%	10	5.5%
	Other	5	4.4%	8	4.5%	9	4.9%
	Missing	5	4.4%	6	3.4%	6	3.3%
	Total N	114	100%	177	100%	182	100%

Table 3

Students' Age, GPA and Course Final Grade by Group

	Group A			Group B			Group C		
Variables	N	M	SD	N	M	SD	N	M	SD
Age	114	25.46	7.38	177	25.53	7.72	182	25.49	8.72
GPA*	108	3.01	.55	156	2.97	.61	157	3.02	.71
Course Grade	114	83.18	25.87	177	81.55	22.50	182	82.18	23.17

*6 students in Group A did not provide their GPA. 21 students in Group B did not provide their GPA, and 24 students in Group C did not provide their GPA.

Table 4

Descriptive Statistics of Students' Age, GPA and Course Final Grade by Group

	Group A					
Variables	Campus			Online		
	N	M	SD	N	M	SD
Age	54	24.23	8.12	54	26.72	8.64
GPA	53	3.12	.51	55	2.95	.40
Course Grade	55	82.72	24.13	59	83.64	27.65

	Group B					
Variables	Campus			Online		
	N	M	SD	N	M	SD
Age	90	22.02	4.24	87	29.16	8.77
GPA	88	2.91	.68	68	3.04	.51
Course Grade	90	81.25	22.32	87	81.85	23.10

	Group C					
Variables	Campus			Online		
	N	M	SD	N	M	SD
Age	86	25.53	8.69	96	25.48	8.91
GPA	83	3.12	.79	75	2.92	.59
Course Grade	86	81.52	23.46	96	82.84	22.85

Instrumentation

Students' Writing Performance

Students' writing performance in producing an online textbook article (Wiki article) was measured using the Wiki Article Evaluation Rubric. The rubric was used by both instructors and students to evaluate student articles. This rubric was created by course instructors after considerable refinement over a number of semesters. Students can receive a score ranging from 1 to 5 for each criterion, with the maximum total score being 20 points. The four criteria are: (a) information importance, or the degree to which information presented in the article is important and relevant to pre-service teachers' professional knowledge; (b) interest of information or the degree to which the information presented in the article engages readers; (c) credibility of sources or the degree to which information used in the article is cited from reliable sources, and the degree to which the article excludes unjustified personal opinion; and (d) effectiveness of writing or the degree to which the article demonstrates rhetorical fluency and clarity of expression.

Content validity has been addressed through consultation among various professors at the research university specializing in educational foundations and other educational disciplines during the two years prior to the current study. This consultation resulted in a series of substantive changes to the assessment rubric – specifically, the addition of specific (rather than holistic) rating criteria. Reliability was checked using Cronbach's Alpha. The internal consistency (alpha coefficient) for this instrument is .84 derived from 231 university students' article evaluations in Fall 2007. In addition, the author checked the internal consistency reliability of this instrument ($\alpha = .87$) with Spring

2008 and Fall 2008 student article evaluations (a total 357 evaluations). The author also checked (the inter-score reliability) the inter-rater reliability between instructor and evaluator, which shows a correlation (r) of .93. The distribution of points in each scale of the criteria is presented in the Wiki Article Evaluation Rubric in Appendix A.

Quality of Student Written Feedback

Students' written feedback skills were measured by using the Rubric of Quality of Feedback, which was created by a course instructor and the author. The rubric consists of five criteria, and each criterion has three rating scales. Students can receive a score ranging from 1 to 6 for each criterion with the maximum total score of 30 points. The five criteria are as follow: (a) feedback includes all components and contains at least two compliments of the article strengths, at least two suggestions for the article improvement, and at least 150 words of overall feedback; (b) feedback reflects standards or criteria, which relates student's current performance to these standards; (c) feedback is understandable, in that it is expressed clearly and directly; (d) feedback is specific, in that it includes specific examples of positive and negative qualities; and (e) feedback is achievable, in that it provides logical suggestions for improvement. The distribution of points in each scale of criteria is presented in the Quality of Student Feedback in Appendix B.

The criteria and scales of this instrument were developed based on the literature of constructive feedback (Nicol & Macfarlane-dick, 2006; Sadler, 1989; Sluijsmans et al., 2002). Content validity was addressed through expert review. This expert review resulted in a series of substantive changes. The reliability of this instrument, the internal

consistency reliability (alpha coefficient) was .86 calculated from the grading scores of the author over the three semesters. In addition, the inter-rater reliability between instructor's and the author's grading was checked, which shows a correlation (r) of .92.

Students' Satisfaction with Peer Assessment Method

Students' satisfaction with the peer assessment method was measured by the Peer Assessment Satisfaction Questionnaire – to compare the extent of student satisfaction among the groups. This questionnaire includes two parts. The first part consists of a series of demographic items and four multiple choice items to gather student demographic information.

The second part of the questionnaire includes thirty-one close-ended questions, which are divided into two scales: Satisfaction with the Assessment System (15 items) and Satisfaction with Peer Feedback (16 items). A four-point Likert scale was used in the questionnaire; ranging from (1) "strongly disagree" to (4) "strongly agree". Students could receive a score from 1 to 4 for each item, and the maximum total score will be 124 points for 31 items.

Students' satisfaction with the assessment system consists of three subscales, which are satisfaction with peer feedback/rating system, satisfaction with Wiki technology system used in peer assessment, and satisfaction with the convenience of doing peer assessment. There are five items in each subscale. The questionnaire blueprint of students' satisfaction with peer assessment system is presented below in Table 5.

Table 5

Students' Satisfaction with Peer Assessment System Blueprint

Subscale	Close-ended Items (four-point Likert scale)	Total Items
Satisfaction with peer feedback/rating system	#1, #2, #3, #5, #6	5
Satisfaction with Wiki technology system used in peer assessment	#10, #14, #15, #12, #13	5
Satisfaction with the convenience in doing peer assessment	#4, #7, #8, #9, #11	5
Total		15

Students' satisfaction with peer feedback also includes three subscales, which are satisfaction with the quality of peer feedback, satisfaction with benefits from providing and receiving feedback, and satisfaction with the process of providing and receiving feedback. The questionnaire blueprint of students' satisfaction with peer feedback is presented below in Table 6.

Table 6

Students' Satisfaction with Peer Feedback Blueprint

Subscale	Close-ended Items (four-point Likert scale)	Total Items
Satisfaction with the quality of peer feedback	#22, #23, #24, #26	4
Satisfaction with benefits from providing and receiving feedback	#20, #21, #25, #27, #28, #29, #30	7
Satisfaction with the process of providing and receiving feedback	#16, #17, #31, #18, #19	5
Total		16

The questionnaire was developed by the author and based on the literature to reflect the value that participants placed on their peer assessment experiences in the course. Ten items of were adapted from the Student Learning Satisfaction Questionnaire (Lu, 2005), which reported an average internal consistency (alpha coefficient) .89 (α = .89) derived from 110 student tests.

Content validity of the Peer Assessment Satisfaction Questionnaire was addressed through expert consultation. Reliability was checked using Cronbach's Alpha. The internal consistency (alpha coefficient) for satisfaction with the peer assessment system and satisfaction with peer feedback is .72 and .80 respectively from 201 undergraduate university students (Xiao & Lucking, 2008).

Additionally, the author calculated the internal consistency reliability of this instrument again with Spring 2008 and Fall 2008 student responses to the questionnaire (a total of 357 students). The alpha coefficient of the instrument shows that students'

satisfaction with the assessment system is .74 and satisfaction with the peer feedback

is .85. Table 7 shows the items corresponding to each of the subscales. The detailed item

information about the Peer Assessment Satisfaction Questionnaire is presented in

Appendix C.

Table 7

Coefficient Alphas and Student Satisfaction with Peer Assessment Scale

Scale	Items Comprising the Scale	α
Satisfaction with the Assessment System	15 (#1 – #15)	.74
Satisfaction with the Peer Feedback	16 (#16 – #31)	.85

Validity of Student-generated Assessment Scores

In this study, the validity of student-generated assessment is defined as validation

of student-generated assessment scores from instructor' perspective, that is validating

students' ratings against those of teachers or evaluators as a standard. In order to examine

the validity of student generated assessment scores, the author used a Pearson Product-

moment Correlation to determine if a relationship between student rating (assessment)

scores and those rating scores assigned by instructors. Two sets of rating scores (the

average scores of student summative peer assessment and instructor grading scores) were

compared to examine how closely student assessments mirror those of the instructors;

this method is commonly used for determining systematic relationships between groups

of assessors (Miller, 2003; Cho et al., 2006; Xiao & Lucking, 2008). Several previous

studies have used this method to detect the validity of student-generated assessment

scores in peer assessment (Chen & Warren, 1999; Cho et al., 2006; Mowl & Pain, 1995;

Saito & Fujita, 2004, Stefani, 1994; Xiao & Lucking, 2008).

Reliability of Student Generated Assessment Scores

Reliability of student-generated assessment scores, in current study, is defined as

the consistency of student-generated assessment scores applying a scoring rubric to

writing evaluation. In order to examine the reliability of students' rating scores, students'

first and second round rating scores were analyzed by using intraclass correlations (ICC),

a common measure of reliability of either different judges or different items on a scale

(Cho et al., 2006; Shrout & Flesiss, 1979; Xiao & Lucking, 2008). According to Cho et

al., (2006), the ICC (C, k) Case 2 formula would effectively estimate the consistency of k

reviewers combined (a case of random reviewers, and random articles). Therefore, using

(C, k) Case 2 formula is appropriate to determine the reliability of student rating scores in

this setting.

Peer-Assessment Training

Prior to the peer assessment activities, students in the comparison group (Group A)

had an 80-minute brief logical peer assessment introduction and practices in course

support section. However, the students in the experimental groups received two forms of

structured peer-assessment skill training. Students in the Group B (students in the Spring

semester of 2008) received principle-based peer-assessment skill training (principle-

based training) that spanned two weeks of instruction and follow-up experiences, while

students in the experimental Group C (students in the Fall semester of 2008) received

target-criteria-based peer-assessment skill training (target-criteria-based training) that

also spanned two weeks and follow-up exercises. The training for both groups was based

on three main assessment skills identified by Sluijsmans & Van Merrienboer (2000).

These are defining assessment criteria, judging the performance of a peer, and providing

feedback for future learning.

Principle-based Peer-assessment Training

There were four peer-assessment tasks that were designed for the Group B

students during training. These tasks were integrated into the course content and the

pedagogy domain, but they were closely related to the student major assignment

concerning writing an article for contribution to an online textbook. The training focused

on (a) learning the rationale of peer-assessment; (b) defining assessment criteria; (c)

providing effective feedback for learning and (d) judging the performance of a peer.

In Task 1, students were introduced to the rationale of peer assessment, the type

of peer assessment that they were going to use in the course, and the product that they

were going to assess after peer-assessment training. After this introduction, the instructor

provided some examples of articles that were written by former students and asked

students to discuss the criteria or standards for a good article.

In Task 2, the skill "defining criteria" was addressed. Examples of valid and

invalid criteria were presented. The instructor asked students to do an exercise in the

classroom individually. Students had to create a rough draft of criteria or standards for an

article that students were going to write. Then the instructor randomly selected

individuals to present their draft of the criteria and discuss it in the classroom. After

selected students presented their criteria for what determines a good article, the instructor

presented the Wiki Article Evaluation Rubric that would be used by the students and the

instructor to evaluate student articles, explaining the criterion and rating scales in detail

and providing examples.

In Task 3, the skill "provide effective feedback for learning" was discussed. The instructor first asked students about (a) their ideas concerning feedback and criticism and (b) their opinions of criteria for good feedback. After a short discussion, the instructor presented to the students expert assessment feedback for a former student article and explained the Rubric of Quality of Feedback, which would be used for peer assessment in the course. Before the end of Task 3, all students received a sample article written by a former student and were asked to provide feedback according to the Rubric of Quality of Feedback in the class. After students finished providing feedback, the instructor asked some students to present their feedback in the class and questioned particular exemplars of the feedback. The instructor also explained why certain feedback was more appropriate than others, and worked to help students understand the criteria of feedback.

In Task 4, the students were trained to "judge the performance of a peer." In this final task, the three prior tasks were integrated. An expert assessment report for a former student's article was presented to the students. This time the instructor focused on explaining the form of assessment report and the language used in feedback. Then students discussed these issues. After discussion, students were given twenty minutes to assess the former student article and write an assessment report including feedback and rating. After students completed this assignment, the instructor questioned some students as to why a particular score and written feedback was given, explained why certain scores were more appropriate than others, and helped students to understand the criteria and the certain aspects of written expression.

Target-criteria-based Training

Target-criteria-based training consisted of the same four peer-assessment tasks as for group B. However, for Group C students, these tasks were not only integrated into course content and the pedagogy domain but also more closely related to the students' major assignment – writing a Wiki article. Although the training contents and tasks were similar to principle-based training, the target-criteria-based training focused on more specific skills and their mastery. For example, in task 2, "defining criteria," the instructor focused training on the defining each criterion for the Wiki article (the students' major writing assignment in the course); determining the credibility of Wiki article sources (one of criterion of the Wiki Article Evaluation Rubric), and evaluating an article's grammar and mechanics. Another example is that in Task 4, "judge the performance of a peer," instructors required all students to write a specific explanation of their ratings of a Wiki article. Additionally, after each task training, students were required to complete a formal assessment training assignment; this consisted of (a) assessing three articles written by former students (without specific assessment criteria or rubric); (b) creating criteria for students' Wiki articles that they would write and use during this course; (c) using a rubric to evaluate one article that was written by a former student and writing an assessment report to submit; and (d) rating three articles that were written by the former students and writing a report explaining each rating.

The major differences between the principle-based training and the target-criteria-based training are as follows: first, students in the target-criteria-based training group were required to do peer-assessment skill assignment exercises, the results of which were evaluated by the instructor after each of the four training tasks; each task focused on

specific criteria of student Wiki article writing assignment. In contrast, students in the principle-based training group were not required to complete any formal peer-assessment skill assignment exercises outside the classroom. The target-criteria-based training closely focused on the requirements (criteria) of writing a Wiki article, such as assessing information and credibility of an article; it also addressed the requirements for multiple-choice question. (Writing four multiple-choice questions along with writing a Wiki article was one of the requirements of writing the article). In contrast, the principle-based training focused more on general peer-assessment skills.

Procedure

Peer Assessment

The use of a Wiki context had been institutionalized at this university over a period of three years. During this time, students in this foundations of education course had been participating in a collaborative student-authored online text project, compiled using the Wikibooks environment. A major student assignment was to write an article for inclusion in an online textbook for the course; this online source is Wikibook. According to a Web definition, "Wikibooks, previously called Wikimedia Free Textbook Project and Wikimedia-Textbooks, was a Wikimedia Foundation wiki for the creation of free content textbooks and manuals" ("Wikibooks," 2008, Section 1). Specifically, each student in this foundation course was required to write a 1000-word article on a given course topic (a total of four students are allowed to write about a single topic) and these articles became resource materials used throughout the duration of the course. In the process of writing the assigned article, students were required to participate in two rounds of peer assessment exercises. The article assignment was worth 30% of the total course

grade, while the peer assessment assignment was worth 5% of the total course grade.

The aim of the peer assessment in this course was threefold: to improve student writing skills, to improve the overall quality of a students' articles, and to improve students' assessment skills such as providing effective feedback in preparation for their future careers as teachers. The peer assessment component consisted of both a formative and summative component. In preparation for the submission of their articles to the class, students completed peer-reviews and peer ratings, providing formative feedback from which other students were expected to make improvements. After the articles were submitted to the online Wikibook, students rated one another using the Rubric of Student Academic Paper; this served as the summative assessment.

Peer Assessment Procedure

Preparation stage

Prior to the peer assessment, the students in the experimental groups had two weeks of structured peer-assessment skills training and had practiced these skills during entire peer assessment process. Students in experimental Principle-Based Training Group (Group B) received two weeks of principle-based training and follow-up experiences. In contrast, students in experimental Target-Criteria-Based Group (Group C) received two weeks of target-criteria-based training and follow-up exercises before they engaged in peer assessment. Both groups' training tasks were integrated in course content and the pedagogy domain, but were closely related to the student major assignment concerning the writing of an article for contribution to an online textbook. Students in the comparison group (Group A), however, had a brief logical introduction to the peer review and rating process and some opportunities for practice in an 80-minute course support

section in week three of the fall 2007 semester. Activities included the introduction to the peer assessment method, explanation of detailed criterion, rating scales about both written feedback and rating, and the requirements of the process of peer assessment.

Formative peer assessment stage

In the first round peer assessment, all student groups employed a peer review-plus-rating peer assessment method to assess peers' articles after students completed their first drafts and posted them to the class's online Wikibook. Four students from both the comparison and experimental groups were randomly assigned to a peer assessment group to assess three peer articles. In the assessment process, students provided written feedback to peer articles according The Rubric of Quality of Feedback on a "discussion page" provided within the Wikibook. Students then gave a rating to their peers' articles according to the Rubric of Student Academic Paper in an online database. After students completed peer assessment activities by providing written feedback and rating their peers' articles, instructors provided each student with a detailed report of each student's article's peer ratings to date, including the rating scores for each of the four rating categories (the degree to which reviewers regard the article as important, the degree to which reviewers are interested in article contents, the credibility of the article, and the clarity/fluency of writing). After the feedback and rating reports were received, all students were required to revise their articles. Specifically, students revised their articles according to peer feedback and their own self-assessments. This process lasted for three weeks.

Summative peer assessment stage

When both the experimental and comparison groups had completed posting their finalized articles to the online Wikibook, the second round of the peer assessment process

began. In this stage of the process, all students used the Rubric of Student Academic

Paper to give a summative rating to peers' articles; students were divided into four groups

at this phase, and each group was asked to rate a different set of peer-authored articles

from among those posted to the Wikibook. Through this process, each student's article

received at least ten ratings (often more than 20). At the end of the second round of peer

assessment, the Student Satisfaction with Peer Assessment Questionnaire was

administered using Blackboard—an online teaching and learning management platform,

to investigate the levels of students' satisfaction with the peer assessment methods that

were employed. The peer assessment process was carried out using Wiki technology.

Ratings were conducted anonymously. However, qualitative feedback provided to articles

were somewhat identifiable based on the Wikibook username (avatar) of the student

providing the feedback. This process lasted two weeks. The specific peer assessment

procedure is presented below in Table 8.

Table 8

Peer Assessment Sequence of Events

Week	Students' activities	Instructor activities
Week 3	• Students in three groups prepare their article • Students in Group B and C prepare peer assessment: receive structured peer-assessment skill training – task 1 and task 2 • Students in Group C complete 2 peer-assessment skill training assignments	• Lecture • Provides structured training • Grades target-criteria-based group students assignments • Provides support for the students who need help
Week 4	• Three groups' students prepare	• Lecture

	• their article • Three groups' students prepare peer assessment: experimental group students receive structured peer-assessment skill training – task 3 and task 4 • Group C students complete 2 peer-assessment skill training assignments • Group A students receive a logical peer-assessment introduction	• provide structured training • Gives a logical peer-assessment introduction • Grades target-criteria-based group student assignments • Provides support for the students who need help
Week 5	• Three groups' students post their drafts to the class's online Wikibook (before Monday 11:00 am) • Three groups' students participate first round peer assessment – each student assesses three peer articles, provides written feedback and gives rating scores for his/her peers' articles.	• Gives lecture • Monitors student peer assessment • Provides support for the students who need help
Week 6	• Three groups' students analyze peer feedback and do self-assessment. • Three groups' students revise their articles.	• Provides each student with a detailed report of their article's peer ratings. (before Monday 11:00 am) • Gives lecture • Monitors student peer assessment • Provides support for the students who need help
Week 7	• Three groups' students analyze peer feedback and do self-assessment. • Three groups' students revise their articles • Three groups' students submit their assessment report (for first round peer assessment before	• Gives lecture • Monitors student peer assessment • Provides support for the students who need help • Grades student assessment report

	Sunday 11:59 pm).	
Week 8	• Students submit their finalized articles and post their articles on the online Wikibook (before Monday 11: 00 am) • Three groups' students rate their peer articles	• Gives lecture • Monitors student peer assessment • Provides support for the students who need help
Week 9	• Three groups' students rate peer articles • Three groups' students complete a survey -- the Peer Assessment Satisfaction Questionnaire (before Sunday)	• Gives lecture • Monitors student peer assessment • Administers the survey – Peer Assessment Satisfaction Questionnaire • Grades students' articles and qualitative feedback

(Group A = Comparison Group, Group B = Principle based Training Group, Group C = Target-criteria-based Training Group)

This study employed a web-based, online approach to facilitate peer assessment for each of the following reasons: first, peer assessment is a major part of the collaborative student authored online Wikibook project in this course, and Wiki interactive software is integrated in the course design as a web-based online approach (O'Shea, Baker, Allen, Curry-Corcoran, & Allen, 2007); second, using such web-based online peer assessment allowed online students to carry out peer assessment at different distance sites; third, by using web-based online peer assessment, students could ensure their own anonymity and facilitate a willingness to be critiqued by evaluating peer work through the web, and finally instructors could monitor student progress throughout the assessment process(Lin et al., 2001; Li & Steckelberg, 2004). Thus, this peer assessment

is both formative and summative, both anonymous and identified, and asynchronous in nature.

<center>*Data Collection and Analyses*</center>

Data Collection

The quantitative measures for this study included (a) students' summative rating scores on students' Wiki articles; (b) instructors' grades on student Wiki articles; (c) researcher grades on student Wiki articles; (d) instructor grades on students' qualitative feedback and (e) students' responses on the Peer Assessment Satisfaction Questionnaire. The data collection procedure is presented in the following paragraphs.

Students' writing scores

Students' final writing scores were collected from class Wiki article grading sheets by the instructor and evaluators. An individual student Wiki article grade was scored from 0 point to 20 total points according to the Wiki Article Evaluation Rubric with four criteria and five scale points.

Students' rating scores

Students' Wiki article rating scores were retrieved from in a database connected each article of student authored online class text (Wikibook) after each round peer assessment (two rounds peer assessment). Likewise, an individual student Wiki article rating was scored from 0 point to 20 total points according to Wiki article Evaluation Rubric with four criteria and five scale points.

Students' written feedback scores

Students' written feedback scores were collected from written feedback grading sheets scored by instructors and the evaluator. An individual student feedback grade was

scored from 0 to 30 total points based on the Quality of Feedback Rubric with five criteria and three scale points.

Students' satisfaction with peer assessment scores

Students' responses to the Peer Assessment Satisfaction Questionnaire were downloaded from Old Dominion University Blackboard database, an online teaching and learning management software. An individual student response to the questionnaire was scored from 31 points to 124 points based on the 4-point Likert scale used in the questionnaire. Responses to the 4-point Likert scale, (strongly disagree, disagree, agree and strongly agree) were coded as 1, 2, 3 and 4 points respectively.

Data Analyses

This study utilized quantitative data analyses techniques. The analyses that were used to address each of the research questions are described below:

Research questions #1 and #2

1. Is there a difference in student's writing performance among students receiving target-criteria-based peer-assessment training, students receiving principle-based peer-assessment training, and students receiving no peer-assessment training?

2. Will target-criteria-based training group students receive higher writing performance scores than principle-based training group students?

In order to answer the two questions concerning the students' writing performance a one-way analysis of variance (ANOVA) was performed to assess differences in mean scores of the final major writing assignment (Wiki article) among the three groups.

Research question #3

3. Is there a difference in the quality of students' written feedback among students receiving target-criteria-based peer-assessment training, principle-based peer-assessment training, or no peer-assessment training?

In order to answer the second research question on the quality of student qualitative feedback, likewise, a one-way analysis of variance (ANOVA) was applied to assess differences in mean scores of the quality of students' written feedback among the three groups. Follow-up tests- post hoc was also used to compare the differences among groups.

Research question #4

4. Will training in peer assessment lead to valid and reliable student-generated peer-assessment scores?

In order to answer the third research question the correlation analyses was performed to assess the differences of the validity and reliability of student-generated rating scores on student Wiki articles among the three groups. In order to assess the validity of student-generated rating scores on student Wiki articles a Pearson product-moment correlation was computed. This was done by comparing student rating scores against instructor scores to determine how well the two set scores correlate. After each group's correlation was calculated, a Fisher r-to-z transformation technique was used to test significance of the difference between two correlation coefficients to determine whether students who had peer-assessment skill training had more valid student-generated assessment scores than those students who had no peer assessment skill training.

In order to assess the reliability of student-generated rating scores on students'

wiki articles an intraclass correlation was used to assess the internal consistence

reliability of student generated assessment scores. Likewise, after each group's

correlation was calculated, a Fisher r-to-z transformation technique was performed to test

significance of the difference between two correlation coefficients to determine whether

students' who had peer assessment skill training had more reliable student-generated

assessment scores than those students who had no peer assessment skill training.

Research question #5

5. Is there a difference in students' satisfaction with the implementation of this peer

   assessment method among students receiving target-criteria-based peer-

   assessment training, principle-based peer-assessment training, and no peer-

   assessment training?

In order to answer the final research question concerning students' satisfaction

with peer assessment a multivariate analysis of variance (MANOVA) was conducted to

assess the differences mean scores on the student satisfaction with peer assessment

conducted in this course. The independent variable in this analysis was Peer Review-

Plus-Rating with structured training, which has three levels, principle-based training,

target-criteria-based training and without training. The dependent variables were the

scores of student satisfaction with peer assessment system and students' satisfaction with

peer feedback.

A MANOVA was utilized in the study instead of univariate analysis for the

different dependent variables in an attempt to avoid Type I errors. Were significant

differences found, univariate (ANOVAs) analysis would be conducted as a follow up to

separately examine the effects of the independent variable on the two dependent

variables.

CHAPTER 4: RESULTS OF THE STUDY

The purpose of this study was to investigate the effects of training in peer assessment on university students' writing performance, peer assessment quality, and students' satisfaction with peer assessment in an online environment. Data were collected from three classes of students enrolled in Fall 2007, Spring 2008 and Fall 2008 semesters respectively at a Large Ease-Coast Urban University. This chapter presents the results of the study by research question and hypothesis.

Impact of Peer Assessment Training on Students' Writing Performance

To address the first and the second research questions regarding students' writing performance, a one-way analysis of variance (ANOVA) was performed to assess difference in mean scores of students' final major writing assignment based on instructor and evaluator's assessment of the students' final articles. The difference between groups was significant, $F(2, 470) = 7.54, p = .001$. However, the effect size was small ($\eta^2 = .031$).

Follow-up tests were conducted to evaluate pairwise differences between the three groups' mean scores. The results showed that there was a significant difference between the means of group A (that did not receive structured peer-assessment skill training) and Group B (that received a principle-based peer-assessment skill training). The results also showed that there was a significant difference between the means of Group A and Group C (that received a target-criteria-based peer-assessment skill training), but no significant difference in the means between Group B and Group C. These results support the first hypothesis that students who receive target-criteria-based peer-assessment training and principle-based peer-assessment training will have higher writing scores than those who

receive no structured peer-assessment training, but not support the second hypothesis that the target-criteria-based training group students will receive higher writing performance scores than the principle-based training group students. Stated differently, these results indicate that students' peer-assessment skill positively impacts students' writing performance when using peer assessment as a strategy to improve students' writing. However, these results also revealed that Group C students did not receive significantly higher final writing scores than Group B students.

The descriptive statistics for the students' final writing scores shown in Table 9 revealed that students in Group B and Group C received substantially higher final writing scores than students in Group A. However, students in Group C did not receive substantially higher final writing score than students in Group B. The mean difference was very small (.06), which indicated that using target-criteria-based training method had no apparent superiority to using principle-based training method in the peer assessment process.

Table 9

Descriptive Statistics for Students' Final Writing Scores by Group

Group	Students' Writing Final Scores		
	N	M	SD
Group A (without structured training)	114	15.50	1.17
Group B (principle-based training)	177	16.01	1.43
Group C (target-criteria-based training)	182	16.07	1.29

Impact of Peer Assessment Training on Students' Written Feedback

In order to answer the third research question, "Is there a difference in the quality of students' written feedback among students receiving target-criteria-based peer-assessment training, principle-based peer-assessment training, or no peer-assessment training", a one-way analysis of variance (ANOVA) was used to assess difference in mean scores of students' feedback to their peers' articles based on evaluator's assessment. The difference between groups was significant, $F (2, 451) = 76.21$, $p = .001$, $\eta^2 = .253$. The strength of the relationship between the peer-assessment training treatment and the scores of students' feedback, as assessed by squared partial Eta, was strong, with the training factor accounting for 25.3% of the variance of the dependent variable.

Follow-up tests were conducted to evaluate pair-wise differences between the three groups' mean scores. The results showed that there was a significant difference between the means of Groups A and B, and Groups A and C, but no significant difference between the means of Groups B and C. These results support the third hypothesis that students who receive target-criteria-based peer-assessment training and principle-based peer-assessment training will have higher written feedback scores than those who receive no structured peer-assessment training and indicate that peer-assessment training has a positive effect on students' assessment skills.

The descriptive statistics for the students' feedback scores shown in Table 10 demonstrate that students in Group B and Group C received significant higher feedback scores than students in Group A. The mean differences were 2.6 and 2.84, respectively. However, according to follow-up test, the difference between the mean scores of students' feedback in Groups B and C (mean difference = .24), but it was not significant. This

finding indicated that use of target-criteria-based training method has no apparent

superiority to use of a principle-based training method in the peer assessment process.

Table 10

Descriptive Statistics for Students' Feedback Scores by Group

Group	Students' Feedback Scores		
	N	M	SD
Group A (without structured training)	114	23.46	2.02
Group B (principle-based training)	175	26.06	2.10
Group C (target-criteria-based training)	165	26.30	2.00

Impact of Peer Assessment Training on Quality of Student-generated Assessment

Scores

The fourth research question examined whether peer assessment training led to

improved quality (validity and reliability) of student-generated assessment scores. First,

to determine the validity of student-generated assessment scores, a Pearson product-

moment correlation was conducted to measure the relationship between student second

round rating scores and evaluator grading scores for the three Groups respectively. These

results yielded significant correlation coefficient, $[r (114) = .809, p < .001$ for Group A, r

$(177) = .822, p < .001$ for Group B, and $r (182) = .843, p < .001$ for Group C], which

indicated that the relationship between student rating scores and evaluator grading scores

was strongly correlated in the three groups, although there are some small differences

between groups on the validity measure.

A Fisher r-to-z transformation technique was used to test the difference between groups' correlation coefficients. The results showed that there was no significant difference between the correlation coefficients of Groups A and B, and Groups A and C, and Groups B and C ($z = .32$, $p = .75$, $z = .89$, $p = .37$, $z = .64$, $p = .52$). Thus the results do not support the fourth hypothesis that students who receive target-criteria-based peer-assessment training and principle-based peer-assessment training will have higher level of validity of their generated assessment scores than those who receive no structured peer-assessment training. The results of the statistical analysis of Pearson product-moment correlation are displayed in Table 11. Clearly, these data reveal three group students' rating scores strongly positive correlated to instructor and evaluator's grading scores, while Group C students had a highest coefficient (r) among the three groups.

Table 11

Descriptive Statistics on Validity of Student-Generated Assessment Scores by Group

Group	Student article numbers	r	p
Group A (without structured training)	114	.809	<.001
Group B (principle-based training)	177	.822	<.001
Group C (target-criteria-based training)	182	.843	<.001

In order to determine the reliability of student-generated assessment scores, an intraclass correlation (ICC) analysis was performed. The rating scores on the first and second-round peer assessment (three raters in the first round assessment and twenty raters

in the second-round assessment) were analyzed. The results of the intraclass correlation analysis revealed that the coefficient of ICC in Group A was moderate for the first-round assessment ($r = .62, p < .01$), but stronger for the second-round assessment ($r = .75, p < .01$). The coefficients of ICC between Groups B and C were also moderate strong for the first-round peer assessment ($r = .65, p < .01, r = .66, p < .01$), but strong for the second-round peer assessment respectively ($r = .84, p < .001, r = .85, p < .001$).

A Fisher r-to-z transformation technique was employed to test the difference between groups' correlation coefficients in first-round assessment and second-round assessment. The results showed that in the first-round assessment there was no significant difference between the correlation coefficients of Groups A and B, and Groups A and C, and Groups B and C ($z = .41, p = .68, z = .56, p = .58, z = .16, p = .87$). However, the results showed that in the second-round assessment there was a significant difference between the correlation coefficients of Groups A and B, and Groups A and C ($z = 2.04, p < .05, z = 2.34, p < .05$), but no significant difference between the correlation coefficients of Group B and C ($z = .33, p = .74$). These results partially support the fifth hypothesis in that student who received target-criteria-based peer-assessment training and principle-based peer-assessment training had higher levels of reliability of their generated assessment scores on the second-round assessment than those who receive no structured peer-assessment training.

The ICC statistics for the three groups are shown in Table 12. These findings indicated that students who had received peer assessment training had more reliable scores on the second round of assessments than those students who had not received peer assessment training. It appears that groups that received peer assessment training became

more reliable with their practice. The principle-based peer-assessment training method and the target-criteria-based training method had almost equally positive effects on the reliability of students' generated assessment scores during students' peer assessment practice.

Table 12

Descriptive Statistics on Consistency of Student-Generated Assessment Scores by Group

Group	First-round assessment		Second-round assessment	
	r	p	r	P
Group A (without structured training)	.62	$< .01$	.75	$< .001$
Group B (principle-based training)	.65	$< .01$	.84	$< .001$
Group C (target-criteria-based training)	.66	$< .01$	.85	$< .001$

Impact of Peer Assessment Training on Students' Satisfaction with Peer Assessment

Methods

The fifth research question examined the effects of peer assessment training on students' satisfaction. To address this question, scores on the Students Satisfaction with Peer Assessment Questionnaire were compared among the three Groups (Group A = 100, Group B = 72, Group C = 182). Student satisfaction with peer assessment method was measured in two domains: student satisfaction with the assessment system and student satisfaction with peer feedback. These factors served as the two dependent variables. A multivariate analyses of variance (MANOVA) was employed to determine the effects of

three types of peer assessment training (without structured peer-assessment skill training, principle-based training, and target-criteria-based training) on the two dependent variables. The results of the statistical analysis are displayed in Table 13.

Table 13

MANOVA Results for Student Satisfaction with Peer Assessment

Dependent variable	MANOVA			
	df	*F*	*p*	η^2
Wilks's Lambda = .96	4, 351	3.61	.01	.020
Satisfaction with assessment system	2	1.07	.35	.001
Satisfaction with peer feedback	2	4.66	.01	.026

The MANOVA analysis showed significant differences among the three types of peer-assessment skill training on the dependent measures, Wilks's Λ = .96, F (4,351) = 3.61, p =.01. The multivariate η^2 based on Wilks's Λ was small (.020). Analyses of variances (ANOVA) on each dependent variable were conducted as follow-up tests to the MANOVA. Each ANOVA was tested at α = .025 level. The results of these ANOVAs showed that differences in score of satisfaction with the peer assessment system were not significant, F (2,351) = 1.07, p = .35. Therefore, the result does not support the sixth hypothesis that students who receive target-criteria-based peer-assessment training and principle-based peer-assessment training will have higher levels of satisfaction with the assessment system than those who receive no structured peer-assessment training.

However, differences between scores of student satisfaction with peer feedback were significant, F (2,351) = 4.66, p = .01. Post hoc analyses of the univariate ANOVA

for the students' satisfaction with feedback scores consisted of conducting pairwise comparisons to find which training method most strongly affected students' satisfaction. The results of Scheffe's post hoc test showed that there was a significant mean difference between Groups A and C, but no significant mean differences between either Groups A and B, or Groups B and C. Thus, students who received a target-based peer-assessment training were more satisfied with the peer feedback they received than students who ether received a principle-based training or no structured training. As such, the seventh hypothesis was partially supported. Students who receive target-criteria-based peer-assessment training and principle-based peer-assessment training will have higher levels of satisfaction with peer feedback than those who receive no structured peer-assessment training. These results indicated that use of target-based peer-assessment training method most profoundly affected students' satisfaction on peer feedback.

The descriptive statistics (see Table 14) showed that the mean scores for both Groups B and C were higher than those of the Group A in students' satisfaction with peer feedback. However, the mean scores of students' satisfaction with assessment system were almost equal between Groups A and B, as well as between Group B and C.

Table 14

Descriptive Statistics on Student Satisfaction with Peer Assessment by Group

Peer-assessment Skill Training Method	Satisfaction with assessment system		Satisfaction with Peer feedback	
	M	SD	M	SD
Without structured training (Group A)	43.87	4.04	47.81	4.35
Principle-based training (Group B)	43.71	2.78	49.35	4.32
Target-criteria-based training (Group C)	44.49	5.20	49.49	4.76

Summary

In this chapter, all findings related to the research question and hypotheses are presented. Table 15 presents a summary of the results of the statistic analysis corresponding to the hypotheses.

Table 15

Summary of the Results of the Statistic Analysis Corresponding to the Hypotheses

No	Hypotheses	Results
1	Students who receive target-criteria-based peer-assessment training and principle-based peer-assessment training will have higher writing scores than those who receive no structured peer-assessment training.	Supported

2	The target-criteria-based training group students will receive higher writing performance scores than the principle-based training group students.	Failed to support
3	Students who receive target-criteria-based peer-assessment training and principle-based peer-assessment training will have higher written feedback scores than those who receive no structured peer-assessment training.	Supported
4	Students who receive target-criteria-based peer-assessment training and principle-based peer-assessment training will have higher level of validity of their generated assessment scores than those who receive no structured peer-assessment training.	Failed to support
5	Students who receive target-criteria-based peer-assessment training and principle-based peer-assessment training will have higher level of reliability of their generated assessment scores than those who receive no structured peer-assessment training,	Partially supported
6	Students who receive target-criteria-based peer-assessment training and principle-based peer-assessment training will have higher levels of satisfaction with the assessment system than those who receive no structured peer-assessment training.	Failed to supported
7	Students who receive target-criteria-based peer-assessment training and principle-based peer-assessment training will have higher levels of satisfaction with the peer feedback than those who receive no structured peer-assessment training.	Partial supported

In conclusion, the findings from the quantitative data analyses indicate that peer assessment skill training has positive effects on students writing performance, student assessment skills, and the reliability of student-generated peer assessment scores on the second round assessment. These results also suggest that students in the Group C that had the target-criteria-based peer-assessment training have a higher level satisfaction with peer assessment feedback. In addition, the results reveal that using target-criteria-based training method has no apparent superiority to use principle-based training method in peer assessment. Furthermore, the results show that peer-assessment skill training has no apparent positive impacts on the validity of students generated assessment scores during peer assessment practice. Finally, the results suggest that students in the three groups (Groups A, B and C) have almost the same levels of satisfaction with peer assessment system.

CHAPTER 5: DISCUSSION

The focus of this research has been the effects of peer-assessment skill training on four major areas: students' writing performance, students' assessment skills, the validity and reliability of student-generated assessment scores, and students' satisfaction with the peer assessment method. All data have been collected from the various sections of a single education course over the span of three semesters. Discussion of the quantitative findings for each of these areas will be provided in four sections to allow for a more comprehensive discussion of the findings. The limitations of the study and recommendations for future research and practices will be discussed. Finally, conclusions and recommendations will be presented.

Student Writing Performance

This study addressed two research questions concerning the effects of peer-assessment skill training on students' writing performance. One question led the researcher to compare the writing performance of two groups of students – those who received peer-assessment skill training and those who did not receive peer-assessment skill training. The second question was focused on which peer assessment training method more effectively impacted students' writing performance. Regarding the first question, the researcher hypothesized that students who received target-criteria-based peer-assessment training and principle-based peer-assessment training would have higher final writing scores than those who received no structured peer-assessment training. The findings did support this hypothesis. Participation in the peer-assessment training had significant effects on students' final writing scores. In the second case, the researcher hypothesized that the target-criteria-based training group students would receive higher

final writing performance scores than the principle-based training group students. However, the findings showed that there was no significant difference in the means of students' writing scores between students in target-criteria-based training group and students in principle-based training group. The target-criteria-based training group students had only slightly higher average writing scores than the principle-based training group students. This finding indicated that using the target-criteria-based training method had no apparent superiority to the principle-based training method in the peer assessment process.

The results the one-way analysis of variance (ANOVA) showed that students who participated in peer-assessment training achieved better writing performance than those who did not participate in peer-assessment training. This finding provides empirical support to the arguments of some researchers, who believe that peer-assessment skill is one of the variables that influence the quality of peer-assessment, because most students lack assessment skills and knowledge and suggested that students should have peer-assessment skill training and instructor's support before they engaging in peer assessment (Cheng & Warren, 1997; Falckikov & Goldfinch, 2000; Sluijsmans et al., 2002; Topping, 1998). Cheng & Warren (1997) noted that students doubted the objectivity of peer assessment in the peer assessment process and claimed to have no training in such assessment practices. Falckikov & Goldfinch (2000) also found the students lacked confidence in doing peer assessment because they lacked the peer-assessment skills and doubted the quality of peer assessment. Accordingly, Sluijsmans et al., (2002) stressed that peer assessment requires students to incorporate complex skills for which students often need to be supported or trained before students can effectively engage in peer

assessment because most students are naïve. Topping (1998) has asserted that peer

assessment is a reflexive act, and has pointed out that in the context of peer assessment,

learning, which is an enterprise traditionally achieved solely by teaching, can now also be

accomplished by assessing. In another study, Topping et al., (2000) found that formative

peer assessment seems likely to be most helpful for students' learning if it produces rich

and detailed qualitative feedback information about strengths and weakness of students'

products not merely a mark or a grade. The peer assessment process requires the

assessors to review, summarize, clarify, give feedback, diagnose misconceived

knowledge, identify missing knowledge, and consider deviations from the ideal (Van

Lehn et al, 1995). All these activities incorporate complex assessing skills and meta-

cognitive learning process. In contrast to the assessor's role, the assessees' learning can

be improved if assessees' products' strengths or merits are rewarded or positively

reinforced and if assessees' weaknesses or shortcomings are negatively reinforced.

Accordingly, peer-assessment skills play an important role in peer assessment. It is

important to train students to develop the capability to ask intelligent, adaptive questions

in peer assessment (Topping, 1998). The results of the present study demonstrate that

peer-assessment training has a positive impact on students' writing performance; thus, the

present findings help fill the void in the literature because support for the benefits of

peer-assessment skill training on students' writing performance are tenuous based on

previous studies.

The finding that peer-assessment skill training positively impacted students'

writing performance is consistent with a related study's findings (Sluijsmans et al., 2002),

which revealed that pre-service teachers who had peer-assessment skill training in the

experimental group received higher scores on their learning task, designing creative

lessons, than those who did not receive peer-assessment skill training in the comparison

group. In addition, this result also expanded the literature on the effects of peer

assessment on students' writing performance and the strategies used in peer assessment

on a purpose to improve the quality of peer assessment as well. Richer (1992) compared

the effects of two kinds of feedback, peer-directed and teacher-based, on college students'

writing proficiency in an experimental study. The results showed that there was a

significant difference in writing proficiency in favor of the peer-feedback-only group.

When comparing three methods for teaching student writing, Plutsky and Wilson (2004)

found that peer feedback helped students become proficient writers. Xiao & Lucking

(2008) also found that participants in the group using a rating-plus-peer-feedback method

demonstrated greater improvement in their academic writing than those in the rating-only

group when comparing two peer assessment methods on university students' academic

writing performance. Working with college computer science major students, Lin et al.,

(2001) compared the effects of two kinds of peer feedback – specific and holistic – on

students' writing according to students' different thinking styles in a web-based online

peer assessment environment. One finding showed that students with high executive

thinking styles who received holistic and specific peer feedback, and students with low

executive thinking styles who received specific peer feedback group significantly

improved their writing. However, students with low executive thinking styles who

received holistic peer feedback did not improve their writing. Miller (2003) compared

two kinds of assessing criteria (general vs. specific). The results showed that students

using specific criteria (25 criteria) earned higher assessment results and produced more

critical feedback than those students using general criteria (5 criteria). Fagot (1991) argues that multiple ratings are superior to single ones, and when individual students were poor judges, increasing the number of raters improved the reliability of averaged scores. Cho et al., (2006) found that six peer reviewers produced higher effective reliabilities than three or four peer raters. In an earlier study, Haaga (1993) suggested that using anonymous peer review was an effective strategy in helping students to improve their papers. Similar findings indicate that anonymous peer review produced more honest, accurate, and critical feedback that helps student improve their writing skills (Zhu, 1994). Additionally, Lu and Bol (2007) compared the effects of anonymous and identifiable electronic peer review on college student writing performance. The results of the study showed that there was a significant difference in students' writing performance in favor of students participating in anonymous e-peer review. Based upon a thorough review of the related literature, the present study adds to the discussion of peer assessment, in general, and of peer-assessment skill training, in particular, aimed at improving students' writing performance.

In this study, the results showed that the scores on students' final written articles were significantly different between groups, but the effect size was small ($\eta^2 = .031$), which indicated that students in Groups B and C showed small gains in writing scores after training compared to students who did not receive training in Group A. Several factors could account for this result based on the author's observations during these semesters. First, writing an expository essay (Wiki article) in this course is one of eight major assignments. In addition, the peer assessment task was only one part of the writing (Wiki article) project. Considering students' and GAs' substantial workload, the course

instructor did not give any additional assignments to reinforce and assess the peer

assessment training content immediately after each training session. Instead, Group B

students (those who had principle-based training) did practice exercises in the classroom,

and Group C students (those who had target-based-training) did practice exercises outside

the classroom and then engaged in the two-part assessment process on their own. The

failure to assess students' peer-assessment training exercises could impact the

effectiveness of training. More preliminary practice and strict assessment could improve

students' confidence and expertise in peer assessment.

Groups B and C students' small gains in writing scores after training may also be

caused by some students' failure to critically evaluate their peers. Because students

provided feedback to their peers under a semi-identified condition in the first round

(formative assessment) of the peer assessment process, some students might have been

dishonest or lenient because of interpersonal relationships or friendships, although the

peer assessment required that students identify at least two weaknesses from their peers'

writing. Indeed, students' responses to the questionnaire indicated that about 29% of

students felt reluctant to give negative feedback to their classmates, which indicated they

felt uncomfortable in assessing their peers although the instructor stressed the functions

of critical feedback in the training process. A small number of students still provided very

lenient negative feedback to their peers. This fact is consistent with previous studies

related to the effects on assessment condition and critical feedback, and the relationship

of critical feedback to students' learning (Kerr et al., 1995; Lu & Bol, 2007; Zhu, 1995).

Zhu (1995) argues that if students do not approach their peers' writing critically, they will

fail to provide meaningful and useful feedback. Kerr et al., (1995) asserts that students

who take a critical approach when reading and scoring peers' work are likely to be more

critical of their own work, and thus create improved writing. A study by Lu and Bol

(2007) revealed that students participating in anonymous e-peer review received

significantly higher writing scores than those of students in identifiable e-peer review. In

addition, students in the anonymous e-peer review group provided significantly more

critical feedback to their peers than did students participating in the identifiable e-peer

review. Clearly, critical feedback impacts writing and anonymous assessment condition

may help students avoid complications associated with their peer relationships in the

assessment and provide more critical feedback.

The finding of no significant difference between Groups B and C in writing

scores may be attributable to the design and implementation. A problem in the design of

the present study could be that the two types of training methods were too similar.

Although there were some differences between the two training methods, students in both

Group B and Group C had the same training tasks, the same training time, and similar

training contents. In addition, students in Group C (target-based training group) had peer-

assessment skill assignment exercises outside the classroom on a purpose to master

assessment skills (one of the differences in the training methods), but lacked strict

assessment and feedback from instructor (only checked as complete or incomplete),

which might diminish the impact of the training exercises. More training exercises along

with more substantive instructor assessment may have resulted in more positive effects

on students' learning products.

Quality of Students' Peer Feedback

The third research question investigated the effects of the peer-assessment skill training on the quality of students' peer feedback in the peer assessment process. To better inform the discussion, it is necessary to re-state the terms used in this study. Peer feedback refers to a means of students' communication – assessors' comments related to assessees' writing performance and standards, which were posted online in a Wikibook discussion page. Peer rating refers to students' grading the article writing performance of their peers using relevant criteria. Students' peer feedback was assessed using a specific rubric developed for this study -- the Rubric of Quality of Peer Feedback that reflected three aspects of peer assessment skills: (1) defining assessment criteria or using criteria; (2) judging the performance of a peer reflecting upon and identifying the strengths and weaknesses in a peer's writing; and (3) providing feedback for improvement and future learning. The results revealed that students who had peer-assessment training received significantly higher scores than those students who had not received peer-assessment training. The strength of the relationship between the peer-assessment training treatment and the scores of students' feedback was strong (η^2 = .253). The follow-up test results showed that there was a significant difference between the means of Groups A and B, and Groups A and C, but there was no significant difference between the means of Groups B and C. These results support the third hypothesis that students who received target-criteria-based peer assessment training and principle-based training would have higher written feedback scores than those who did not receive structured peer-assessment training. The results indicate that peer-assessment training had a positive effect on students' assessment skills.

This finding is consistent with previous studies of peer-assessment training on students' assessment skills. Zhu's (1994) experimental study reported that training students for peer revision led to more and higher quality peer feedback after students were trained through special teacher-student conferences in small groups. Sluijsmans et al., (2002) found that peer-assessment training developed and improved pre-service teachers' peer-assessment skills. Their study results showed that pre-service teachers in the experimental group were more likely to use the criteria and give more constructive comments than those in the comparison group that had received no peer assessment training. Also, the pre-service teachers in the experimental group scored higher on feedback structure and used fewer naïve words than those in the comparison group. Additionally, Sluijsmans et al., (2004) found that pre-service teachers who received peer-assessment training, focusing on defining criteria, were more capable in using the set criteria determined during the peer-assessment tasks than those in the comparison group who did not receive peer-assessment training. In addition, this result also expanded the literature on the effects of quality peer feedback on the quality of peer assessment and students' writing. Previous studies indicated that lower quality of peer feedback is mostly associated with the students' writing ability, assessment skills, and types of peer feedback. In evaluating a 2-year program involving peer grading and peer feedback to their peers' essays in a microeconomics course, Kerr et al., (1995) found that students with better writing ability were better at the task of grading the essays of their peers (in better agreement with the instructor grading) than those with lower writing ability; students with better writing ability provided better quality feedback than those with lower writing ability. In an experimental study, Lin et al., (2001) compared the effects of different peer

feedback formats (holistic vs. specific) and different thinking styles (high executive thinking styles vs. low executive thinking styles) on students learning outcomes. The findings of this study showed that thinking style and feedback format interactively affected student learning. Low executive students receiving specific feedback significantly outperformed those receiving holistic feedback. Similarly, high executive students receiving specific feedback did slightly better than those receiving holistic feedback, but not significantly. The authors argued that while executive thinkers could overcome the disadvantages of holistic feedback, the low executive thinkers could not.

Sluijsmans and colleagues' studies of peer-assessment training on students' assessment skills focused on evaluating quantitative use of criteria, positive and negative comments, and naïve word use when students provided feedback to their peers. In contrast, in the current study of peer-assessment skill training, the evaluation of students' assessment skills centered on the quality of peer feedback since many previous studies demonstrate that students' peer feedback positively impacts students' writing performance (Lin et al., 2001; Plutsky & Wilson, 2004; Richer, 1992; Xiao & Lucking, 2008). The design of the evaluation rubric and the content of peer-assessment training are in agreement with both Sadler's perspective and Nicol & Macfarlane-Dick's perspective. Sadler (1989) identified three essential conditions necessary for students to benefit from feedback in their learning tasks. He asserts that students have to know what good performance is, how current performance relates to good performance, and how to act to close the gap between current and good performance. Nicol & Macfarlane-Dick (2006) also argues that good feedback should help clarify what good performance is, facilitate the development of self-assessment (reflection) in students' learning, deliver high quality

information to students about their learning, encourage positive motivational beliefs and self-esteem, and provide opportunities to close the gap between current and desired performances.

The present study results showed that students in Group B and Group C received significantly higher feedback scores than those of students in Group A did. Based on the scores that students gained from using this five criteria rubric (six points for each criterion), the results indicated that students who had structured peer-assessment skill training performed better on each criterion of the evaluation rubric than those students who did not have structured peer-assessment training. Students who had peer-assessment skill training were more likely to use the writing assignment criteria when they provided feedback to their peers. The feedback provided by students who had peer-assessment skill training was more specific and understandable than that submitted by students who did not receive structured peer-assessment training. In addition, the suggestions in feedback provided by students who had peer-assessment skill training were more achievable than those submitted by students who did not receive peer-assessment skill training. Although the assessment skills are a set of complex skills, especially in providing useful feedback, students can acquire and improve their assessment skills through training. The findings further support the previous studies of peer-assessment skill training on students' assessment skill (Sluijsmans et al., 2002; Sluijsmans et al., 2004; Zhu, 1994).

The finding that there was no significant difference between Groups B and C in students' written feedback scores (Group B, M = 26.06 and Group C M = 26.30), which may be resulted from the design and implementation. A problem in the design of the present study could be that the two types of training methods were too similar to each

other. Although the design of the target-based training (Group C) was improved based on the principle-based training (Group B), students in both Group B and Group C had the same training tasks and training time; they also had similar training contents. In addition, students in Group C (target-based training group) had peer-assessment skill assignment exercises outside classroom on a purpose to master assessment skills (one of the differences in training methods), but lacked strict assessment and feedback from the instructor (only checked as complete or incomplete), which might diminish the positive impact of the training exercises. If the students in Group C were given more time in training and more exercises along with more substantive instructor assessment, their written feedback scores would be higher than those students in Group C who had the principle-based training.

In short, the present study showed that peer-assessment skill training has positive effects on students' assessment skills. Students' assessment skills can be developed and improved through a structured training. If students had more training time and students had more exercises along with instructor assessment, the training results would more positively impact students' assessment skills.

The Validity and Reliability of Student-generated Assessment Scores

Quality of peer assessment has been the concern of many researchers and practitioners. In the current study, the fourth research question sought to examine whether peer-assessment training leads to valid and reliable student-generated assessment scores. Nearly 500 students' rating scores were collected from two rounds of peer assessment during the three semesters included in the study. In the first round peer assessment, students were assigned to groups of three students to provide peer feedback and ratings

based on the Wiki Article Evaluation Rubric, the first round peer assessment was formative, while the second round peer assessment was summative. In this stage, groups of 20 students rated peers' articles. However, for practice, the instructor encouraged students to do more peer ratings in addition to those required of their group, so each student received more than 20 peer ratings in fact.

According to the results of data analysis, students who received peer-assessment training in Groups B and C had more reliable assessment scores in the second round of peer assessment than students who did not receive peer-assessment training in Group A, but the three groups of students had similarly reliable assessment scores in the first round peer assessment. These results partially support the hypothesis that students who receive target-criteria-based peer-assessment training and principle-based peer-assessment training will have higher levels of reliability of student-generated assessment scores than those who receive no structured peer-assessment training. The results also showed that the student-generated assessment scores were equally valid for those students who received training as well as those who did not. This result does not support the hypothesis that students who received target-criteria-based peer-assessment training and principle-based peer-assessment training will have scores with higher levels of validity than those who received no structured peer-assessment training.

The Validity of Student-Generated Assessment Scores

The Pearson product-moment correlation results yielded strong correlation coefficients across groups ($r = .809$ for Group A, $r = .822$ for Group B, and $r = .843$ for Group C), which indicated that the relationship between student rating scores and evaluator grading scores was significant and strongly correlated among the three groups.

These findings are consistent with previous studies on the validity of student-generated assessment scores in writing. Cho et al., (2006) found that the validity of student-generated rating scores was relatively high and similar among 708 students across 16 courses in different subject areas over three years. Stefani (1994) and Saito and Fujita (2004) found high validity of student-generated assessment scores in their studies. Likewise, Xiao and Lucking (2008) found high validity of student-generated assessment scores when students assessed their peers' writing in an online environment.

The results indicated that all participating students, regardless of whether or not they had any peer-assessment training, generated assessment scores with equally high levels of validity. The training seems to have had no effect on the validity of students' rating scores, which is an unexpected result.

The validity of students' grading or rating in the summative function of peer assessment has been explored by many teachers and researchers. Some studies have been conducted to investigate the validity of students' grading or rating from instructors' perspective (Chen & Warren, 1999; Cho, Schunn, & Wilson 2006; Falchikov, 1986; Mowl & Pain, 1995; Saito & Fujita, 2004; Stefani, 1994; Topping, et al., 2000; Xiao and Lucking, 2008). Almost all these studies on effects of peer assessment on student academic writing used quantitative methods, except for one study that employed qualitative methods (Topping, et al., 2000). However, the techniques used to calculate the value of validity were not consistent in these quantitative studies. Most of these studies computed validity by using average means of peer generated scores against an instructor's scores, but one study (Falchikov, 1986) used percentage agreement between the individual student's rating and an individual instructor's rating. Among them, three

studies did not describe exactly how validity was assessed, but it is likely that validity was computed in the usual way by comparing students' rating scores against instructor-assigned grades. All these studies' results showed that the validity of students' rating scores was diverse. Cheng and Warren (1999), and Mowl and Pain (1995) reported low validity. In contrast, Stefani (1994), Cho, Schunn and Wilson, Saito & Fujita (2004) and Xiao and Lucking (2008) found high validity of students-generated rating scores. It is worth noting that the two studies that reported low validity did not describe exactly what criteria or rubric was used for students to assess their peers' work. One study mentioned that global judgment considering some aspects was used, and the other stated that a five-ponit-scale rating form was used, but it gave no further specific information. In contrast, those studies that reported high validity described or discussed exactly the criteria or rubric with rating scales for both students and instructors. The only exception was Stefani (1994), where global judgments were used for peer assessment. In addition, another noticeable fact is that in Mowl and Pain's (1995) and Falchikov's (1986) studies, single raters were used to assess a peer's product. Mowl and Pain's study yielded low validity of peer assessment, while Falchikov's study generated moderately high validity of peer assessment, but used the percentage agreement method to assess validity, which was criticized for inability to detect quality and inconsistence with other studies (Cho, Schunn, & Wilson 2006). In Cheng and Warren's study, the authors did not describe how many raters assessed each peer's product, but it seems that multiple raters were used. In contrast, other studies that reported high validity of students' assessment scores stated that multiple assessors were employed for assessing each peer's products, yielding high peer assessment score validity.

The present results showed that the validity of student-generated assessment scores was almost equally high in each of the three student groups regardless of whether or not they received peer-assessment training. There is little empirical research related to the effects of peer-assessment training on the validity of students' assessment scores. In the author's perspective, the variables that influence the validity of students' assessment scores are complicated. Several other variables may have much more influence on the validity of students' assessment scores than the peer-assessment skill training does. First, the validity of student assessment scores may rely on the validity and reliability of the assessment criteria or rubric and scoring system from a traditional assessment perspective. If the scoring criteria cannot be understood due to vagueness or difficulty, or if the rating scale offers very few choices for scoring, then an inaccurate, unfair judgment (too low or too high) can be made, which causes inaccurate and unfair assessment to those being assessed. In addition, if the instruments preclude the discrimination of performance, they lack formative or summative value for the students who are assessed (Miller, 2003). MacAlpine (1999) found when student raters used four items (criteria) rated on a 5-point Likert scale instead of a single letter (A, B, C, D and E) in peer assessment, students were better able to discriminate their peers' performance. Miller (2003) found that increasing the number of criteria (from 5 criteria each using 5-point-rating scale to 25 criteria each using 5-point-rating scale) decreased the mean scores of students' grades and increased the standard deviation of the peer and self-assessment, which improved students "over-grading" in peer and self-assessment. Miller (2003) suggested that more specific criteria may have allowed the assessors to reflect on more aspects of the quality of the performance. Saito and Fujita (2004) found that students using valid and reliable

assessment rubric (r = .91) with multiple raters in peer assessment generated assessment scores whose validity was high and not affected by "over-grading". A similar study, Xiao and Lucking (2008), also reported that in their peer assessment, students using valid and reliable assessment rubric (r = .84) with multiple raters generated assessment scores with high levels of validity.

Second, ensuring that students fully understand criteria and including students in the development of assessment may increase the validity of their assessment scores (This is one of the elements of peer-assessment skills training). Some studies suggested that when students are involved discussion about the assessment criteria and negation about performance there is a positive effect on quality of peer assessment (Falckikov, 1995a; Orsmond & Reiling, 2000; Stefani 1994; Toping, 1998). Falchikov and Goldfich (2000) reported that student familiarity with and involvement in creating assessment criteria is a key to the enhancement of peer assessment validity after evaluating 48 studies that focused on the quality of students' rating or grading scores. Third, peer assessment in an anonymous assessment condition may enhance the validity of students' assessment scores. Some studies has argued that anonymous peer assessment can be used to minimize opportunities for students to reward friends or otherwise game the system during peer review process (Haaga, 1993; Kerr et al.; Lu & Bol, 2007; Vinson, 1996; Zhao, 1998). Finally, multiple raters assessing each student product may produce higher validity of students' assessment scores. Using multiple raters may increase the accuracy of the students' assessment scores by averaging and also may lessen the effects of individuals' biases. Many previous studies suggested that multiple assessors can be used to balance the uneven quality of peer assessment, both peer reviews and peer ratings (Cho et al.,

2006; Falchikov & Goldfinch, 2000; Kerr et al., 1995; Nilson, 2003; Quible, 1997; Robinson, 1999; Topping, 1998; Xiao & Lucking, 2008).

In the present study, the three groups of students experienced two rounds of peer assessment in nearly the same assessment conditions: using same assessment criteria rubric ($r = .87$), same numbers of raters assessing each student's article, the same assessment condition, and involving discussion of assessment criteria. The difference is that Group A had a brief peer, loosely-structured assessment introduction that focused on understanding the assessment criteria and peer assessment process, while Groups B and C had structured peer-assessment skill training. According to the current study's results the validity of student-generated assessment scores was almost equally high in the three groups. Perhaps, the variables that were discussed above may affect the validity of student-generated assessment scores much more profoundly than the structured peer-assessment skill training does. A replication research is recommended. The validity of student-generated scores would be increased if the training time is extended and the training method is improved.

Reliability of Student-Generated Assessment Scores

Although the results of the current study did not support the hypothesis that students who receive target-criteria-based peer-assessment training and principle-based peer-assessment training will generate peer assessment scores with higher levels of validity than those scores generated by students who receive no structured peer-assessment training, the results showed that students in groups B and C who received structured peer-assessment training yielded higher levels of reliability of students' assessment scores in the second round peer assessment than the students who did not

receive peer-assessment training in the Group A. The reliability of student-generated

assessment scores was defined in Chapter Three of the current study as the consistency of

students' assessment scores applying a scoring rubric to writing evaluation. The students'

first and second round rating scores were analyzed by using intraclass correlations (ICC),

which is a common measure of reliability of either different assessor or different items on

a scale (Cho et al., 2006; Shrout & Flesiss, 1979; Xiao & Lucking, 2008). According to

Cho et al., (2006), the ICC (C, k) Case 2 formula would effectively estimate the

consistency of k reviewers combined (a case of random reviewers, and random articles).

Therefore, using the (C, k) Case 2 formula is appropriate to determine the reliability of

student rating scores in this setting.

The results of the ICC analysis showed that the ICC in Group A was moderately

strong for the first-round (three student ratings) assessment ($r =.62$), stronger for the

second-round (20 student ratings) assessment ($r = .75$), and the coefficient of ICC

between Groups B and C was also moderately strong for the first-round peer assessment

($r = .65$, $r = .66$) and strong for the second-round peer assessment respectively ($r = .84$, r

$= .85$). The results of second-round assessment showed that students who had peer-

assessment training generated more reliable assessment scores than those who did not

receive peer-assessment training. The reliability of all three groups' students' scores in

this study was high, a finding consistent with previous studies on reliability of student

generated scores in writing (Cho et al., 2006; Haaga, 1993; Xiao & Lucking, 2008).

The results also showed that twenty raters (second round) yielded higher

correlation coefficient values than those of three raters (first round) in the three groups

respectively. These results concur with the findings of previous studies (Cho et al., 2006;

Xiao & Lucking, 2008). Cho et al., (2006) found that six assessors produced rating scores with higher reliability than those of three or four assessors in an online peer assessment environment. Xiao and Lucking, (2008) also found that twenty raters yielded higher reliability of students' assessment scores than those three raters in an online peer assessment environment. These results supported Ferguson's (1966) assertion that the use of multiple raters tends to improve reliability by increasing the ratio of true score variance to error variance and Fagot's (1991) argument that multiple ratings are superior to a single one and further supports the findings of Magin (1993) that if individual students are poor judges, the reliability of averaged scores can be increased by increasing the number of raters.

In contrast to prior research conducted on validity in writing, there are fewer studies that have explored reliability. In addition, the techniques used to calculate the value of reliability were not consistent. Haaga (1993) reported relatively high reliability ($r = .55$) of student-assigned grades when comparing two student evaluators' assessment scores using the Pearson product-moment correlation. Marcoulides and Simkin (1995) reported that peer reviewers seemed to be consistent evaluators, when each paper was reviewed by three students by using a percentage-of-variance approach. Cho et al., 2006 and Xiao and Lucking, (2008) reported high reliability of student-generated assessment scores by using ICC to determine the degree of consistency of student-generated assessment scores. However, no previous studies of peer-assessment skill training have investigated reliability of students-generated assessment scores; therefore, this study extends the scholarship inquiry in this field. It is also important to acknowledge that although the two experimental groups of students (Groups B and C) yielded almost the

same levels reliability of students-generated assessment scores, the two experimental groups' student-generated assessment scores were stronger than those students in Group A who had no peer-assessment training in the second round peer assessment.

The result of this study showed that in the first round peer assessment, there is no apparent difference in the reliability of students-generated assessment scores between students who received training and those who did not. This may account for the effects of training practice. Specifically, when students were rating their peers' articles in first round peer assessment, the training effects might not be expressed in increased levels of student-generated score reliability, but as they continued to practice, the reliability of their assessment scores improved.

In short, the variables that impact on the validity and reliability are complex. In the author's perspective, perhaps the most important influences the quality of students-generated assessment scores are the quality of assessment criteria and rubric or scoring system, students participation and discussion of the assessment criteria and scoring system, assessment conditions, the numbers of assessors and students' assessment skills. Since there are few studies that focus on peer-assessment skill training and its impact on the validity and reliability of students-generated assessment scores, more research is needed in this area. In this sense, the results of the current study may provide some reference for future researchers.

Student Satisfaction with Peer Assessment

The last research question, in current study, examined the effects of peer-assessment training (peer assessment method) on students' satisfaction. Students' satisfaction with peer assessment was measured in two domains: student satisfaction with

the assessment system and student satisfaction with peer feedback. The results of the analyses revealed that differences in scores of satisfaction with the peer assessment system were not significant between students who had peer assessment training and students who did not. This finding did not support the sixth hypothesis. However, the results showed that the students who received target-criteria-based peer-assessment skill training in Group C had significantly higher scores of satisfaction with the peer feedback than the students in other groups had. The finding partially supported the seventh hypothesis. These results indicated that students who received target-criteria-based peer-assessment skill training were more satisfied with peer feedback they received and almost equally satisfied with peer assessment system in class when compared to the students in Group A and B.

These results are consistent with previous research into the effects of peer-assessment skill training on students' satisfaction (Stuijsmans et al., 2002; Stuijsmans et al., 2004; Zhu, 1994). Zhu (1994) reported that after a special teacher-student conference designed to train students in developing and practicing strategies to conduct revisions of peers' writing tasks, students in an experimental group demonstrated improved attitudes toward peer revision and engaged in peer revision tasks more actively. Stuijsmans et al., (2002) reported that after four sections of peer-assessment skills training that focused on defining criteria, giving constructive feedback and judging the performance of a peer and peer assessment practices, pre-service teachers in experimental groups exhibited higher levels of satisfaction with peer assessment than those in the comparison groups. Likewise, Sluijsmans et al., (2004) reported that pre-service teachers in the experimental group who received peer-assessment training that only focused on defining criteria showed more

positive attitudes toward peer assessment than those in the comparison group. The results also showed that pre-service teachers who received peer-assessment training were more satisfied with the class than those who did not receive peer-assessment training.

The results of the current study indicated that students who had target-criteria-based peer-assessment skill training had higher levels of satisfaction with peer feedback than those students in other groups. It is noteworthy that students who had peer-assessment training in Groups B and C had higher scores in satisfaction with the quality of peer feedback and satisfaction with benefits from providing and receiving feedback than those students in Group A. This finding indicated that peer-assessment skill training affected students' satisfaction most strongly on these two subscales: the quality of peer feedback and the benefits of providing and receiving feedback. These findings support the contention that it is important that this process includes feedback and not just grades or scores (Davies, 2006; Stefani, 1998) and "formative assessment seems likely to be most helpful if it yields rich and detailed qualitative feedback information about strengths and weaknesses, not merely a mark or a grade" (Topping, 2000, p.150). However, students in the three groups had similar scores on satisfaction with the process of providing and receiving feedback, which implied that the ways of providing and receiving peer feedback in this online peer assessment environment did affect the students' satisfaction with peer feedback for students who received training as well as those who did not.

In the current study, students experienced two rounds of peer assessment, the first round being formative and the second summative. Overall, all students in the three groups expressed a high level of satisfaction with this online peer assessment. Over 80%

of students not only showed positive attitudes toward peer assessment, but they valued peer assessment as a worthwhile activity and acknowledged benefit from providing and receiving peer feedback. These results concur with previous results related to students' attitude toward peer assessment (Li & Steckelberg, 2004; Liu et al., Prins et al., 2005; 2001 Xiao & Lucking, 2008). Based on a survey, Li and Steckelberg (2004) reported that students expressed a high level of satisfaction toward computer- meditated peer assessment process in which they were actively involved. Liu, Lin, Chiu, and Yuan (2001) reported that nearly 70% of participants stated that they preferred using peer review for their writing assignments, and most of them viewed peer review as equally effective to instructor's feedback -- all positive responses to the peer assessment. In a case study aimed to determine students' attitudes toward peer assessment in a computer supported collaborative learning environment, Prins, Sluijsmans, Kirschner, and Strijbos (2005) found that students had a positive attitude toward the use of peer assessment and were actively involved in the peer assessment assignments. Xiao and Lucking (2008) also found that participants in the group of using a rating-plus-qualitative-feedback method of peer assessment exhibited higher levels of satisfaction with peer assessment than those in the rating-only group, and over 80% of participants showed positive attitudes toward peer assessment in an online assessment environment. It is worth mentioning that although students experienced summative peer assessment in the second round peer assessment, the findings did not indicate that students has developed a negative attitude toward peer assessment, a finding that corresponds to that of a previous study. Saito and Fujita (2004) investigated the effects of college students' attitudes toward peer assessment. The study's results showed peer feedback ratings do not seem to influence student attitudes toward

peer assessment. It is also worth noting that over 85% students agreed or strongly agreed that giving feedback to their peers and taking feedback from their peers was an effective approach to improving their critical thinking skills, which supports Topping's (1998) contention that peer assessment might strengthen students' critical thinking skills.

Limitations

Although this study was carefully designed and strategies have been used to minimize threats, as with other studies of this type, some threats to the study's findings' external validity and internal validity exist. The major threats exposed in this study were problems of sampling, subject selection, instrumentation and ecological external validity, that is, equality of study environmental conditions.

Sampling could be considered a problem since there was no attempt made to randomize the selection of subjects that participated in this research. This study simply employed a convenience sampling strategy from the accessible population. Because the subjects were not randomly selected, the characteristics of subjects might not be the same as the target population in terms of characteristics like socioeconomic status, age, and ability. Since this is a quasi-experimental design, it is difficult to minimize the effects of selection biases, because attempting to draw subjects randomly from the target population was extremely difficult as is often the case with quasi-experimental studies in education settings. In addition, the peer assessment method involved a unique two-round peer assessment in an online, Wikibook environment. As a result, the generalizability of this study's findings is limited.

Formation of groups is a threat to internal validity due to the fact that random assignment of participants to groups was not conducted. As a result, students enrolled in

Spring 2007 were assigned as comparison group and students enrolled in Spring 2008

and Fall 2008 were assigned as experimental groups, and each group consisted of on

campus and online students. It is possible that there were some differences between the

groups, such as students' course load, technology back ground and attitudes toward peer

assessment. These potential threats may impact internal validity of this study's findings,

because the students' success in this course, especially, in this particular online peer

assessment, much relied on these factors. However, to attempt minimize the effects of

selection, student enrollment GPA, age, gender and ethnic background information were

collected to check differences between groups. These data indicated that there were no

significant differences in means of students' GPA and age between groups and also no

significant differences in means of students' GPA between online students and on

campus students in each group before they entered the experiment. This attempted

checking on student background helped assure general equality of groups.

Chief among the issues of validity that might have played a part in this study is

the interaction of external validity and internal validity in the context of threats due to

history and those due to ecological validity. Because the three groups' subjects were

drawn from different semesters and the experimental period lasted throughout three

semesters, the conditions might be slightly different from semester to semester. Some

threats may impact the external and internal validity. First, the differences in course

schedule across semesters may impact the validity. Students enrolled in semesters of Fall

2007 and 2008 had morning class schedule (8:00 to 9:15 AM), while students enrolled in

semester of Spring 2008 had an afternoon class (1:00 to 2:15 PM). Different class

schedules may impact student learning success results. Likewise, different class

schedules also may impact the instructor's teaching results. Second, the support of teaching assistants' for the classes may not have been equal in the three semesters due to the change of assessments in this course. A teaching assistant with rich experience was substituted by a new assistant during the beginning of the Fall 2008 semester. The new, untrained teaching assistant was unfamiliar with this particular online, Wikibook technique used in this course and also unfamiliar with the online peer assessment structures at beginning of that semester. Support for students' learning was likely weaker than that available to students during other semesters. Third, another threat to the validity of this study may come from the course instructor. Although the three groups of students were taught by the same instructor, it did not guarantee that the prepared course plan, curriculum content and related information were presented in exactly the same ways to the three groups. These threats may limit the generalizability of this study's findings.

Instrumentation may also be a threat to internal validity. On the one hand, assessing student writing and student assessment feedback is a formidable task in this study with such a large number of subjects (n = 473). There is no guarantee that an evaluator would grade the last paper with exactly the same criteria as he/she graded the first paper, even though the same instrument was used throughout. In addition, because of the large numbers of students' articles and assessment feedback, the evaluation lasted several weeks due to the feedback not being returned in timely manner. Additionally, measuring student satisfaction with peer assessment involved the use of self-reported data. Even though students were informed that their responses would not affect their course grades or instructor, some students might consider some negative consequences to them or to the instructor caused by their responses to the questionnaire. Likewise, because

there was no connection between the course grade and completing of the questionnaire, and because responding to the questionnaire was voluntary, some students might not have taken it seriously. Therefore, all these instrumentation issues may impact the study's internal validity.

Although several strategies were employed in order to minimize threats to external and internal validity associated with extraneous variables, some extraneous variables could not be completely controlled, and their potential impacts must be considered. In summary, due to the threats to the external validity of this current study, the findings may not generalize to all other subjects or tasks. With all of the limitations of this study, implications for practice and suggestions for future study will be presented in following sections.

Recommendations

Recommendations for Instructional Practice

The findings of the present study suggest that peer-assessment skill training has positive effects on university students' writing, peer-assessment skills, and quality of peer assessment. Based on the results of this study, some recommendations for implementing peer-assessment skill training in higher education may be made. Given that the primary reason for peer-assessment training is to improve students' assessment skills and their learning, the following strategies may be useful:

1. Design peer-assessment skill training integrating in the course learning objectives (Sluijsmans et al., 2002; Sluijsmans et al., 2004).

2. Provide several sessions of training over a sustained period, allowing time for practice exercises between sessions.

3. Spend more time training students to understand and negotiate the criteria of a specific product, and let all students know what a good performance is and what a poor performance is; it is important to know that understanding the criteria of a specific product in a certain subject area can lead to improved feedback for peers and also improved awareness of the quality of one's own work.

4. Provide opportunities for students to practice assessment skills in a non-threatening environment, such as by providing formative feedback to their peers.

5. Strengthen students' accountability throughout the peer assessment process.

6. Include a strategy for reviewing students' peer assessments, either by having instructors evaluate the peer assessment or by having the student authors report on the usefulness of the assessment completed by their peers.

7. Have all students use pseudonyms or avatars in the peer assessment (no use of real names) to ensure the anonymity of peer assessment by using online interactive software technology.

8. Use multiple reviewers in peer assessment to ensure high quality of peer-assessment (Falchikov & Goldfinch, 2000; Cho, et al,. 2006; Xiao & Lucking, 2008).

It is also important to remember that peer-assessment skill training in peer assessment has many benefits in terms of improvement to students' learning and assessment skills (Sluijsmans et al., 2002; Sluijsmans et al., 2004). In addition, the quality of students' feedback has vital effects on students' product improvement.

Therefore, the priority of training success is to train students to understand the assessment criteria and improve their ability to analyze the quality of the peers' products. Finally, previous studies (Cho, et al., 2006; Xiao & Lucking, 2008) and the current study provide evidence that peer-generated grades are sufficiently reliable and valid from instructor's perspective and can be used to augment instructor's assessment in the classroom.

Suggestions for Future Research

Implementations of peer-assessment skill training are likely to continue in this area. Further research needs to ascertain what peer-assessment skill training strategies are most effective in providing students an opportunity to be more successful in their writing. Although the findings of the current study partially filled the gap in this area and may provide useful information, the study of effects of structured peer-assessment training on students' writing is still tenuous, and there exists a great potential for further research in this area.

First, further study should avoid the limitations of the current study. In order to the increase the external validity, it is highly recommended that researchers conduct replication across different courses, grade levels, students' course tasks, all with random subject selection and assignment to groups. In addition, to minimize selection bias, randomly selecting experimental and comparison groups in the same semester is also recommended. The strategy of sampling and selection change in design would logically improve or help eliminate the threats to external and internal validity and make the findings more generalizable to similar situations.

Second, the design of this study was based on descriptive and inferential statistics to provide quantitative evidence to explain study findings for the effectiveness of peer

assessment skill training. For the current study, it cannot completely explain how and why peer-assessment skill training positively impacts students' writing, students' peer-assessment skills, quality of student generated scores, and students' satisfaction with peer assessment. Incorporating qualitative methods in conjunction with the quantitative methods would allow for the collection of richer, potentially more descriptive data regarding the effects of peer-assessment skill training.

Third, this study found that students who had peer-assessment training outperformed their counterparts who did not receive peer-assessment training in writing, and also found that students who had peer-assessment training had higher peer-feedback scores than those how had not received peer-assessment skills training. However, it is not clear what length of peer-assessment skill training time can maximally improve students' writing performance and assessment skills. Sluijsmans et al., (2002) suggested, more than four hours training section would better improve students' subject products and assessment skills. Therefore, different length of training section groups and pre-test and post-test design is recommended for future research. These findings would provide valuable information to assist teachers in determining a length for peer-assessment skill training that would maximally benefit students.

Fourth, since the current study did not investigate students' satisfaction with peer-assessment skill training, the author does not know specific information about students' satisfaction with the training, such as the length of period of training and contents of training. Students' satisfaction with training in further research would be highly recommended in order to ascertain what changes in training are needed to provide students maximal benefit from peer-assessment skill training.

Fifth, this study invested the effects of peer-assessment skill training on the quality of student-generated assessment scores. The results showed that training had positive effects on reliability of students' second round assessment scores, but had no positive effects on the validity of students-generated assessment scores. The author hypothesized that these variables -- assessment criteria or rubric and scoring system, students' understanding of criteria, an anonymous assessment condition, and multiple raters -- may prominently impact on the validity of students-generated assessment scores. Further research may consider incorporating these variables in peer-assessment skill training study to find out which variables would most impact the validity of students-generated scores.

A final issue that deserves further research is the relationship between content knowledge and assessing skills. Specifically, the issue concerns whether students' subject knowledge affects the quality of students' assessment. As a complex skill, peer assessment can only be demonstrated inside a particular subject matter domain (Sluijsmans et al., 2002). According to findings from the Student Satisfaction with Peer Assessment Questionnaire, some students (25%) responded that their peers had inadequate knowledge to evaluate their Wiki articles in the groups that had structured peer-assessment skill training. Sluijsmans et al., (2002) also found that a student reported this issue by saying that he had problems with assessing a peer's product that is of a higher level than his own product. Then he realized that he does not have enough domain knowledge to criticize his peer's product. It is plausible that the more subject expertise a student has, the more capable he or she is of assessing their peers' work. Therefore,

further research should explore whether a student's subject knowledge affects their performance in peer assessment or the development of their assessment skills.

Conclusions

The current study examined the effects of peer-assessment skill training on university students' writing performance, peer assessment skill, the quality of student-generated assessment scores, and students' satisfaction with peer assessment in an online environment. The findings from this study showed that students who had peer-assessment training in the experimental groups had higher final writing scores than those who did not receive peer-assessment training in the comparison group. The findings of the study also revealed that students who had peer-assessment training in the experimental groups received higher peer feedback scores than those who did not receive structured peer-assessment training in the comparison group. Additionally, the findings showed that students who had structured peer-assessment training generated more reliable assessment scores than those who did not receive structured peer-assessment training in second round peer assessment. Next, the current study's findings demonstrated that students who had target-criteria-based peer-assessment training exhibited higher levels of satisfaction with peer assessment feedback than those who had principle-based peer-assessment training but did not receive peer-assessment training. Furthermore, the findings indicated that the majority of students were actively engaged in peer assessment and benefited from playing the role of both assessor's and assessee's. Despite the fact that peer-assessment skill training has positive effects upon students' writing performance, students' assessment skills, and the quality of peer assessment, data analysis shows that the validity of student-generated scores in experimental groups was not significantly higher than that in the

comparison group. This factor invites more consideration about how to improve the quality of training and peer assessment procedures. Finally, the findings of current study provide evidence that student-generated assessment scores are sufficiently reliable and valid from instructor's perspective and can be used to augment instructor's assessment in the classroom.

FEFERENCES

Bandura, A. (1997). *Self-efficacy: The exercise of control*. New York: W. H. Freeman.

Berg, I. V. D., Admiraal, W., & Pilot, A. (2006). Peer assessment in university teaching: evaluating seven course designs. *Assessment and Evaluation in Higher Education, 31*(1), 19-36.

Birkeland, T. S. (1986). *Variations of feedback related to teaching paragraph structure to technicians*. Retrieved January 7, 2008, from http://proquest.umi.com/pqdweb?index=0&did=752276871&SrchMode=1$sid=3&Fmt=2.

Black, P. & Wiliam, D. (1998). Assessment and classroom learning, *Assessment in Education, 5*(1), 7-74.

Bhalerao, A., & Ward, A. (2000). Towards electronically assisted peer assessment: A case study. Retrieved January 19, 2008, from Warwick University, Department of Computer Science Web site: http://www.dcs.warwick.ac.uk/~abhir/research/papers/new-alt-j.pdf

Bloxham, S., & West, A. (2004). Understanding the rules of the game: marking peer assessment as a medium for developing students' conceptions of assessment. *Assessment and Evaluation in Higher Education, 29*(6), 721-733.

Boud, D. (1990). Assessment and the promotion of academic values. *Studies in Higher Education, 15*, 101-111.

Boud, D. (1995). *Enhancing learning through self-assessment*. London: Kogan Page.

Boud, D., Cohen, R. & Sampson, J. (1999). Peer learning and assessment, *Assessment and Evaluation in Higher Education, 24*(4), 413-426.

Brindley, C., & Scoffield, S. (1998). Peer assessment in undergraduate programmes. *Teaching in higher education, 3(1), 79-89.*

Brown, S., & Knight, P. (1994). *Assessing learning in higher education.* London: Kogan Page.

Bump, J. (1990). Radical changes in class discussion using networked computers. *Computers and the Humanities, 24,* 49-65.

Butler, D. L., & Winne, P. H. (1995). Feedback and self-regulated learning: A theoretical synthesis. *Review of Educational Research, 65,* 245-281.

Calado, S. M. R. (1994). The impact of system design factors on users' reaction to peer feedback. *Dissertation Abstracts International, 56,* 1428.

Campbell, K., & Zhao, Y. (1996). Refining knowledge in a virtual community of learners. *Journal of Technology and Teacher Education, 4*(3/4), 263-280.

Carson, J. G., & Nelson, G. L. (1996). Chinese students' perception of ESL peer response group interaction. *Jounal of Second Language Writing, 5(1),* 1-19.

Chaudron, C. (1983). *Evaluating writing: Effects of feedback on revision.* Paper presented at the 17th Annual TESOL Convention, Toronto, Ontario, Canada.

Chang, K., Chen, I., & Sung, Y. (2002). The effect of concept mapping to enhance text comprehension and summarization. *The Journal of Experimental Education 71*(1), 5-23.

Cheng, W. & Warren, M. (1997). Having second thoughts: Student perceptions before and after a peer assessment exercise. *Studies in Higher Educatrion, 22*(2), 233-239)

Cheng, W. & Warren, M. (1999). Peer and teacher assessment of the oral and written tasks of a group project. *Assessment and Evaluation in Higher Education, 24*(3), 301-314.

Chi, M. T. H. (1996). Constructing self-explanations and scaffolded explanations in

tutoring. *Applied Cognitive Psychology, 10*, S33-S49).

Cho, K., Schunn, C. D., & Wilson, W. (2006). Validity and rliability of scaffolded peer assessment of writing from instructor and student perspectives. *Journal of Educational Psychology, 98*(4), 891-901.

Csikszentmihalyi, M. (1978). Instrinsic rewards and emergent motivation. In M. R. Lepper & D. Greene (Eds.), *The hidden costs of reward* (pp. 205-216). Hillsdale, NJ: Lawrence Erlbaum Associates.

Davies, P. (2006). Peer assessment: judging the quality of students' work by comments rather than marks. *Innovations in Education and Teaching International, 43*(1), 69-82.

Dierick, S & Dochy, F. (2001). New lines in edumetrics: new forms of assessment lead to new assessment criteria, *Studies in Educational Evaluation,27*, 307-329.

Dweck, C. S. (1975). The role of expectations and attributions in the alleviation of learned helplessness. *Journal of Personality and Social Psychology, 31*, 674-685.

Eisenberg, M. B., & Ely, D. P. (1993). Plugging into the Net. ERIC review, 2(3), 2-10.

Fagot, R. (1991). Reliability of ratings for multiple judges: intraclass correlation and metric sales. *Applied Psychological Measurement, 15*(1), 1-11.

Falchikov, N. (1986). Product comparisons and process benefits of collaborative self and peer group assessment. *Assessment and Evaluation in Higher Education, 11*(2), 146-166.

Falchikov, N. (1995a). Peer feedback marking: developing peer assessment. *Innovations in Education and Training International, 32*(2), 175-187.

Falchikov, N. (1995b). Improving feedback to and from students. *Assessment for*

Learning in Higher education, 1, 157-166.

Falchikov, N. & Goldfinch, J. (2000). Student peer assessment in higher education: A meta-analysis comparing peer and teacher marks. *Review of Educational Research, 70*(3), 287-322.

Fagot, R. (1991). Reliability of ratings for multiple judges: intraclass correlation and metric scales. *Applied Psychological Measurement, 15*(1), 1-11.

Farh, J. L., Cannella, A. A., & Bedeian, A. G. (1991). The impact of purpose on rating quality and user acceptance. *Group and Organization Studies, 16*(4), 357-386.

Ferguson, G. A. (1966). *Statistical analysis in psychology and education* (2nd ed.). New York: McGraw Hill.

Ferris, S., & Wilder, H. (2006). *Uses and potentials of Wikis in the classroom.* Retrieved August 2, 2008, from

http://www.inovateonline.info/index.php?view=article&id=258.

Ford, B. W. (1973). The effects of peer editing/grading on the grammar-usage and theme-composition ability of college freshmen. *Dissertation Abstracts International, 33,* 6687.

Ghorpade, J., & Lackritz, J. R. (2001). Peer evaluation in the classroom: A check for sex and race/ethnicity effects. *Journal of Education for Business, 76*(5), 274-282.

Graesser, A. C., Peason, N., K., & Magliano, J. P. (1995). Collaborative dialogue patterns in naturalistic one-to-one tutoring. *Applied Cognitive Psychology, 9,* 495-522.

Guilford, W. H. (2001). Teaching peer review and the process of scientific writing. *Innovations and Ideas, 25*(3), 167-175.

Haaga, D. A. F. (1993). Peer review of term papers in graduate psychology course.

Teaching of Psychology, 20(1), 28-32.

Hanrahan, S. & Isaacs, G. (2001). Assessing self- and peer assessment: the students'

views, *Higher Education Research and Development, 20*(1), 53-70.

Hunter, D., & Russ, M. (1995). Peer assessment in performance studies. In S. Griffiths, K.

Houston, & A. Lazenblatt (Eds.), *Enhancing student learning through peer*

tutoring in higher education: Section3 – Implementing (Vol. 1, pp. 63-65).

Coleraine, Northern Ireland: University of Ulster.

Kafai, Y., & Resnick, M. (1996). *Constructionism in practice: designing, thinking, and*

learning in a digital world. Mahwah, NJ: Lawrence Erlbaum Associates.

Keller, J. M. (1987). Strategies for stimulating the motivation to learn. *Performance*

and Instruction Journal, 1-8.

Kerr, N. L., & Bruun, S. E. (1983). Dispensability of member effort and group motivation

losses: free-rider effects. *Journal of Personality and Social Psychology, 44*(1), 78-

94. Kerr, P. M., Park, K. H., & Domazlicky, B. R. (1995). Peer grading of essays

in a principles of microeconomics course. *Journal of Education for Business,*

July/August, 356-361.

Kerr, P. M., Park, K. H., & Domazlicky, B. R. (1995). Peer grading of essays in a

Principles of microeconomics course. *Journal of Education for Business,*

July/August, 356-361.

Lasater, C. A. (1994). Observation feedback and analysis of teaching practice: Case

studies of early childhood student teachers as peer tutors during a preservice

teaching practicum. *Dissertation Abstracts International, 55*, 1916.

Latane, B., Williams, K., & Harkins, S. (1979). Many hands make light work: the causes

and consequences of social loafing. Journal of Personality and Social Psychology, 37(6), 822-833.

Lave, J., & Wenger, E. (1991). *Situated learning: Legitemate peripheral participation.* New York: Cambridge University Press.

Leidner, D. E., & Jarvenpaa, S. L. (1995). The use of information technology to enhance management school education: A theoretical view. *MIS Quarterly, 19*(3), 265-291.

Li, L. K.(2001). Some refinements on peer assessment of group projects. *Assessment & Evaluation in Higher Education. 26,* 6-18.

Li, L., & Steckelberg, A. (2004). *Using peer feedback to enhance student meaningful learning.* Chicago, IL: Association for Educational Communications and Technology. (ERIC Document Reproduction Service No. ED48511).

Lin, S. S. J. Liu, E. Z. F., & Yuan, S. M (2001). Web-based peer assessment: feedback for students with various thinking-styles. *Journal of Computer Assisted Learning, 17,* 420-432.

LINRO, (2005), *Peer Review Definition.* Retrieved August 5, 2008, from http://www.linfo.org/peer_review.html.

Liu, E. Z., Lin, S. S., Chiu, C. H., & Yuan, S. M. (2001). Web-based peer review: The learner as both adapter and reviewer. *IEEE Transactions on Education, 44(3),* 246-251.

Lopez-Real, F. & Chan, Y. R. (1999). Peer assessment of a group project in a primary mathematics education course, *Assessment & Evaluation in Higher Education, 24,* 67-79.

Liu, N. F. (2005). *Hong Kong academics' and students' perceptions of assessment*

purposes and practices. Learning Oriented Assessment Project report, Hong Kong Institute of Education.

Lu, R. L. (2005). *A comparison of anonymous e-peer review versus identifiable e-peer review on college student writing performance and learning satisfaction.* Retrieved October 2, 2007, from http://proquest.umi.com/pqdweb?index=0&did=920928121&SrchMode =1&sid=1&Fmt=6&VInst=PROD&VType=PQD&RQT=309&VName=PQD&T S=1206972687&clientId=3505.

Lu, R. L. & Bol, L. (2007). A comparison of anonymous versus identifiable e-peer review on college student writing performance and the extent of critical feedback. *Journal of Interactive Online Learning, 6*(2), 100-115.

Johnson, R. (1999). Helping students to understand and use assessment for learning. In S. Avery, C. Bryan & G. Wisker (Eds.), *Innovations in teaching English and textual studies,* 181-197.

Mabrito, M. (1991). Electronic mail as a vehicle for peer response. *Written Communication, 8*(4), 509-532.

MacAlpinil, J. M. (1999) Improving and encouraging peer assessment of student presentations. *Assessment & Evaluation in Higher Education, 24* (1), 15-25.

MacLeod, L. (1999). Computer-aided peer review of writing. *Business Communication Quarterly, 62*(3), 87.

Magin, D. (1993). Should student peer ratings be used as part of summative assessment? *Higher Education Research and Development, 16,* 537-542.

Mangelsdorf, K. (1992). Peer reviews in the ESL composition classroom: What do the students think? *ELT Journal, 46*(3), 275-285.

Marcoulides, G. A., & Simkin, M. G. (1995). The consistency of peer review in student

writing projects. *Journal of Education for Business, 70,* 220-223.

McDowell, L. (1995). The impact of innovative assessment on student learning.

Innovations in Education and Training International, 32, 302-313.

McIsasc, C. M. & Sepe, J. F. (1996). Improving the writing of accounting students: A

cooperative venture. *Journal of Accounting Education, 14*(4), 515-533.

Mehrens, W. A., Popham, W. J. & Pyan, J. M. (1998). How to prepare students for

performance assessments. *Educational Measurement: Issues and Practice, 17*(1),

18-22.

Miller, P. J. (2003). The effect of scoring criteria specificity on peer and self-assessment.

Assessment & Evaluation in Higher Education, 28, 383-394.

Mowl, G., & Pain, R. (1995). Using self and peer assessment to improve students' essay

Writing – A case study from geography. *Innovations in Education and Training

International, 32*(4), 324-335.

Nicol, D. J., & Macfarlane-Dick, (2006). Formative assessment and self-regulated

learning: a model and seven principles of good feedback practice. *Studies in

Higher Education, 31*(2), 199-218.

Nilan, K. J (1983). Astudy of recipient reactions to receiving developmental feedback

from a peer group of co-trainees. *Dissertation Abstracts International, 44,* 29-30.

Nilson, L. B. (2003). Improving student peer feedback. *College Teaching, 51*(1), 34-39.

Papert, S. (1993). *The children's machine: Rethinking school in the age of the computer.*

New York: Basic Books.

Paulus, T. M. (1999). The effect of peer and teacher feedback on student writing. *Journal*

of Second Language Writing, 8(3), 265-289.

Piagrt, J. (1969). *Science of education and the psychology of the child.* New York: Norton.

Pintrich, P. R. (1995). *Understanding self-regulated learning.* San Francisco, CA.

Jossey-Bass.

Pintrich, P. R. & Zusho, A. (2002). Student motivation and self-regulated learning in the

college classroom, in: J. C. Smart & W. G. Tierney (Eds) *Higher Education:*

handbook of theory and research (vol. XVII) (New York, Agathon Press).

Plutsky, S., & Wilson, B. A. (2004). Comparison of the three methods for teaching and

evaluating writing: A quasi-experimental study. *The Delta Pi Epsilon Journal,*

46(1), 50- 61.

Prins, F. J., Sluijsmans, D. M. A., Kirschner, P. A., & Strijbos, J. (2005). Formative peer

assessment in a CSCL environment: a case study, *Assessment & Evaluation in*

Higher Education, 30 (4), 417-444.

Oldfield, K. A. & MacAlpine, J. M. (1995). Peer and self-assessment at the tertiary level

– an experimental report, *Assessment & Evaluation in Higher Education, 20,* 125-

132.

Orsmond. P., Merry. S. & Reiling, K (1996). The importance of marking criteria in the

use of peer self-assessment. *Assessment & Evaluation in Higher Education, 20,*

239-250.

Orsmond, P., Merry, S. & Reiling, K. (2000). The use of student derived making criteria

in peer and self-assessment. *Assessment and Evaluation in High Education, 25*(1),

23-38.

Orsmond, P., Merry, S., & Reiling, K. (2002). The use of examplars and formative

feedback when using student derived marking criteria in peer and self-assessment. *Assessment and Evaluation in Higher Education, 27*(4), 309-323.

O'Shea, P. M., Baker, P. B., Allen, D. W. Curry-Corcoran, D. E., & Allen, D. B., (2007). New levels of student participatory learning: A WikiText for the introductory course in education. *Journal of Interactive Online Learning, 6*(3), 228-235.

Quible, Z. K. (1997). The efficacy of several writing feedback system. *Business Communication Quarterly, 60*(2), 109-124.

Ramsey, P. G., Wenrich, M. D., Carline, J. D., Inui, T. S., Larson, E, B., & Logerfo, J. P. (199) Feasibility of hospital-based use of peer ratings to evaluate the performances of practicing physicians. *Academic Medicine, 71,* 364-370.

Raymond, E. (2000). Cognitive Characteristics. *Learners with Mild Disabilities* (pp. 169-201) Needham Heights, MA: Allyn & Bacon, A Pearson Education Company.

Richardson, W. (2006). *Blogs, wikis, podcasts, and other powerful web tools for classrooms.* Thousand Oaks, CA: Corwin Press.

Richer, D. L. (1992). The effects of two feedback systems on first year college students' writing proficiency, *Dissertation Abstracts International,* 53, p. 2722. Richardson, W. (2006). *Blogs, wikis, podcasts, and other powerful web tools for classrooms.* Thousand Oaks, CA: Corwin Press.

Robinson, J. M. (1999). Anonymous peer review for classroom use: Results of a pilot project in a large science unit. *Teaching and Learning Forum 99.* Sadler, D. R. (1989). Formative assessment and the design of instructional systems, *Instructional Science, 18,* 119-144.

Sadler, D. R. (1989). Formative assessment: revisiting the territory, *Assessment in Education, 5* (1), 77-84.

Saito, H. & Fujita, T. (2004). Characteristics and user acceptance of peer rating in EFL writing classrooms. *Language Teaching Research 8*(1), 31-54. Stefani, L. A. J. (1994). Peer, self and tutor assessment: relative reliabilities. *Studies in Higher Education, 19*(1), 69-75.

Schaffer, J. (1996). Peer response that works, *Journal of Teaching Writing, 15*(1), 81-89.

Schrunk, D. H., & Zimmerman, B. J. (1994). Self-regulation in education: Retrospect and prospect. In D. H. Schrunk & B. J. Zimmerman (Eds.), *Self-regulation of learning and performance*. Hillsdale, NJ: Erlbaum.

Schrunk, D. H., & Zimmerman, B. J. (2008). *Motivation and self-regulated learning: theory, research, and applications*. New York: Lawrence Erlbaum Associates Taylor & Francis Group.

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

Smith, D, Campbell, J & Brooker, R. (1999). The impact of students' approaches to essay writing on the quality of their essays. *Assessment and Evaluation in Higher Education, 24*(3), 327-338.

Sluijamans, D., Dochy, F & Moerkerke, G. (1999). Creating a learning environment by suing self-peer- and co-assessment, *Learning Environments Research, 1*(3), 293-319.

Sluijsmans, D., Moerkerke, G., Dochy, F. & Van Merrienboer, J. J. G. (2001). Peer

assessment in problem based learning, *Studies in Educational Evaluation, 27*(2), 153-173.

Sluijsmans, D., Brand-Gruwel, S., & Van Merriënboer, J. J. G. (2002). Peer assessment training in teacher education: Effects on performance and perceptions. *Assessment and Education in Higher Education, 27*(5), 443-454.

Sluijsmans, D., Brand-Gruwel, S., Van Merrienboer, J. J. G., & Martens, R. L. (2004). Training teachers in peer-assessment skills: effects on performance and perceptions. *Innovations in Education and Teaching International, 41*(1), 59-78.

Stefani, L. A. J. (1994). Peer, self and tutor assessment: relative reliabilities. *Studies in Higher Education, 19*(1), 69-75.

Stefani, L. A. J. (1998). Assessment in partnership with learners. *Assessment and Evaluation in Higher Education, 23*(4), 339-350.

Tillema, H. H., Kessels, J. W. M. & Meijers, F. (2000) Competencies a building blocks for integrating assessment with instruction in vocational education: a case from the Netherlands, *Assessment and Evaluation in Higher Education, 25*(3), 265-278.

Topping, K. (1998). Peer assessment between students in colleges and universities. *Review of Educational Research, 68,* 249-276.

Topping, K., & Ehly, S. (1998). *Peer assisted learning.* Mahwah, NJ: Lawrence Erlbaum Associates.

Topping, K., Smith, F. F., Swanson, I., & Elliot, A. (2000). Formative peer assessment of academic writing between postgraduate students. *Assessment and Evaluation in Higher Education, 25*(2), 149-169.

Van Lehn, K. A., Chi, M. T. H., Baggett, W., & Murray, R. C. (1995). *Progress report:*

Towards a theory of learning during tutoring. Pittsburgh, PA: Learning Research

and Development Center, University of Pittsburgh.

Venables, A. & Summit, R. (2003). Enhancing scientific essay writing using peer

assessment. *Innovations in Education and Teaching International, 40*(3), 281-290.

Vygotsky, L. S. (1971). *The psychology of art.* Cambridge, MA: MIT Press.

"Wikibooks." (2008). *Encyclopedia.* Retrieved January 21, 2008 from

http://www.reference.com/search?q=Wikibook.

Woolf, H. (2004). Assessment criteria: reflections on current practices. *Assessment and*

Evaluation in Higher Education, 29(4), 479-493.

Xiao, Y. & Lucking, R. (2008). The impact of two types of peer assessment on students'

performance and satisfaction within a Wiki environment. *Internet and Higher*

Education, 11(3-4), 186-193.

Zariski, A. (1996). Student peer assessment in tertiary education: Promise, perils and

practice. *Teaching and Learning Within and Across Disciplines,* 1890103.

Zimmerman, B. J. & Schunk, D. H. (2001). *Self-regulated learning and academic*

achievement: theoretical perspectives. Mahwah, NJ, Lawrence Erlbaum

Associates.

Zhao, Y. (1998). The effects of anonymity on computer-mediated peer review.

International Journal of Educational Telecommunication, 4(4), 311-345.

Zhu W. (1994). Effects of training for peer revision in college freshman composition

classes. *Dissertation Abstracts International, 55,* 209.

Zhu, W. (1995). Effects of training for peer response on students' comments and

interaction. *Written Communication, 12*(4), 492-528.

APPENDICES

Appendix A: Wiki Article Evaluation Rubric

	No Importance 1	Little Importance 2	Some Importance 3	Quite Important 4	Very Important 5
How importan t was the information presented on this topic to you as a teacher education student?	Information is not relevant to future teachers. Information is entirely the author's opinion. The information is obsolete. Only irrelevant details; or only common knowledge	One useful point A few facts but mostly the author's opinion. Most of the information is irrelevant in today's schools. Focused on unimportant subtopics OR overly general with few specifics	A couple useful points; some irrelevant information About half of the information is the author's opinion. Some out-dated information; may not reflect current practice Provided good information but an incomplete understandin g of the topic or key issues	Includes ideas relevant to future teachers Mostly based on researched information. Applicable to today's students Provided a good general overview OR effective discussion of key ideas	Covers key ideas crucial for future teachers to know Based on researched information. Highly relevant to today's students Provides a general overview & specific details; leads to a good understandin g of the topic
	Of no interest 1	Of little interest 2	Reasonably Interesting 3	Quite Interesting 4	Extremely Interesting 5
	No side bar included.	Sidebar repeats what is already in the article	Sidebar includes new information related to the topic.	Sidebar includes new information that enhances understandin	Sidebar includes new information I was motivated to

				g of the topic	read/view
How interesting was the article to read?	No perspective is given. Nothing new No elaboration or explanation included No clear points	Only the typical view is presented. One new idea or bit of information Information presented without analysis or interpretation At least one clear point	One interesting perspective is presented A couple new ideas or pieces of information Some interpretation / analysis included Some clear points are made.	At least two perspectives were presented Some new information/ideas Reasonable interpretation / analysis makes the information easier to understand Points are explained with examples	Multiple perspectives are included Mostly new information/ideas Insightful interpretation / analysis – inspires a new perspective on the issue. Points are made with compelling examples
	Not credible 1	Limited credibility 2	Reasonable credibility 3	Substantial credibility 4	Completely credible 5
How credible do you think the information is?	Missing two or more sources OR sources used but not cited or listed. Most sources are untrustworthy or biased. All sources and information reflect a single viewpoint	Missing a source OR 5 sources are used but not cited and/or included in a reference list. Some sources are untrustworthy or biased. Sources lack diversity OR information from divergent	Required sources are included; APA format is not used or has many errors. Sources are reasonable. A variety of sources listed but the information primarily reflects a single viewpoint	Required sources are included; a couple of formatting errors Sources are reasonably reputable and current. Information from a good variety of sources is included It is clear	Required sources are properly cited and included in a reference list in APA format. Sources are reputable, minimally biased and current. Information from diverse sources representing

| | The entire article is biased and opinion based without acknowledgment of this perspective. | sources is only superficially mentioned

Author routinely states her opinion as fact, ignores own biased | Sometimes the author forgets and states his own opinion as fact. | when the author is presenting his own opinion or perspective; he doesn't try to pass if off as fact. | multiple perspectives is included

Author clearly indentifies his own ideas, biases and opinions |
|---|---|---|---|---|---|
| | Poorly written

1 | Fairly written

2 | Reasonably written

3 | Well written

4 | Excellently written

5 |
| How well do you think this article was **written?** | Questions are missing or not multiple-choice.

Learning target is missing or unrelated to content

Nothing in the article grabs the reader's attention

Article is entirely "pasted" together from other sources.

Poor organization, sentence structure and/or grammatical errors made | 4 multiple-choice questions are included.

Learning targets generally related to the content are stated

At least one part of the article is interesting

Cited information is "pasted" into the article with little to no explanation.

The organization was difficult to follow, sentences were | Multiple-choice questions (2 application & 2 knowledge) asses key points.

Reasonable learning targets are stated; the content relates to these goals

Parts of the article capture attention

Some of the cited information is discussed

A few areas were hard to follow, confusing or oddly | Multiple-choice questions (2 application &2 knowledge) align with the learning targets, and asses key points.

Specific and reasonable learning targets are stated; the content aligns with these goals

Captures attention initially and periodically throughout

Cited information is discussed or explained. | Multiple-choice questions (2 application & 2 knowledge) align with the learning targets, asses key points, and are written according to the guidelines by Bothell (see R4)

Specific, appropriate and observable learning targets are stated; the content is clearly organized to help the reader achieve these goals |

	it very difficult to understand the content.	awkward and/or there were several distracting mechanical errors.	organized. There were a few distracting errors.	The article flowed pretty well and there were just a few mechanical errors.	Captures and maintains attention throughout Cited information is introduced, elaborated on and explained Writing is organized, easy to read, and contains few to no mechanical errors.
Total					

Appendix B: The Rubric of Quality of Feedback

Assessment Criterion	Need Improvement (1-3 points)	Average (4 points)	Above Average (5-6 points)
Includes All Components	Feedback includes less than two compliments on the article's strengths, or less than two suggestions for the article's improvement, and less than 150 words of overall feedback.	Feedback includes two compliments on the article's strengths, two suggestions for the article's improvement, and no less than 150 words of overall feedback.	Feedback includes more than two compliments on the article's strengths, and more than two suggestions for the article's improvement, and more than 150 words of overall feedback.
Reflects Standards or Criteria	Feedback is unrelated to the criteria listed in the rubric	Feedback generally relates to the criteria listed in the rubric	Feedback specifically describes how the article meets or does not meet the criteria listed in the rubric
Understandable	The feedback is vague, confusing or difficult to understand	Some of the feedback is clear; some is vague, confusing or difficult to understand	Feedback is written in a clear and direct manner that is easy to understand
Specific	The feedback is general; it does not reference specific passages of the text as strong or weak points	At least one specific passage in the text is noted as a strength or weaknesses; much of the feedback is general	Many specific passages in the text are referred to as strengths and/or weaknesses
Achievable	Feedback contains no specific suggestions to achieve or contains suggestions that are unrealistic given the capabilities of the author/the time allotted for revision	The feedback contains at least one reasonable suggestions; may also contain suggestions that are unrealistic given the capabilities of the author/the time allotted for revision	The feedback is positively stated and includes several suggestions that can be reasonably made by the author to improve the quality of the article

Appendix C: Peer Assessment Satisfaction Questionnaire

Directions: The following questions ask about your satisfaction with peer assessment. **Remember there are no right or wrong answers, just answer as accurately as possible.** SD = Strongly disagree, D = Disagree, A = Agree, and SA = Strongly agree. Use the scale below to answer the questions. If you strongly agree with the statement, circle 4; if you strongly disagree with the statement, circle 1. Try to find the appropriate degree of agreement for your each answer.

	Satisfaction with the Assessment System	SD	D	A	SA
1	The peer feedback/rating system motivated me to do my best work.	1	2	3	4
2	The peer feedback/rating system is appropriate for the Wikibook article project.	1	2	3	4
3	The peer feedback/rating system created a learning environment in which I felt comfortable.	1	2	3	4
4	*The peer feedback/rating system was too demanding.	1	2	3	4
5	The peer feedback/rating system made me feel responsible for my own learning.	1	2	3	4
6	The peer feedback/rating system made me feel responsible for others' learning.	1	2	3	4
7	It is easy for me to complete my feedback/rating assignments.	1	2	3	4
8	I give my peers feedback/ratings by the stated due dates.	1	2	3	4
9	My peers give me feedback/ratings by the stated due dates.	1	2	3	4
10	The use of Wikibook technology for doing peer assessment is efficient with regard to the overall class structure.	1	2	3	4
11	*Too much learning time was spent doing peer feedback/rating activities.	1	2	3	4
12	I feel confident in my ability to evaluate others' work during the peer feedback/rating activities.	1	2	3	4
13	I feel confident in my ability to evaluate my own work during peer feedback/rating activities.	1	2	3	4
14	The technology behind the feedback/rating system allows me to rate and provide feedback to my peers' articles quickly.	1	2	3	4
15	The Wikibooks technology allowed the feedback/rating process to be helpful in improving the textbook's quality.	1	2	3	4

	Satisfaction with Peer Feedback				
16	I enjoy giving peer feedback.	1	2	3	4
17	I enjoy receiving peer feedback.	1	2	3	4
18	I believe that it is important for me to learn how to give	1	2	3	4

feedback.

19	I believe that it is important for me to learn how to take feedback.	1	2	3	4
20	Giving feedback is an effective approach to improve my critical thinking skills.	1	2	3	4
21	Taking feedback is an effective approach to improve my critical thinking skills.	1	2	3	4
22	I'm satisfied with the overall quality of the feedback I've received.	1	2	3	4
23	I'm satisfied with the overall quality of the feedback I've given.	1	2	3	4
24	My peers provided sufficient amount of feedback on my Wikibook article.	1	2	3	4
25	The peer feedback I received was helpful to improve my Wikibook article.	1	2	3	4
26	Peers have adequate knowledge to evaluate my Wikibook article.	1	2	3	4
27	I have benefited from rating to my peers' Wikibook articles.	1	2	3	4
28	I have benefited from providing feedback to my peers' Wikibook articles.	1	2	3	4
29	I think I have learned more from peers' feedback than from the instructors' feedback.	1	2	3	4
30	The average rating scores my peers gave me accurately reflected the overall quality of my work.	1	2	3	4
31	*I felt reluctant to give negative feedback to my classmates.	1	2	3	4