The Pennsylvania State University
The Graduate School
College of Education

EFFECT OF QUALITY CHARACTERISTICS OF PEER RATER ON THE

RELIABILITY AND VALIDITY OF PEER ASSESSMENT

A Dissertation in
Educational Psychology

by

Xiuyan Guo

© 2017 Xiuyan Guo

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy,

August 2017
The dissertation of Xiuyan Guo was reviewed and approved* by the following:

Hoi K. Suen  
Distinguished Professor of Education  
Dissertation Advisor  
Chair of Committee

Pui-Wa Lei  
Professor of Education

Mindy L. Kornhaber  
Associate Professor of Education

Mosuk Chow  
Senior Scientist and Professor of Statistics

Peggy Van Meter  
Associate Professor of Education,  
Professor-In-Charge, Graduate Program in Educational Psychology

*Signatures are on file in the Graduate School
ABSTRACT

There is a large body of research on influential factors of the reliability and validity of peer assessment scores. Far less research has been done on the effect of peer raters’ quality characteristics. This study investigated the effects of the following three characteristics of peer rater on the reliability and validity of peer assessment: knowledge of course content measured by the assessment, rater training experience, and rater motivation in making responsible judgments. A quasi-experimental, three-way completely crossed factorial design was used to analyze data from 838 students in 24 classes in a typical Chinese high school. The students rated four well-selected essays written by their peers using the standard Gaokao (National College Entrance Examination) Chinese Writing Rubric. In this study, knowledge of course content is measured by an exam score in the assessed domain. Training experience is measured by (1) the number of times a student has previously participated in a peer assessment, and (2) whether the student received self-paced training in this study. Motivation of the peer rater is measured by (1) whether they were offered monetary incentives, and (2) whether they have been told about the general benefits of completing peer assessment seriously before conducting the peer assessment in this study. It was hypothesized that peer raters with high quality characteristics would provide more reliable and valid scores than their counterparts. All the data were be analyzed using a series of two-level, hierarchical linear models.

The findings confirmed the hypothesis, and yielded clear evidence that peer raters with more knowledge in the assessed domain, more peer rating experience, or higher rating motivation can provide more reliable and more valid scores than those with less knowledge in assessed domain, less peer rating experience, or lower rating motivation. This study also found significant interaction effects between peer raters’ content knowledge and rating motivation, and rating
experience and rating motivation. Implications for in person and online peer assessments are discussed, as well as suggestions for further research.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>List of Figures</td>
<td>vi</td>
</tr>
<tr>
<td>List of Tables</td>
<td>vii</td>
</tr>
<tr>
<td>Acknowledgements</td>
<td>viii</td>
</tr>
<tr>
<td>Chapter 1 Introduction</td>
<td>1</td>
</tr>
<tr>
<td>1.1. Background</td>
<td>1</td>
</tr>
<tr>
<td>1.2. Statement of the problem</td>
<td>5</td>
</tr>
<tr>
<td>1.3. Present Study</td>
<td>7</td>
</tr>
<tr>
<td>Chapter 2 Review of Literature: Peer assessment, reliability and validity</td>
<td>10</td>
</tr>
<tr>
<td>2.1. Understanding peer assessment</td>
<td>10</td>
</tr>
<tr>
<td>2.1.1. Theoretical foundation of peer assessment</td>
<td>10</td>
</tr>
<tr>
<td>2.1.2. Desirable circumstance for peer assessment</td>
<td>13</td>
</tr>
<tr>
<td>2.1.3. Problems with peer assessment</td>
<td>23</td>
</tr>
<tr>
<td>2.2. Reliability and validity of peer assessment</td>
<td>26</td>
</tr>
<tr>
<td>2.2.1. Empirical evidence for reliability and validity of peer assessment</td>
<td>27</td>
</tr>
<tr>
<td>2.2.2. Factors that influence reliability and validity of peer assessment</td>
<td>36</td>
</tr>
<tr>
<td>Chapter 3 Review of the literature: Peer rater quality characteristics</td>
<td>41</td>
</tr>
<tr>
<td>3.1. Effect of peer rater’s content knowledge in assessment domain</td>
<td>42</td>
</tr>
<tr>
<td>3.2. Effect of peer rater’s training experience</td>
<td>46</td>
</tr>
<tr>
<td>3.3. Effect of peer rater’s rating motivation</td>
<td>51</td>
</tr>
<tr>
<td>Chapter 4 Methodology</td>
<td>59</td>
</tr>
<tr>
<td>4.1. Research questions and hypotheses</td>
<td>59</td>
</tr>
<tr>
<td>4.2. Research design</td>
<td>65</td>
</tr>
<tr>
<td>4.3. Participants</td>
<td>67</td>
</tr>
<tr>
<td>4.4. Procedure</td>
<td>68</td>
</tr>
<tr>
<td>4.4.1. High school Chinese language mid-term examination and teacher rating</td>
<td>68</td>
</tr>
<tr>
<td>4.4.2. Selection of training materials and final peer assessment essays</td>
<td>69</td>
</tr>
<tr>
<td>4.4.3. Peer rater training method</td>
<td>72</td>
</tr>
<tr>
<td>4.4.4. Motivation intervention</td>
<td>74</td>
</tr>
<tr>
<td>4.4.5. Peer assessment</td>
<td>75</td>
</tr>
<tr>
<td>4.5. Instrumentation and measures</td>
<td>76</td>
</tr>
<tr>
<td>4.5.1. Content knowledge in assessment domain</td>
<td>76</td>
</tr>
<tr>
<td>4.5.2. Training experience</td>
<td>77</td>
</tr>
<tr>
<td>4.5.3. Rating motivation</td>
<td>78</td>
</tr>
<tr>
<td>4.5.4. Dependent variables</td>
<td>79</td>
</tr>
<tr>
<td>4.6. Data analysis</td>
<td>84</td>
</tr>
<tr>
<td>Chapter 5 Results</td>
<td>90</td>
</tr>
</tbody>
</table>
5.1. Effect of peer rater’s quality characteristics on inter-rater reliability of peer assessment ................................................................. 95
5.2. Effect of peer rater’s quality characteristics on intra-rater reliability of peer assessment ................................................................................ 101
5.3. Effect of peer rater’s quality characteristics on rating validity of peer assessment .................................................................................. 107

Chapter 6 Discussion, Implications, Limitations and Future Research ................................................. 113

6.1. Discussion of the overall findings ................................................................................. 114
6.2. Implication for educational practice .......................................................................... 122
6.3. Limitations and future research .................................................................................. 125
   6.3.1. Sampling and generalizability of findings in this study ........................................ 125
   6.3.2. Treatment fidelity of peer rater training .............................................................. 126
   6.3.3. Motivation incentive methods in peer assessment ............................................ 127
   6.3.4. Peer rater’s demographic and psychological characteristics ............................. 128
   6.3.5. Other types of rater error .................................................................................. 129
   6.3.6. Identify high quality peer raters and improve the accuracy of peer assessment ...................................................................................... 129

References ................................................................................................................................ 131

Appendix A The Gaokao Chinese Writing Rubric ...................................................................... 144
Appendix B Original Chinese Version of Essay Topics and Sample Essays ................................. 148
Appendix C Original Verbal Persuasion Lecture in Chinese ...................................................... 158
Appendix D Practicing Scoring Sheet ...................................................................................... 159
Appendix E Peer Assessment Scoring Sheet ........................................................................... 161
Appendix F Motivation Questionnaire .................................................................................... 162
LIST OF FIGURES

Figure 4-1. Histogram of Participants’ Age Distribution .......................................................... 67
Figure 5-1. Boxplot of Inter-rater Error across 24 Classes ...................................................... 92
Figure 5-2. Boxplot of Intra-rater Error across 24 Classes ...................................................... 93
Figure 5-3. Boxplot of Criterion Error across 24 Classes ........................................................ 94
Figure 5-4. Matrix Scatter of Continuous Variables ............................................................... 95
Figure 5-5. Normal Q-Q Plots of Residuals at Level 1 and Level 2 (Inter-rater Error) ........... 99
Figure 5-6. Scatter plot of Residual vs. Fitted Value at Level 1 and Level 2 (Inter-rater Error) .......................................................... 100
Figure 5-7. Boxplot of Residuals at Level 1 and Level 2 (Inter-rater Error) ................. 100
Figure 5-8. Normal Q-Q Plots of Residuals at Level 1 and Level 2 (Intra-rater Error) .......... 105
Figure 5-9. Scatter plot of Residual vs. Fitted Value at Level 1 and Level 2 (Intra-rater Error) ...................................................................................................................... 106
Figure 5-10. Boxplot of Residuals at Level 1 and Level 2 (Intra-rater Error) .................... 106
Figure 5-11. Normal Q-Q Plots of Residuals at Level 1 and Level 2 (Criterion Error) ........ 111
Figure 5-12. Scatter plot of Residual vs. Fitted Value at Level 1 and Level 2 (Criterion Error) ...................................................................................................................... 112
Figure 5-13. Boxplot of Residuals at Level 1 and Level 2 (Criterion Error) ...................... 94
LIST OF TABLES

Table 4-1. Assignment of Experimental Condition ................................................................. 66
Table 4-2. Independent Variables and Their Measures ........................................................... 78
Table 5-1. Descriptive Statistics .............................................................................................. 91
Table 5-2. Results from the Unconditional Model (Inter-rater Error) ................................. 96
Table 5-3. Model Comparison Results from the Hypothesized Model (Inter-rater Error) ...... 96
Table 5-4 Results from the Intercepts-and Slopes-as-Outcome Model (Inter-rater Error) .... 98
Table 5-5. Results from the Unconditional Model (Intra-rater Error) ................................. 101
Table 5-6. Model Comparison Results from the Hypothesized Model (Intra-rater Error) ...... 102
Table 5-7. Results from the Intercepts-and Slopes-as-Outcome Model (Intra-rater Error) .... 104
Table 5-8. Results from the Unconditional Model (Criterion Error) ...................................... 107
Table 5-9. Model Comparison Results from the Hypothesized Model (Criterion Error) ........ 108
Table 5-10. Results from the Intercepts-and Slopes-as-Outcome Model (Criterion Error) .... 110
This dissertation would not have been possible without the support of many people. I own my deepest appreciation to the members of my committee for their wisdom, encouragement and guidance throughout my dissertation process. You have all inspired me, challenged me, and supported me, especially during the toughest and most trying work of my dissertation.

First and foremost, I would like to thank my academic advisor, Dr. Hoi K. Suen. Without your patient guidance and everlasting support, I would never have reached this point. Your wisdom, insight, kindness, and encouragement were essential to me to complete this arduous but exciting journey. I will always feel indebted to you for the great amount of effort you put into cultivating me and this dissertation. It is my greatest blessing and honor to be your advisee.

Dr. Pui-Wa Lei, thank you for the training, mentoring, and support you’ve given me. You have always been so encouraging and supportive, since my very first day in the Educational Psychology program. I am deeply impressed by your rigorous scientific attitude and your expertise in measurement and statistics. I feel honored to have learned from you and worked with you.

I would also like thank Dr. Mindy L. Kornhaber. Your helpful critiques have made this research much more meaningful and significant. From you, I see a real life example of a great mind and a humble heart. Your kindness and encouragement warmed my heart and cheered me up during my toughest times. I feel privileged to learn from you and to have you as my committee member.

I am very grateful to Dr. Mosuk Chow for your unfailing patience and everlasting support during the entire research process. Your wisdom in life and insights into statistics have inspired
me tremendously. Without your guidance, I could not have completed my master in statistics and reached this point.

A sincere note of thanks go to Yuan Wan. You have been so kind and passionately ready to help me whenever and wherever I have needed it. Without your help, it would have been impossible for me to recruit sufficient qualified participants to conduct this study and to collect data. I feel so lucky to have worked with you, and look forward to continuing our friendship.

Special thanks must also go to my dearest peers, Yao Xiong, Debora Lee, and Xiaoli Jiang. I am so lucky to have you as classmates in the Educational Psychology program at Penn State. Thank you for assisting me in every aspect of my life.

My deepest gratitude goes to my family and friends for their unconditional love, steadfast support, and endless encouragement. Special thanks goes to my husband, Yongzheng, for his understanding and support. Long distance is challenging, but we have handled it successfully. I would especially like to thank my daughters, Richelle and Rebecca, who are like two little angels that bring so much joy to my life. I greatly appreciate and enjoy every minute we spend together.

Last but not least, acknowledgements must also go to the teachers and students who participated in this study. Without your willingness to volunteer your time, this study would not have been possible.
Chapter 1

Introduction

1.1. Background

Peer assessment is defined as “an arrangement in which individuals consider the amount, level, value, worth, quality or success of the products or outcomes of learning of peers of similar status” (Topping 1998, p. 250). The practice has been widely applied in both educational and non-educational settings for decades. In educational settings, peer assessment is an activity through which students engage with criteria and standards and apply them to make judgments on their peers’ work (Falchikov & Goldfinch, 2000). Peer rating, peer grading, peer appraisal, and peer evaluation are alternative terms for peer assessment. Conditions, contexts, and circumstances that are conducive to using peer assessment are those in which students are involved in the assessment process for active learning, students have unique and accurate view of other members’ performance/contribution in a team project, and the grading load is too heavy for instructors to complete timely feedback. Each of these conditions will be discussed in more detail in the following chapters.

First, peer assessment has been widely adopted as an effective pedagogical method to improve learning processes and learning outcomes. Peer assessment requires peer raters to exercise various cognitive activities such as reviewing, summarizing, clarifying, diagnosing errors or deviations from the ideal, and identifying missing knowledge (Falchikov & Goldfinch, 2000; Mowl & Pain, 1995). When involved in peer assessment, students have an opportunity to apply the knowledge they have learned from a course to criticize and evaluate the work of their peers and develop their own critical thinking skills. Peer assessment also provides students with
opportunities to learn through cognitive modeling of their peers’ work (Lin, Liu, & Yuan, 2001). Due to its cognitive, pedagogical, meta-cognitive, and affective benefits on students’ learning (Lu & Law, 2012), peer rating has frequently been used as a strategy for formative assessment (Cheng & Warren, 1999) by teachers to improve effective teaching (e.g. Purhcase, 2000), teach students assessing skills (e.g. Bloxham & West, 2004), encourage active participation of students in the classroom (e.g. Stefani, 1998), manage social control (e.g. Gibbs, 1999), or any combination of these.

Second, peer assessment is also widely used in collaborative learning environments to provide an effective way to hold students accountable for their behavior and performance in team projects, and to provide a fair and valid way to assign individual grades for a team project. Under this condition, peer rating is both formative and summative in nature. In spite of the amassed substantial evidence on its cognitive and affective benefits (Bartle, Dook, & Mocerino, 2011; Caulfield & Caroline, 2006; Rau & Heyl, 1990), collaborative/team learning suffers from two weaknesses: (1) Some students may have little motivation to engage in the team project and thus, they may learn little from the collaborative project; and (2) often, instructors cannot determine each team members’ contribution to a group project, which makes it very difficult to individually grade each group members’ performance. However, students can be expected to have unique and possibly more accurate views of their team members’ effort, performance, and contributions in a team project. Not only can peer rating provide an effective way to reduce disengagement and frustration from collaborative course activities (Cheng and Warren, 2000; Weaver & Esposto, 2012), it can also serve as a valid way to assign individual marks for group projects (Zhang & Ohland, 2009).

Third, peer rating has been presented as an alternative solution to reduce instructors and teaching assistants’ heavy workload of assessing student work, especially in large classes. The task of assessing exam papers or writing assignments is extremely time- and effort-intensive. In
some large classes it is very difficult, if not impossible, for instructors and teaching assistants to complete such an assessment and give students feedback in a timely manner. Cho, Schunn, & Wilson (2006) explained that despite progress made in the past two decades through the Writing in the Disciplines movement, courses rarely include serious writing tasks because of the excessive workload involved in assessing students’ writing skills. Peer assessment can contribute to reducing this workload, and may also improve learning quality by assigning the assessment task to students (Rada, Michailidis, & Wang, 1994; Cho, Schunn, & Wilson, 2006; Bouzidi & Jaillet, 2009).

Although peer assessment has been used by teachers for many years, it was once challenged for its possible violation of the Family Education Rights and Privacy Act (FERPA). The challenge was raised in the case of Owasso Independent School District v. Falvo in 2000, in which a mother of a six grader, Kristja Falvo, sued the Owasso Independent school district on the grounds that the use of peer grading by her child’s school violated students’ rights to privacy. Falvo argued that student grades are an education record and should not be released to others in the class without a parent’s permission. While her complaint was supported by three judges on the Tenth Circuit Court of Appeal, the Supreme Court ruled that the grades generated by peer assessment did not satisfy the definition of an “education record” and that peer assessment activities do not violate FERPA. The ruling dispelled teachers’ concerns, and provides solid legal evidence that peer assessment can be used in educational settings (Parry, 2002).

Recently, peer assessment has drawn more attention from many researchers and educators due to the development and expansion of massive open online courses (MOOCs). Since George Siemens developed the first MOOC with his colleagues at the University of Manitoba in 2008, MOOCs have been generating worldwide interest and lively discussions in higher education (Mackness, Mak, &Williams, 2010). Numerous instructors worldwide have since
developed and offered MOOCs at no financial cost. There are more than 16.8 million students who have enrolled in MOOCs to date (Shah, 2014).

The most distinguishable difference between MOOCs and traditional online courses is the teacher-to-student ratio. For each MOOC there can be as many as over one hundred thousand enrolled students, so instructor-to-student interaction and feedback are fiscally and physically impossible. Furthermore, the teach-learn-assess cycle is broken. As a result, it is difficult for students and instructors to engage in appropriate and timely feedback to ensure proper learning. To find a way to assess the achievements of thousands of students, or to measure their mastery of course material, MOOC providers and researchers propose and utilize a variety of approaches including multiple choice testing, on-line discussion forums, automated essay scoring algorithms, and flipped classrooms. After examining these assessment methods, Suen (2014) concluded that peer assessment provides the most universally applicable method to introduce assessment back into the learning cycle in MOOCs. Coursera, which offers almost half of all MOOCs and is the largest MOOC platform provider, has recommended peer assessment as a formative pedagogical tool and a source of summative assessment.

In sum, peer assessment is desirable in several situations. It has been widely used as a formative pedagogical tool to help students to become active, responsible, and reflective learners in both traditional face-to-face classrooms and online courses because of the amassed substantial evidence of its cognitive, pedagogical, meta-cognitive, and affective benefits. Also, peer assessment provides an effective way to hold students accountable for their behavior and performance in team projects, and to provide a fair and valid way to assign individuals grades for team projects. Thus, peer assessment functions both as a formative and as a summative assessment in collaborative learning environments. Finally, peer assessment is commonly used to reduce instructors’ heavy workload of assessment in large classes. In MOOCs, peer assessment is
especially crucial because it allows instructors to measure students’ achievement and progress without having to rely on a computerized grading method.

1.2. Statement of the problem

Although peer assessment is desirable in many circumstances, and there is a substantial body of literature documenting diverse learning benefits for student participation in peer assessment (e.g. Falchikov, 1995; Topping, Smith, Swanson, & Elliot, 2000), instructors are reluctant to implement peer assessment in their course design, especially as a tool for summative assessment. One reason for this phenomenon is instructors’ concern regarding the reliability and validity of peer assessment outcomes.

In peer assessment, one way of defining reliability and validity is that reliability is about the consistency and stability of scores, whereas validity is the degree of agreement or similarity between the scores assigned by peers and those assigned by instructors (Falchikov & Goldfinch, 2000). Within this perspective, there are two aspects of reliability: Inter-rater reliability, which is the extent of agreements in ratings between different raters of the same performance; and intra-rater reliability, which is rating consistency within an individual rater. According to Standards for Educational and Psychological Testing (2014), the inter-rater reliability and intra-rater reliability in this study are not score reliabilities, but rather are measures of rater interchangeability and rater consistency. Since the main concern of most researchers is the degree of agreement between the scores given by peers and by instructors, given the above definitions, many studies that claim to have investigated peer assessment reliability actually investigated the validity of peer assessment (Bouzidi & Jailllet, 2009; Stefani, 1994; Falchikov, 1986). When peer rating is used as a summative assessment, high reliability and validity is necessary for meaningful and fair score interpretation.
Therefore, an important endeavor in educational research is to determine what factors can improve the reliability and validity of peer assessment results. Some efforts have focused on factors related to peer assessment settings. For example, accumulated evidence indicates that the correlation between peer rating and teacher rating for paper-based peer assessment is significantly higher than computer-assisted peer assessment, and graduate courses generate significantly highly agreement between teacher and peer raters (Li et al., 2015). Others emphasized the management of the assessment procedure. Falchikov and Goldfinch (2000), for instance, found that peer rating is closer to teacher rating when global judgments (i.e., holistic scoring) are used rather than when rating involves assessing several individual dimensions (i.e. analytic scoring). Similarly, peer rating better resembles faculty assessment when students are involved in the development of rating criteria as opposed to when they are not (Orsmond, Merry, & Reiling, 1996). The accuracy of peer assessment can also be significantly improved when it is non-anonymous and voluntary (Li et al, 2015). However, all of the aforementioned factors can only explain a small proportion of variations in the agreement between peer ratings and teacher ratings (see the meta-analysis conducted by Li et al, 2015 for more details), which indicates a need to investigate other potential factors.

As Overall and Magee (1992) suggest, the reliability and validity of scores based on a rating scale, regardless of how carefully constructed, is critically dependent on the raters who operate it. By definition, the largest difference between peer ratings and teacher ratings is the quality of the rater. In the case of teacher ratings, the raters are teachers who, relative to the students, are experts in the assessment domain and have rich experience in rating. Student raters, on the other hand, are novices both in the content knowledge domain and in assessment skills. Also, student peer raters may lack motivation to assess their peers’ work seriously, which may lead to meaningless evaluation results. Therefore, it is reasonable to propose that peer rater’s
quality characteristics are potential determinants of the reliability and validity of peer assessment outcomes.

### 1.3. Present Study

In order to provide reliable and valid peer assessment results, peer raters need (1) good content knowledge in order to understand their peer’s work, (2) good rating skills in order to make accurate judgment about their peer’s work, and (3) enough motivation to make a responsible and serious judgment. There has been relatively little research dedicated to examining peer rater’s characteristics. It would seem potentially fruitful to examine the impact of these quality characteristics of peer raters on the reliability and validity of peer assessment results. To this end, the main goal of the present study is to investigate the effect of peer raters’ quality characteristics on the rater reliability and rating validity of peer assessment.

Specifically, this study intends to examine how peer raters’ content knowledge, training experience, and motivation of making serious and responsible judgment influence the inter-rater reliability, intra-rater reliability, and rating validity of peer assessment scores when judging written compositions. It also intends to examine if there are interaction effects among peer raters’ content knowledge, training experience, and rating motivation. The following research questions are addressed:

1. Does peer raters’ content knowledge in the assessed domain affect the rater reliability and rating validity of peer assessment?
2. Does peer raters’ training experiences lead to more reliable and valid peer assessment scores?
3. Does peer raters’ rating motivation influence the reliability and validity of peer assessment?
4. Is there an interaction effect between peer raters’ rating motivation and training experience on the reliability and validity of peer assessment?

5. Is there an interaction effect between peer raters’ rating motivation and content knowledge on the reliability and validity of peer assessment?

6. Is there an interaction effect between peer raters’ training experience and content knowledge on the reliability and validity of peer assessment?

The corresponding hypotheses and the literature involved in generating these hypotheses are described in Chapter 2. In order to address these research questions, a randomized cluster trial design study was conducted with a sample of 838 students from 24 classes in a typical Chinese high school. Each participant was expected to rate four well-selected essays written by peers for a Chinese language mid-term examination using the standard Gaokao (National College Entrance Examination) Chinese Wring Rubric. Before performing peer assessment, students in half of the classes were randomly selected to receive a self-paced training module on peer rating. Similarly, to manipulate students’ rating motivation, randomly chosen classes of students were offered motivational incentives which included either (1) monetary rewards for accurate peer rating results, or (2) teachers’ verbal persuasion on the value of performing peer assessment. Because the control groups did not receive any motivational incentives, it was possible for them to experience feelings of resentment and demoralization if they knew such incentives were provided to the experimental group, which could lead to an abnormally low performance. In order to eliminate this possibility, the control group performed peer assessments one day earlier than the experimental groups. This $2 \times 2 \times 2$ completely crossed factorial design was employed with two levels of training intervention (i.e., self-paced training vs. no training), two levels of monetary reward (i.e. monetary incentives vs. no monetary incentives, and two levels of verbal persuasion (i.e., verbal persuasion vs. no verbal persuasion).
Peer raters’ content knowledge in the assessed domain was measured by their Chinese language mid-term examination score. Their training experience was measured by their previous peer rating experience and whether or not their classes received such self-paced training. Peer raters’ rating motivation was measured by what motivation intervention they received (e.g., only monetary incentives, only verbal persuasion, both, or none). A detailed description of the design, procedures, and instrumentations and measures is provided in Chapter 3. All of the data are analyzed using a series of two-level, hierarchical linear models (HLM). Chapter 4 reports the analyzed results. Finally, Chapter 5 discusses the overall findings and their implications for traditional and online peer assessments. Limitations of the dissertation and important areas for future research are also addressed.
2.1. Understanding peer assessment

This section articulates the theoretical foundation of peer assessment, describes where and why peer assessment is desirable, and addresses the problem of peer assessment with a focus on its possible low reliability and low validity.

2.1.1. Theoretical foundation of peer assessment

Peer assessment is grounded in the theory of social constructionism, as it often involves the joint construction of knowledge through discourse. It may also be seen as a manifestation of philosophies of active learning (Falchikov & Goldfinch, 2000). Peer assessment activities have been found as an effective way to promote learning processes and outcomes, and as Falchikov and Goldfinch note, “it is this aspect which commonly forms the rationale for introducing peer assessment into courses” (p.288).

Constructivist theory emphasizes a learner’s role in building knowledge and skills. A basic assumption of teaching and learning, according to constructivist theory, is that knowledge and skills cannot simply be transferred from instructor to learners; instead, learners must be engaged in constructing their knowledge and skills (Von Glaserfeld, 1989). The learners’ minds respond to new knowledge either by assimilation, when their current knowledge is an adequate basis to interpret the new experience, or by accommodation, when there is a conflict between the expectations that derive from their current knowledge and actual experience. The process of accommodation in response to such disequilibrium puts the learners in an active position and
makes them change to conform to the new knowledge (Piaget, 1977). Constructivist instruction is expected to provide a social and cultural context which encourages a learner to “internalize knowledge, whether 'discovered,' transmitted from others, or experienced in interaction with others” (Lave & Wenger, 1991, p.47). In doing so, a learner can experience an ownership which enables the learner to understand the knowledge, or acquire skills in an intimate way that cannot be achieved by rote learning (Dori & Belcher, 2005).

Active learning is initially promoted by Dewey (1924) and it typically goes hand in hand with constructivist ideas. Practitioners of active learning theory also believe that knowledge is not imposed from the outside, but is gained by learners’ active engagement. They place less emphasis on transmitting information and more on developing students’ skills (Keyser, 2000), encouraging students to engage in problem solving, sharing ideas, giving feedback, and teaching one another (Johnson, Johnson & Smith, 1991). Peer assessment can be used as one of active learning strategies to advance students learning.

Compared with the original conception of constructivism, social constructivism puts more emphasis on the role of cultural and social integration. Social constructivists believe that higher mental functions begin in the social world and learner interacts with the members of society to construct knowledge actively. Learning is defined as development through internalization of the content and tools of thinking within a cultural context (Duit & Treagust, 1998; Vygotsky, 1963). The role of social interaction is central to learning. Peers help each other by offering alternatives and sustaining reasoning activities, and individuals can integrate knowledge from peers and cultural context and use their learned knowledge and cognitive skills privately or interpersonally (Vygotsky, 1978). Therefore, the growth of cognitive proficiency involves appropriating intellectual wealth from the social and cultural space. This insight highlights the power of group processes to advance educational aims.
In peer assessment, social constructivism can provide a theoretical explanation for a learner’s role in building knowledge and skills. From the social constructivist point of view, individuals learn through dialectical processes. Dialectical processes are interaction patterns in which learners exchange ideas, share understanding, negotiate meanings, and apply their knowledge to understanding new content (Vygotsky, 1963). Peer assessment provides an environment for such interaction since it allows students to construct knowledge through social sharing and negotiating (Liu, Lin, & Yuan, 2001).

In the process of peer assessment, students take dual roles both as assessors and as the assessed. Social constructivism provide theoretical framework for both roles. As assessors, students are expected to apply their knowledge to evaluate their peers’ work and give feedback. They review, summarize, clarify, diagnose errors, identify missing knowledge or deviation from the ideal, and discussion with peers or instructors (Falchikov & Goldfinch, 2000; Van Lehn, et al., 1995). Through exercising these cognitive and social activities, assessors can have deeper understanding of content knowledge covered in the peer assessment domain, construct new knowledge based on the interaction with other peers, and develop important cognitive skills such as problem solving, evaluating, decision making, and communicating.

When taking the role of the assessed, the learner’s main task is to deal with feedback given by peer assessors. This process usually requires the assessed to exercise various cognitive activities such as analyzing, comparing, reflecting, exchanging, negotiating, and revising (Topping, 1998; Xiao & Lucking, 2008). According to social constructivism, a learning approach related to a student’s role in receiving peer assessment is scaffolding, which is the application derived from Vygotsky’s theory of the Zone of Proximal Development (ZPD). Scaffolding refers to the process in which those who are more competent assist the learner in completing aspects of a task that he or she can only complete with help (Bruner, 1963, 1975). Typically, scaffolding is completed by focusing on the gradual “release of responsibility” from more competent others to
the learner, resulting in a learner eventually becoming fully responsible for his or her own performance. This gradual release of responsibility is accomplished by continuously decreasing the degree of assistance provided by more competent others without altering the learning task itself. Through peer assessment, the assessed can benefit from their peers’ scaffolding feedback to facilitate learning (Xiao, 2010). However, not all feedback is from more competent peer assessors. Only high quality peer feedback (e.g. accurate, specific, practical, critical and achievable feedback) can help the assessed substantively acquire the scaffolding to advance their learning (Xiao, 2010). To improve the quality of peer feedback and to scaffold learning by the peer receiving assessment, peer assessors need to be guided by teachers to implement scaffolding before they perform peer assessment (Orsmond, Merry, & Reiling, 2000; Sluijsmans, Brandgruwel, & van Merrienboer, 2002).

2.1.2. Desirable circumstance for peer assessment

Due to its various advantages, peer assessment is desirable in many academic contexts and has been employed in many circumstances. This section will describe several typical circumstances, explain when peer assessments are desirable in these circumstances with examples of the corresponding purpose and peer assessment design, and discuss each option’s respective advantages and disadvantages

2.1.2.1. As an effective pedagogical method

Peer assessment has been widely adopted as an effective pedagogical method to improve learning processes and learning outcomes. Students are involved in the assessment process for active learning and are expected to apply the knowledge they have learned from a course to
critique and evaluate the work of their peers. Peer assessment also provides students with more opportunities to learn through cognitive modeling of their peers’ work (Lin, Liu, & Yuan, 2001). Moreover, in some peer assessment situations, students are involved not only in the final judgment of peers’ work, but also in the development of assessing criteria (e.g. Orsmond, Merry, & Reiling, 2000; Smith, Cooper, & Lancaster, 2002). There is substantial empirical evidence on the benefits of involving students to develop assessing criteria. The most obvious one is that such involvement can promote a sense of ownership and increase student autonomy (Falchikov, 1995), personal responsibility, and motivation (Topping, 1998). This, in turn, will improve the quality of the learning process and learning outcome. Through involvement in developing a rubric, students may gain a better understanding of the evaluation criteria and have the opportunity to apply them to peer assessment activity more easily (Orsmond, Merry, & Reiling, 1996).

Due to its cognitive, pedagogical, meta-cognitive, and affective benefits on students’ learning (Lu & Law, 2012), peer rating has frequently been used as a strategy for formative assessment (Cheng & Warren, 1999) and for involving students as active learners (Gielen, Peeters, Dochy, Onghena, & Struyven, 2010). For example, Liu, Lin, and Yuan (2001) implemented peer assessment via a web-based peer review (WPR) system, which was a piece of software that was developed by the authors. The participants were 143 computer science undergraduate students in an Operating System class at a Taiwanese university. Each student prepared a project independently and uploaded it to the WPR system; the system then randomly assigned six peer raters for each project. After students rated and commented on peers’ projects, the system distributed the scores and comments back to each submitter and informed the teacher about the general status of the class. The submitter was required to revise the original assignment based on the comments they received. After three rounds, the teacher rated each final project based on the six peer rater’s grades, and also rated each peer rater’s comment quality. To get a high score in this circumstance, a student had to play multiple roles: first as an effective author;
then as a responsible peer rater with critical thinking; and finally as an adapter who benefited from peers’ feedback. To examine the learning effect and students’ perception of peer assessment, the participants were asked to rate their satisfaction on a five-point Likert-type survey scale containing ten questions. The results indicated that students perceived peer assessment as an effective learning tool, and they reported benefiting from reading peers’ projects and peers’ feedback. Many students reported that by comparing their own work with peers’ work, they became more aware of their strengths and weaknesses than in a conventional teacher assessment situation. Additionally, this study reported that 77% of peer raters displayed all four kinds of higher level thinking: planning, monitoring, regulation, and critical thinking.

When used as a formative assessment approach, peer assessment is a pedagogical tool rather than an assessment tool, and its main purpose is for learning and not grading; it provides the rated with information about their progress and empowers them in active and collaborative learning (Liu & Li, 2014). Thus, formative peer assessment has been adopted by teachers to improve effective teaching (e.g., Purhcase, 2000), teach students assessing skills (e.g., Bloxham & West, 2004), encourage active participation of students in the classroom (e.g., Stefani, 1998), and manage social control (e.g., Gibbs, 1999), or any combination of these.

The potential advantages of peer assessment include the development of the skills of evaluating, justifying, and using discipline knowledge (Topping, Smith, Swanson, & Elliot, 2000). Topping et al. (2000) also suggested that peer assessment can increase post hoc reflection, improve student ability to generalize knowledge to new situations, and promote self-assessment and self-awareness. The other side of the coin is that some students feel negatively about peer assessment because raters are also competitors (Lin et al., 2001). During the peer assessment process, students may utilize different assessment criteria that produce inconsistent assessment results. Moreover, students often lack the ability to evaluate each other’s work or may apply judgments that are excessively subjective. Zhao (1998) also pointed out that students commonly
believe that only instructors have the ability and knowledge required to effectively evaluate work and provide critical feedback.

2.1.2.2. As part of a collaborative learning environment

Peer assessment is also widely used in collaborative learning environments to provide an effective way to hold students accountable for their behavior and performance in team projects, and to provide a fair way to assign individual grades for a team project. Under this condition, students have unique and close-up views of other members’ performance and contributions in a team project. As such, peer assessment is both formative and summative in nature. In spite of the substantial evidence amassed on the cognitive and affective benefits of team work (Bartle, Dook, & Mocerino, 2011; Caulfield & Caroline, 2006; Rau & Heyl, 1990), collaborative and team learning suffered from two weaknesses: (1) Some students may have little motivation to engage in the team project and thus, they may learn little from the collaborative project; and (2) often, instructors cannot observe first-hand information of each team member’s contribution to a group project, which makes it very difficult for them to grade each group member’s performance. However, students can be expected to have unique and more direct perceptions of other members’ effort, performance, and contributions in a team project. Not only can peer rating provide an effective way to reduce disengagement and frustration from collaborative course activities (Cheng and Warren, 2000; Weaver & Esposto, 2012), it can also serve as a more informative way to assign individual marks for group projects (Zhang & Ohland, 2009).

In a study investigating the influence of a particular peer rating design on peer rating in a group project, Dingel & Wei (2014) employed a sample 113 undergraduate students enrolled in an entry-level sociology course. To create diverse groups of race and gender, instructors randomly assigned students to 22 groups that ranged in size from four to six students. The project
required each group to analyze a subset of United States census data and write three papers collaboratively. These three papers made up 20% of the course grade for each individual student. Instructors rated each paper and assigned them a group score; each student then rated themselves and each of their group members using a behaviorally anchored, nine-point-scale rating that measured effort and contribution. The average peer rating score was calculated for each individual group member and for the overall group. The instructors divided the individual average by the group average to get an “adjustment factor” for each student. Finally, each individual student’s score was calculated by multiplying their adjustment factor by his or her team paper score. The research found a positive correlation between students’ course performance and peer evaluation scores.

In typical peer assessments for a team project, students confidentially rate their group members only once at the end of the semester. Although students feel less peer pressure when the peer rating process is confidential, rating only at the end of the term may result in undesirable behaviors by group members. For instance, students may tolerate some group members’ negative attitudes and irresponsible behaviors, thinking that they can “burn” these poor performers at the end of the semester on the peer rating (Bacon, Stewart, & Silver, 1999). Under this perception, group members who are poor performers gain little to nothing from peer rating during the whole process of the team project. Such students only face the consequences of a low score at the end of the team project due to their poor performances. Thus, peer rating cannot serve as an effective tool to improve group performance over the duration of a course or learning experience, or to motivate students to engage in the team project (Strong & Anderson, 1990). Also, compared to multiple time-points peer rating, single-time peer rating has less accurate assessment results (Dominick, Reilly, & McGourty, 1997).

To overcome some of the disadvantages associated with peer rating that occurs only at the end of the semester, Brooks and Ammons (2003) developed a peer assessment procedure that
features early implementation, multiple evaluation points, and specific rating criteria. They applied this peer rating design to 340 undergraduate students enrolled in an introductory business course which comprised three modules. During each module, students completed a team project and made a classroom presentation. At the end of each module, students completed a peer rating package outside of classroom hours for confidentiality purposes, placed the peer rating package in a sealed envelope, and handed the envelope to the instructor after the completion of the team project and presentation. The peer rating package allowed students within a group to rate each other based on three different criteria. Students first rated each group member’s performance and contribution on a five-point Likert-type scale (1=Never, 5=Always), taking into consideration things such as attendance and distribution of workload. Secondly, students made qualitative comments about each group member’s performance and contribution. Finally, students distributed numerical points to each group member. The total number of points for distribution equals the number of team member times 100. If an individual did more than his or her fair share, the peer rater would distribute more than 100 points to that person. Likewise, if someone did less, the peer rater would distribute that person less than 100 points. The average peer rating score for each individual student was used to weight the instructor’s grade for the team project. The three projects accounted for a combined total of 31.25% of each student’s course grade.

The authors suggested that using this multiple-point peer rating design that provides qualitative feedback early on and at multiple time points during a team project can reduce the free-rider problem that occurs when one or more students in the group do not participate fully in tasks, thereby contributing almost nothing to the well-being of the group (Brooks & Ammons, 2003). Multiple-point peer rating activities offer students several opportunities to express their dissatisfaction by rating free-riders lower on the scale multiple times; this may, in turn, motivate the free-riders to put forth more effort. Simultaneously, such a design may allow students to identify aspects of poor performance and ways to improve, and lead students to view group
experiences in a more positive light. When peer raters are involved in qualitative feedback regarding an individual’s performance in a team project across time, with multiple rating opportunities, peer ratings tend to be more accurate. However, this multiple-time-point peer rating design demands great effort from both instructors and students. It is not easy for instructors to administer and monitor this kind of peer rating in team projects.

To further facilitate the management and administration of peer rating in group work, several web-based software programs have been developed. Computer Assisted Self and Peer Assessment Rating (CAPSAR) is one such software tool designed by Bournemouth University (Lugosi, 2010). It allows instructors to set up group projects, and students to participate in self- and peer-rating once or multiple times throughout the life of the group project. Alternative rating criteria can be used at different phases of each group project. Besides numerical scores, CASPAR also allows students to provide qualitative feedback, and allows instructors to monitor progress, moderate scores, and collect feedback.

Although peer rating can be administered more effectively with the assistance of web-based software, it still poses several challenges to the reliability and validity of its results when it is used as a summative tool to grade each individual’s performance within a team project. Students may intentionally overrate a team member due to friendship or peer pressure, or underrate a team member due to competition. Moreover, in team project peer rating, team members usually know each other well; students with similar demographic backgrounds, such as gender, age, and race/ethnicity, may hold similar biases that influence their peer rating. Due to the mixed findings regarding the effects of various demographic variables (e.g., Dingel & Wei, 2014; Falchikov & Magin, 1997; Ghorpade & Lackritz, 2001; Watson, BarNir, & Pavur, 2010;) on peer ratings in group work, instructors must be aware of and consider these issues when forming groups.
Peer assessment has been presented as an alternative solution to reducing instructors and teaching assistants’ heavy workload of assessment, especially in large classes. The task of assessing exam papers or writing assignments is extremely time and effort intensive. In some large classes, it is very difficult, if not impossible, for instructors and teaching assistants to complete such an assessment task and give students feedback in a timely manner. Cho, Schunn, & Wilson (2006) explained that because of the excessive workload of assessing writing skills, despite progress made in the past two decades through the Writing in the Disciplines movement, current college courses rarely include serious writing tasks. Peer assessment can contribute to reducing this workload and, possibly, also improve learning quality by assigning the assessment task to students. (Rada, Michailidis, & Wang, 1994; Cho, Schunn, & Wilson, 2006; Bouzidi & Jaillet, 2009).

Bouzidi and Jaillet (2009) carried out an online peer assessment of exam papers among 242 undergraduate students enrolled in three different engineering courses. Students first completed an exam in a supervised class which required simple calculations, mathematical reasoning, short algorithms, and short textual responses. Each exam paper was then collected and digitalized to ensure that students remained anonymous. Each student was asked to rate four digitalized exam papers based on clear assessment criteria developed by instructors. In order to have a higher quality peer assessment, the students were told that the quality of their assessment would be marked and that the assessment would be anonymous. In addition to peer assessments, students also did self-assessments. The results indicate that peer rating is equivalent to teacher rating. The authors suggested that a combination of peer assessment and self-assessment would lead to better accuracy of the results.
Besides assessing exam papers, peer rating has also been widely used to assess writing tasks. For example, Cho, Schunn, & Wilson (2006) reported a peer assessment study that involved 708 students across four universities in which writing tasks were evaluated. The writing tasks assigned to students varied across 16 different courses, and the types of papers included in the courses consisted of the following: the introduction section to a research paper, a proposal for an application of a research finding to real life, a critique of a research paper, and a proposal for a new research study. All of the first drafts were uploaded to Scaffolded Writing and Rewriting in the Discipline (SWoRD), an online system for implementing peer assessments of student writing, and then each paper was randomly assigned to five or six peer raters. Peer raters were given guidance on peer assessment and then used carefully constructed rubrics to grade the assigned drafts on three 7-point evaluation dimensions, as well as to provide comments. When the assessment deadline had passed, SWoRD automatically determined grades for each draft based on the average peer assessment scores. SWoRD also calculated a numerical evaluation for each peer rater based on three automatically determined measures of review consistency. After reviewing the grade of their first drafts, students were asked to revise their papers according to the comments their peers provided, and then to upload their final drafts back into SWoRD. Each final draft was distributed to the same peer raters from the first round for grading. Once the draft had been submitted, each student was asked to rate the helpfulness of each comment he or she received using a 7-point helpfulness scale. These ratings constituted the other half of peer rater’s reviewing grade. An obvious advantage of this peer assessment design is that it provides clear incentives for peer raters to take peer assessment tasks seriously. SWoRD’s automatic grading system for peer raters emphasizes and rewards consistency with other peer raters, thus, it should increase inter-rater reliability. On the other hand, exclusively emphasizing consistency and agreement among peer raters may encourage central tendency, which is a threat to assessment validity.
Researchers have paid significant attention to peer assessment in recent years due to the advent of Massive Open Online Courses (MOOCs). MOOCs have inspired widespread interest and lively academic debate since George Siemens and a colleague developed the first MOOC at the University of Manitoba in 2008 (Mackness, Mak, & Williams, 2010). Instructors worldwide have developed and offered MOOCs at no financial cost to students, allowing millions of people with diverse goals, experiences, languages, cultural backgrounds, and motivations to enroll. By the end of 2014, more than 16.8 million students had enrolled in a MOOC since the first course was offered in 2008 (Shah, 2014).

The teacher-to-student ratio is the most distinguishable difference between MOOCs and traditional online courses; in each MOOC, there can be over one hundred thousand enrolled students with just a single professor. Instructor-to-student interaction and feedback are fiscally and physically impossible in a MOOC, which breaks the teach-learn-assess cycle. As a result, it is difficult for students and instructors to engage in appropriate and timely feedback. MOOC providers and researchers propose and utilize a variety of approaches to assess the achievements and course mastery of thousands of students including multiple choice testing, on-line discussion forums, automated essay scoring algorithms, and flipped classrooms. Suen (2014) examined these assessment methods and concluded that peer assessment provides the most universally applicable method to reintroduce assessment into the learning cycle of a MOOC. Peer assessment has also been recommended by Coursera, which offers almost half of all available MOOCs and is the largest MOOC platform provider.

For example, in 2013, the Pennsylvania State University offered a MOOC on Coursera named The Maps and the Geospatial Revolution, which was a five-week MOOC focusing on geospatial technology, map making, and analyzing geographic patterns using current tools. There were 48,984 students who registered for the course, with 8,707 students remaining active through the last week of the course. In the last week of the course, students were expected to submit an
assignment in which they applied the knowledge they learned in the course to create a map in a storytelling context. In addition, students were asked to rate at least three completed storytelling projects submitted by their peers based on an instructor-designed rubric, and to provide open-ended written feedback to these peers. The median was used to determine the final peer rating score for the assignment; the peer assessment assignment accounted for 20% of each student’s overall grade (Luo, Robinson, & Park, 2014).

2.1.3 Problems with peer assessment

Although peer rating is desirable in many circumstances, it raises two major common concerns among researchers and educators. One is students’ perception of, and attitude towards, peer assessment. The other is reliability and validity concerns.

2.1.3.1 Students’ negative attitude towards peer rating

Due to the amassed substantial evidence of its benefits serving either as a formative pedagogical tool or as a summative assessment method, most teachers hold positive attitudes toward peer assessment (Topping, 1998; Davies, 2000; Wen, Tsai, & Chang, 2006). However, some students have negative feelings about peer assessment. According to Wen and Tsai’s (2006) report of 280 university students’ perceptions of, and attitudes towards, peer assessment, most students would prefer peer assessment scores to constitute a minor part of the total course grade and none of the students favored total reliance on the peer assessment score itself. Students perceived that the major part of the responsibility of grading should belong to teachers. Students reported more resistance to providing numerical scores than providing informal feedback for their peers’ work (McGarr & Clifford, 2012).
Why do students hold negative attitudes about peer rating? Possible reasons include their doubt of peer raters’ expertise, perceived power relations, and concerns about time. First, some students may feel that their classmates, as novice peer raters, are not qualified to make judgments on their work or provide insightful feedback (Liu & Carless, 2006). Others may believe that only teachers have the ability and knowledge to evaluate and provide critical feedback (Zhao, 1998). Therefore, students do not trust the whole process or the final results of peer rating.

The second reason for students’ negative attitude toward peer rating is that peer rating disrupts power relations in a class, where traditionally only teachers had the power to assess students’ work. Students may dislike having power over peers or peers having power over them (Falchikov, 2001). Many students feel uncomfortable rating their friends or classmates (McGarr & Clifford, 2012). Also, in peer assessment, students’ work becomes available to their peers, not just to the teacher. Such privacy concerns have also been raised by parents and teachers. Based on the 2002 US Supreme Court unanimous ruling on the Owasso Public Schools v. Falvo case (2002), peer assessment activities do not violate the Family Education Rights and Privacy Act (FEPPA) because the grades given by peer raters do not satisfy the definition of an “education record” (Parry, 2002). Although, this decision dispels teachers’ concern and provides solid legal grounding that peer assessment can be used in educational settings, students may still feel uneasy because they lose some control over their own work (Liu & Carless, 2006).

Finally, peer assessment is time-consuming and effort-demanding for students. As mentioned earlier, peer assessment can reduce teacher’s heavy burden of assessment and provide feedback to students more promptly by involving students in the assessment process. However, from students’ perspective, they have to share the teacher’s workload, devote additional time, and put a great amount of effort into grading their peer’s work. Thus, the perception of the time-consuming nature of peer assessment tasks may lead to students’ negative attitude towards it.
2.1.3.2. Possible low reliability and validity of peer rating

In addition to students’ attitude, reliability and validity are other major concerns in peer assessment. When peer rating is used as a summative assessment, people worry about its possible low reliability and validity. Cho, Schunn and Wilson (2006) referred to such worries as a problem of face validity. First, student peer raters are novices in their disciplines with respect to content knowledge; therefore, they may not have enough knowledge to discern the quality of the content of their peers’ work. Second, student peer raters are inexperienced in assessing and may lack the relevant skills to make accurate judgments on their peers’ work. Third, student peer raters may lack motivation to assess their peers’ work seriously, which may lead to meaningless evaluations. Fourth, student peer raters are prone to bias due to gender, race, ethnicity, and friendship (Dancer and Dancer, 1992).

These threats, however, do not mean that high reliability and validity of peer assessment is impossible. Indeed, there are theoretical and practical reasons to suspect that peer ratings are just as reliable and valid as instructor ratings. Cho, Schunn, & Wilson (2006) listed several advantages of peer ratings over instructor ratings in terms of reliability and validity. First, the reliability of instructor rating can be impaired by drifting effect. Because instructors usually have to grade a large stack of assignments in a short time period, they are more likely to shift rating criteria over time. This drifting effect is not a concern for peer raters, who are typically given a much smaller set to grade and can spend more time rating each given assignment. Second, under instructor rating, each assignment is usually graded by the instructor alone with, at most, the help of a second rating by a teaching assistant. Under a peer rating system, on the other hand, each paper can be assigned to multiple peers for grading. The combined ratings of multiple raters tend to have higher reliability than ratings from a single rater. Third, validity of instructor ratings can be threatened by certain biases, such as an instructor’s prior knowledge of each students’
performance in class. Peers in a large class, however, are less likely to have previous performance information about randomly selected classmates. Fourth, instructor ratings may have validity problems due to floor effect. Because instructors have expertise in the knowledge domain of their class and may have unrealistically high performance expectations for students, they may have difficulty distinguishing between performance levels at the low end of the scale. Peer raters, who are, on average, in the middle of that distribution themselves, tend to have an easier time perceiving such differences.

2.2. Reliability and validity of peer assessment

This section begins with clarifying the definition of the reliability and validity of peer assessment. It continues with reviewing empirical evidence and methods used to calculate the reliability and validity of peer assessment, and ends with summarizing influential factors that have been investigated to improve the reliability and validity of peer assessment.

Before further discussion of this concern, the confusion surrounding the terms reliability and validity in existing literature must be clarified. Falchikov and Goldfinch (2000) point out that validity is sometimes misreported in the literature as reliability. Based on Falchikov and Goldfinch, reliability means the consistency and stability of scores, whereas validity means the degree of agreement or similarity between the scores assigned by peers and those assigned by instructors. Because the main concern of most researchers is the degree of agreement between the scores given by peers and by instructors, many studies actually investigated the validity of peer assessment rather than the reliability (Bouzidi & Jailllet, 2009; Stefani, 1994; Falchikov, 1986).
2.2.1. Empirical evidence for reliability and validity of peer assessment

2.2.1.1. Reliability

In peer assessment, there are two types of reliability: Inter-rater reliability (i.e., external reliability), and intra-rater reliability (i.e., internal reliability). Inter-rater reliability can be defined as the extent of agreement between different raters of the same performance or assignment; while intra-rater reliability refers to the consistency within an individual rater across different assignments (Chang, Tseng, Chou, & Chen, 2011). In other words, if the primary concern is the extent of exact agreement among raters on the same assignment evaluation, then it is examining inter-rater reliability. If, however, one is interested in the extent to which raters consistently apply a rating rubric to assess their peers’ work, one should examine intra-rater reliability.

Based on designs and the matching patterns of peer raters and peer ratees (e.g., the number of peer raters for one assignment; different vs. same group of peer raters assessing the same set of assignments; peer ranking vs. peer grading with numerical score on continuous scales), various metrics have been used to calculate peer rating reliability. The most common measure of peer rating reliability is the intraclass correlation coefficient (ICC) approach because of its wider applicability. There are different forms of ICC, and the selection of a particular form should depend on the specific circumstance in which it is being used, such as whether the assessed assignments are considered random or fixed, whether the peer raters are considered as random or fixed, whether the interest is inter-rater reliability or intra-rater reliability, whether the same peer raters assess the same assignment, and whether one wants to examine the reliability of a single rater or the reliability of the full set of peer ratings of a given assignment taken together (McGraw & Wong, 1996; Cho, Schunn, & Wilson, 2006). ICC is computed via an analysis of
variance approach. In this analysis of variance, the dependent variable is peer rating, and the
independent variables include peer raters (R) and assignments/exams (A). Essentially, ICC goes
up as the mean square of the assignment effect (MSA) increases, and it goes down as the mean
square of the interaction of peer raters and assignment (MSA×R) increases.

According to McGraw and Wong (1996), if the interest is inter-rater reliability for Single
Measurement (i.e. the agreement among single peer rating scores given by individual raters who
are assigned to rate the same assignment) rather than Average Measurement (i.e. the agreements
among the average peer rating scores calculated for each subgroup raters who are assigned to rate
the same assignment), one can use ICC (A, 1) Case 2, where A indicates agreement, 1 indicates
Single Measurement, and Case 2 refers to the case of two-way random effects (i.e. random rater
and random assignments) with interaction. This is a specific case within the Generalizability
theory (Brennan, 2010). Equation 1 is the formula for estimating inter-rater reliability for single
raters.

\[
ICC (A, 1) = \frac{MS_A - MSR_{\times A}}{MS_A + (n-1)MSR_{\times A} + \frac{N}{n}(MSR - MSR_{\times A})},
\]  
\text{Equation (1)}

where MSA = mean square for assignment effect; MSR = mean square for rater effect; MSA×R =
mean square for interaction of rater and assignment; N is the total number of raters; n= the
number of raters per assignment.

If the intra-rater reliability for Single Measurement is the concern, the ICC (C, 1) Case 2,
where C indicates consistency, 1 indicates Single Measurement, and Case 2 refers to the case of
two-way random effects (i.e. random rater and random assignments) with interaction. This is
another special case within the Generalizability theory (Brennan, 2010), can be calculated using
Equation 2.

\[
ICC (C, 1) = \frac{MS_A - MSR_{\times A}}{MS_A + (n-1)MSR_{\times A}},
\]  
\text{Equation (2)}
where MSA = mean square for assignment effect; MSA×R = mean square for interaction of rater and assignment; n = the number of raters per assignment. This form of ICC is also known as norm-referenced reliability and as Winer’s adjustment of anchor points (Winer, 1971).

It is noteworthy that although various forms of ICC seem to increase its popularity for estimating reliability, the assumption that there is a common population variance for all measurement conditions ought to be met (McGraw & Wong, 1996). Without satisfaction of this assumption, calculating any form of ICC is inappropriate. Also, the reliability calculated from any form of ICC is for the whole peer rating setting and not for scores given by a particular peer rater. This is also the conceptual framework and assumption of Generalizability theory (Brennan, 2010).

Considering the importance of reliability in assessment, and the common concern for possible low reliability in peer rating, it is surprising to find that few studies have examined peer rating reliability. A search through ERIC and ProQuest databases from 1990 to 2016 of peer reviewed journal only found nine prior studies examining the reliability of peer ratings (Chang, Tseng, Chou, and Chen, 2011; Cho, Schunn, & Wilson, 2006; Haaga, 1993; Luo, Robinson, & Park, 2014; Magin, 2001; Marcoulides & Simkin, 1995; Sung, Chang, Chang, & Yu, 2010; DeWever, Keer, Schellens, & Valcke, 2011; Zhang, Johnston, & Kilic, 2008). One possible reason for this relative absence of literature is that the main concern of most researchers and educators has been the degree of agreement between the scores given by peers and by instructors. Thus, many studies actually investigated the validity of peer assessment rather than the reliability. Yet, this absence of reliability measurement can undermine the research findings of peer rating validity (Luo, Robinson, & Park, 2014) because acceptable reliability is a necessary but not a sufficient prerequisite for assessment validity (Raczynski, Cohen, Engelhard, & Lu, 2015).

Mixed results, using different metrics, have been found in the existing the nine studies. The most common approach is using Generalizability theory. Sung, Chang, Chang, and Yu
(2010) used the GENOVA computer program to estimate the reliability of peer rating on the performance of seventh graders playing musical instrument (N=116). When the number of peer raters per assignment was three, they observed a reasonable reliability (the generalizability coefficient =0.84; the index of dependability=0.81). Similarly, Zhang, Johnston, and Kilic (2008) examined the reliability of rating peers’ contributions in group projects with a sample of 134 students enrolled in a teacher education program out of three classes. They found that rater effect was the weakest variance component, and it never accounted for more than 12% of the total variance.

Also under the framework of G theory, Marcoulide and Simkin (1995) examined the reliability of peer rating from a sample of 60 undergraduates by calculating the percentage of variance. Their results showed that 69% of the variance in individual ratings could be explained by overall paper quality effect, 31% by the interaction of rater and paper quality, and 0% by rater effect. The results indicate reliable peer rating because the main source of variability is the quality of the paper (i.e. the objective of measurement) rather than the interaction between peer rater and paper. Cho, Schunn, and Wilson, (2006) looked at ICC to investigate the reliability of peer rating for a sample of 708 students across 16 different courses from four universities. They reported that ICCs ranged from 0.17 to 0.56 (M=0.39, SD=0.16) for graduate courses, and from 0.20 to 0.47 (M=0.34, SD=0.09) for undergraduate courses. The authors drew a conclusion of acceptably reliable peer rating by comparing the above result with a benchmark for the overall reliability of peer evaluation of professionals in peer-reviewed journals (M= 0.27, SD=0.12) reported in Marsh and Ball’s (1989) meta-analysis of various journals. Using a similar approach, Luo, Robinson, and Park, (2014) examined the reliability of peer ratings in a MOOC (N=1825) offered by the Pennsylvania State University on the Coursera platform. They reported a single measure ICC of 0.262 which indicated “low inter-rater agreements” among five randomly selected peer raters when rating the same assignment; and an Average Measure ICC of 0.64, which indicated
“moderate strength” of agreement, when the mean of five individual peer raters’ score was used as an index of measurement. The authors suggested using at least three peer raters to generate moderately reliable peer ratings scores.

Hagga (1993) reported a relatively high Pearson product-moment correlation between pairs of graduate students (N=45) assessing common papers (r =0.55). Using different approaches, a similar result of reliable peer rating of an individual student’s contribution in collaborative group work was observed by DeWever, Keer, Schellens, and Valcke (2011) and Magin (2001).

An opposite result was observed by Chang, Tseng, Chou, and Chen (2010) in a peer assessment involving 72 high schoolers taking a computer course. Kendall’s coefficient of concordance was used to measure inter-rater reliability, and Homogeneity analysis was used for intra-rater reliability. This peer rating was carried out on a group-to-group basis in which each assignment was rated by six students from another group. The non-significant Kendall’s coefficients of concordance indicated that there was no overall agreement among six raters on the same assignment. For homogeneity analysis, two-thirds of the assignments failed to achieve significant level (alpha=0.05) which implies “a lack of consistency” within a rater on various assignments.

In sum, reliability in peer rating can be measured either by the degree of agreement between scores given by different peer raters on the same assignment (inter-rater reliability) or by the extent to which peer raters consistently apply the same rating criteria to assess different assignments (intra-rater reliability). Most research on peer rating examined inter-rater reliability, and only one of the nine previous studies examined intra-rater reliability (i.e. Chang, Tseng, Chou, & Chen, 2011). Based on designs and the matching patterns of peer raters and peer ratees (e.g. the number of peer raters for one assignment; different vs. same group of peer raters assessing the same set of assignment; peer ranking vs. peer grading with numerical score on
continuous scales), various metrics have been used to calculate peer rating reliability, but most of them under the framework of Generalizability theory (Cho, Schunn, & Wilson, 2006; Luo Robinson, & Park, 2014; Marcoulides & Simkin, 1995; Magin 2001; Sung, Chang, Chang, & Yu, 2010; Zhang, Johnston, & Kilic, 2008). When peer rating is conducted for assessing an individual student’s contribution or performance in a collaborative team project, previous studies consistently showed that students can produce reliable rating scores (Magin, 2001; Zhang, Johnston, & Kilic, 2008). However, when peer rating is used to assess an individual’s assignment, existing literature showed mixed results (Haaga, 1993; Chang, Tseng, Chou, & Chen, 2011; Cho, Schunn, & Wilson, 2006; Luo Robinson, & Park, 2014; Marcoulides & Simkin, 1995; Sung, Chang, Chang, & Yu, 2010). Both inter-rater reliability and intra-rater reliability in previous studies are estimated for the scores of peer assessment as a whole, rather than for that of a particular peer rater.

2.2.1.2. Validity

According to the 2014 Standards for Educational and Psychological Testing, validity is “the degree to which evidence and theory support the interpretation of test score for present uses of test” (AERA/APA/NCME, 2014, p.11). One common goal of using peer rating is to relieve the heavy workload of grading assessments on instructors. Within this goal, validity in peer assessment is the degree of agreement or similarity between the scores assigned by peers and those assigned by instructors (Falchikov & Goldfinch, 2000). Validity assessment provides validity evidence based on the relationship with other variables (AERA/APA/NCME, 2014), and may be viewed as an evaluation of the utility of peer assessment. The assignment scores given by peers and given by the instructor are meant to assess the same constructs, so the relationship between peers’ scores and the instructor’s score for a given assignment can be viewed as
convergent validity evidence. Instructors’ ratings are used as the ground truth for validation of peer ratings. Therefore, the validity of peer ratings can be measured by the similarity between the scores recorded by peers and by instructors, although the instructor rating itself has been found to be problematic (Falchikov, & Goldfinch, 2000; Falchikov & Magin, 1997; Newstead & Dennis, 1994) with the possibility of scores that are neither very reliable nor valid indicators of achievement (Guilford & Fruchter, 1973). Instructors, as content experts, are considered high quality raters, and thus instructor ratings seem to be the most reasonable standard for peer rating validation (Chang et al., 2011), especially when peer rating is used as an alternative solution to reducing instructors workload in large-sized traditional classes and MOOCs.

Validity index of peer rating is a measurement of the degree of agreement between two judges: the instructor and the peer (Bouzidi & Jaillet, 2009). The index can be measured as “a linear association between the vector of instructor ratings and the vector of mean student ratings, where the rating of each paper is one element of each vector” (Cho, Schunn & Wilson, 2006, p.895). The most commonly used method is Pearson product-moment correlation between the arithmetic mean of peer ratings and the instructor rating (Luo, Robinson, & Park, 2014; Bouzidi & Jailllet, 2009; Tsai, 2013). Higher positive correlation indicates better validity.

Although the Pearson product-moment correlation has advantages over percentage agreement measures due to its robustness against distributional pattern of rating (i.e. whether some rating are especially common; Cho, Schunn, & Wilson, 2006), it cannot detect systematic deviation. In other words, even if the scores given by peers are very high or very low compared to the scores given by the instructor, as long as the deviation is systematic, the Pearson product-moment correlation coefficient can still be extremely high. To overcome this limitation, independent t-tests have been used together with Pearson product-moment correlation to measure peer validity (Chang, Tseng, Chou, & Chen, 2011; Cheng & Warren, 1999; Stefani, 1994).
Pearson product-moment correlation, when accompanied by a t-test, provides measures of the degree of agreement between peer ratings and teacher ratings. If one is interested in knowing whether one peer rater can perform more valid or accurate peering ratings than another, the validity index for each peer rater is appropriate. Because the similarity of peer given scores and instructor given scores can provide evidence of convergent validity and criterion-related evidence, the validity index for each peer rater can be calculated by the difference in score given by peers and the instructor on the same assignment. Therefore, the validity index for each peer rater can also be described as an accuracy index.

Although existing literatures do not provide a formula to calculate validity index for an individual rater in peer rating, several measures of rater accuracy may be modified to calculate a validity index for each individual peer rater in peer assessment. Instead of measuring the similarity between the instructor scores and peer scores directly, rater accuracy has been examined from an opposite or indirect approach by measuring the difference between instructor scores and peer scores. One measure taking such approach is the root-mean-squared distance (RMSD) between the peer ratings for a given paper and the instructor rating for that paper (Cho, Schunn, & Wilson, 2006). RMSD is the square root of the sum of the squared differences divided by the number of peer ratees/assignments minus 1 to produce an unbiased estimate. It is computed in the following formula

\[
\text{RMSD}_k = \sqrt{\frac{\sum_{i=1}^{r} \left( \frac{\sum_{j=1}^{d} (R_{ij} - R_{ij(k)})^2}{r-1} \right)}{d}}
\]

where \(d\) is the number of items (or domains); \(r\) is the number of ratees (or performances), \(k\) is the subscript referring to the kth rater, \(R\) refers to the observed rating, and \(T\) refers to the True score. When this formula is used in peer assessment, \(R\) is the score given by peer rater \(k\), and \(T\) is the score given by the instructor. A value of RMSD of zero denotes perfect accuracy, indicating that
a peer rater and instructor gave the same score for each assignment rated by that peer rater. A larger value indicates less accuracy or less valid peer rating.

Contrary to the limited number of studies on the reliability of peer rating, there is a large body of research investigating peer rating validity. Although a few studies showed significant discrepancies between peer and instructor assigned scores (Chen, 2010; Mowl and Pain, 1995; Korman & Stubblefield, 1971; Oldfield & Macalpine, 1995), the majority of studies found moderate to high consistency between peer scores and teacher scores (e.g., Chen & Tsai, 2009; Papinczak, Young, Groves, & Haynes, 2007; Salder & Good, 2006; Tseng & Tsai, 2007; Sung, Chang, Chiou, & Hou, 2005; Tsai & Liang, 2009; Gadbury-Amyot et al, 2003; Billington, 1997; Kelmar, 1993; Stefani, 1994).

Falchikov and Goldfinch (2000) conducted a meta-analysis on the validity of peer assessment from 1959 to 1999 by examining 48 quantitative studies that compared the numerical marks or grades awarded by peers and faculty in traditional classrooms. With a correlation coefficient ranging from 0.14 to 0.99, the mean overall value was $r=0.69$, which indicates strong evidence of average correlation between peer and teacher marks. Subsequently, Li et al. (2015) conducted a meta-analysis of 69 studies from 1999 to 2013 about the validity of peer assessment in online classroom environments. The results of the study indicated a correlation coefficient between peer and teacher ratings that ranged from - 0.19 to 0.98 with a moderately strong estimated average Pearson correlation of 0.63.

Evidence can also be found in Massive Open Online Courses. For example, Tsai (2013) examined the validity of peer assessment in a MOOC named “History of the World since 1300” offered by Princeton University on the Coursera platform in 2012. In this study, 170 peer graded assignments were randomly selected, and the average of a group of teaching assistants’ scores was used as the “ground truth” of accuracy. A significant positive correlation between peer’s scores and corresponding teaching assistants’ scores was found with $r = 0.42$. Another study
reported a stronger positive correlation between instructor scores and the median of peer scores in a Coursera MOOC named *Maps and the Geospatial Revolution* offered by the Pennsylvania State University in 2013, with $r = 0.619$ (Luo, Robinson & Park, 2014).

In sum, the validity of peer rating refers to the degree of agreement or similarity between the scores assigned by peers and those assigned by the instructor. It provides validity evidence based on the relationship with other variables (AERA/APA/NCME, 2014); the relationship between scores assigned by peers and the instructor for a given assignment provides convergent validity evidence. Instructor rating is used as an exterior criterion for the validation of peer ratings. Therefore, the validity of peer rating can be measured by the similarity between the scores assigned by peers and by instructors. Pearson product-moment correlation is the common measure of the overall validity of peer rating, whereas RMSD can be modified to estimate the inverse validity index for each individual peer rater. Empirical evidence of moderately high correlations between peer scores and instructor scores show that both in-person peer assessment and online peer assessment produce reasonably valid scores.

### 2.2.2. Factors that influence reliability and validity of peer assessment

Improving the reliability and validity of peer rating for summative assessment is a common concern for many educators and researchers. In addition to examining the empirical evidence of reliability and validity of scores for a summative function in peer assessment, many previous studies have investigated what potential factors can influence the reliability and validity of peer assessment. Most of these previous studies have focused on factors related to peer assessment design. This is likely due to the findings of Falchikov and Goldfinch’s (2000) meta-analysis on 48 quantitative peer assessment studies which showed that high quality studies appear to have better peer-instructor agreement than those with poor design. They explained that because
poorly designed studies involved less than clear implementation, students may have been confused about important elements of the peer rating procedure. Such confusion may result in unreliable and inaccurate rating. Based on existing literature, the controllable/manipulatable influential factors include rating method, development of assessment rubric, matching pattern between raters and ratees, non-anonymity, qualitative comments, and number of peer raters involved in each assignment. Although some uncontrollable/unmanipulatable factors, such as subject area, course level, and the nature of the assessment task, can influence the reliability and validity of peer assessment, they are not the focus of the literature review here because they cannot be controlled and manipulated by instructors or course designers to improve the reliability and validity of peer assessments.

2.2.2.1. Rating method: analytic rating vs. holistic rating

In their meta-analysis, Falchikov and Goldfinch (2000) compared the validity of peer assessments among three different rating methods: making an overall global judgment with no explicit criteria; making an overall global judgment based on several dimensions; and making judgments for each dimension separately. The results showed that the correlation between teacher rating and peer rating is significantly higher when global assessment is used instead of using ratings that consist of evaluating several individual dimensions. Further, between the two global rating methods, making an overall global judgment based on several dimensions produces more valid peer rating than making an overall global judgment with no explicit criteria.
2.2.2.2. Development of assessment criteria

Whether students are involved in the process of developing assessing criteria is an important influential factor of the validity of peer assessment. In a recent meta-analysis comparing peer and teacher ratings, Li et al (2015) found that involving peer raters in developing assessment criteria yielded much higher correlations between peer and teacher ratings than not involving peer raters in developing the criteria. This finding is consistent with results reported by Falchikov and Goldfinch (2000) that student-derived and agreed criteria are associated with better teacher-peer agreement than teacher-supplied criteria. Being involved in developing a rubric can promote a sense of ownership and autonomy (Falchikov, 1995) and increase personal responsibility and motivation (Topping, 1998), which in turn may improve the quality of the learning process and assessment results. Through such involvement, students may also gain a better understanding of the rating criteria and have the opportunity to apply them to peer assessment activity more easily (Orsmond, Merry, & Reiling, 1996).

2.2.2.3. Qualitative Comments

Peer ratings were found to more closely resemble teacher ratings when peer raters provided both scores and comments than when peer raters provided only numerical scores (Li, et al., 2015). Compared to giving a purely quantitative grade of their peers’ work, students feel more comfortable providing qualitative feedback (Cestone, Levine, and Lane, 2008). Giving qualitative comments requires peer raters to put more effort into reflecting on their peer’s work; such reflection processes can provide peer raters a clearer rationale for their numerical rating, which may in turn enhance the reliability and validity of peer ratings (Avery, 2014; Li, et al., 2015).
2.2.2.4. Matching patterns between raters and ratees

Peer ratings better resemble teaching ratings when peer raters and ratees are matched at random than when the matching is not random (Li, et al., 2015). Peer raters who are similar by a variety of measures, including gender, race, prior knowledge, and rating experience, may hold similar biases that influence their peer ratings (Dingel & Wei, 2014). Randomly matching raters and ratees is reasonably helpful to reduce certain systematic biases, and thus leads to more accurate peer ratings. When raters and ratees are matched at random, the negative effect of “friendship” and “hostile attitude” is less likely to occur, and in turn may lead to more objective and valid peer rating scores.

2.2.2.5. Anonymity

There are several advantages of anonymity. For example, students feel more comfortable doing peer rating anonymously, and anonymous environments can eliminate the negative influence of friendship on peer assessment and can lead to fairer and more honest rating (Joinson, 1999). However, such advantages do not necessarily lead to higher reliability and validity. Based on the results of a meta-analysis of peer assessments in 69 recent studies, non-anonymous peer ratings showed higher degrees of agreement with teacher ratings than did anonymous peer ratings (Li, et al., 2015). The authors explained that “non-anonymity may lead peer raters to take peer assessment more seriously and thereby generating more accurate ratings.” (Li, et al. 2015, p.12).
2.2.2.6. The number of peer raters involved

According to the spirit of the Spearman-Brown prophecy formula and the conceptual framework of Generalizability Theory, multiple ratings is better than a single rating, and more raters usually lead to higher reliability. In their quasi-experimental study, Xiao and Lucking (2008) reported that twenty peer raters produced a higher reliability of assessment scores than three peer raters in an online peer assessment environment with 232 college students. However, in real peer assessment activities, having more peer raters rate one assignment typically means that each peer rater needs to rate more assignments, indicating a heavier workload of assessing for each peer rater. For practical concerns, students should not be expected to rate many assignments. Cho, Schunn, and Wilson (2006) examined the validity and reliability of peer assessments of writing assignments with 708 students, and found that six peer raters yielded higher reliability than three or four peer raters. Luo, Robinson, and Park (2014) investigated the reliability and validity of peer assessment in a MOOC, and found that the use of at least three peer raters is necessary to achieve a moderate reliability score.

Based on these aforementioned influential factors, instructors or course designers may enhance the reliability and validity of peer assessment by using holistic assessment, involving students in the development of rating criteria, integrating peer rating with qualitative feedback, randomly matching raters and ratees, keeping peer raters non-anonymous, and optimizing the number of peer raters involved in each assignment. However, all of the factors can only explain a small proportion of variations in the agreement between peer ratings and teacher ratings (See the meta-analysis conducted by Li et al, 2015 for more details), which indicates a need to investigate other potential factors.
Chapter 3

Review of the literature: Peer rater quality characteristics

Regardless of how carefully constructed, the reliability and validity of scores based on a rating scale is critically dependent on the raters who operate it (Overall and Magee, 1992). To improve the reliability and validity of peer assessment, we need to consider the effect of peer raters’ quality characteristics. In the summative function of peer assessment, the most serious concern is whether peer ratings have similar reliability and validity to teacher ratings. Different features of peer rating and teacher rating may be important influential factors in the reliability and validity of peer assessment.

By definition, the largest difference between peer rating and teacher rating is the quality of the raters. In the case of teacher ratings, the raters are teachers who are believed to be experts in the assessment domain and have rich experience in rating. In the case of peer ratings, on the other hand, student raters are novices both in the content knowledge domain and in assessment skills. Student raters may also lack motivation to seriously assess their peers’ work, which may result in meaningless assessment results. In order to provide reliable and valid peer assessment results, peer raters need (1) good content knowledge, in order to understand their peer’s work, (2) good rating skills, in order to make accurate judgment about their peer’s work, and (3) enough motivation to make a responsible and serious judgment. Therefore, it is reasonable to propose that three potential determinants of the reliability and validity of peer assessment outcomes are a peer rater’s content knowledge in assessment domain, training experience in peer rating, and motivation for making a serious rating.
3.1. Effect of peer rater’s content knowledge in assessment domain

Comprehensive content knowledge in the assessment domain is one of the most basic requirements of a qualified rater in almost all assessment situations. Without enough content knowledge, all judgments are no more than guessing, and meaningful feedback is impossible. In most situations, raters are experts in the assessment domain. However, in peer assessment, student raters are novices in the content knowledge domain. Differences in knowledge and cognitive process between expert and novice have been studied in several professional domains, including, education, medicine, the arts, accounting, and computer programing (Adelson & Soloway, 1988; Groen & Patel, 1998; Lawrence 1988; Schenk, Vitalari, & Davis, 1998). These studies provide insight into the knowledge structure and cognitive processes that underlie the difference between novice and expert performance, and confirm that such differences occur across a variety of disciplines (Schenk, Vitalari, & Davis, 1998).

Experts differ from novices in both quantity and organization of knowledge (Chi et al., 1981; Schunk, 2008). Researchers have estimated that expertise requires an enormous amount of knowledge, which is about 50,000 chunks of knowledge. A chunk is a group of meaningful information which can be treat as one unit in memory. In other words, there are more nodes (concepts) and links (relationships) stored in experts’ long term memory (LTM) than in novices’ long term memory (Ross, 2006; Spelke & Kinzler, 2007). Experts also have better hierarchical organization of knowledge and are more likely to organize knowledge in hierarchies, whereas novices often demonstrate little overlap between concepts in a specific domain (Larkin, McDermott, Simon, & Simon, 1980).

These differences in quantity and organization of domain knowledge lead to different cognitive processes as well as to different task performances. According to information processing theory, knowledge that a person is currently thinking about in their working memory
is called active knowledge (Schunk, 2008). A concept can be activated either by direct activation, which occurs when a stimulus encounters the environment that directly corresponds to one stored in LTM; or by spread activation, meaning that concepts can be activated through the link of activation. When one concept is activated, that activation will spread down the links to other concepts. Activating one concept leads to the activation of other concepts that are connected to it. The stronger the association between two concepts, the less the cognitive resource required to activate the associated concepts. A greater amount of knowledge stored in experts’ LTM can lead to more direct activation. Additionally, in experts’ knowledge networks, more strongly associated concepts have links requiring less input to reach threshold levels of activation, which allow for a greater number of concepts to be activated together at the same time. Therefore, it is easier for experts to retrieve knowledge from long term memory than it is for novices. Because of the greater amount of knowledge and better organized LTM structure, experts are able to access information more efficiently and have a better performance. In contrast, due to the incomplete knowledge base and the poor LTM structure, novices have less direct activation, require greater efforts to process information, and tend to be more error-prone (Chi, Glaser, & Rees, 1982). In other words, novices have less active knowledge in working memory when performing a task and are more likely to commit errors.

In academic assessment, it can be reasonably surmised that expert raters (teachers) will generate more valid assessment outcomes than novice raters (students). As has been stated, the most obvious difference between teacher raters and peer raters is their knowledge level in the assessment domain. Compared to teachers (i.e. the expert raters), peer raters are novices in the knowledge domain, and their knowledge of a specific domain might be limited to what the teacher has covered during the class. Nonetheless, the content domain of assignments or exams for peer assessment is usually within the range of content knowledge covered in a particular course. Students are often encouraged to learn not only from the teacher, but also from other more
It is possible for top performing students to master all of the content knowledge covered in an exam or assignment and thereby have enough content knowledge to rate their peers’ work. Further, some gifted students may go beyond the content knowledge covered in class and spend extra time developing their own expertise in the specific domain. A problem is presented by the dramatically varied knowledge levels among peer raters. For mediocre or poorly performing students, even completing the assignment or exam can be a challenge, making sound evaluation of others’ work an obstacle. One possible result of randomly matching peer raters and ratees is to have some of the ratees being rated by a group made of entirely mediocre or poorly performed peers. In such case, the peer scores assigned for these ratees are virtually meaningless. It is reasonable to believe that peer raters with high content knowledge levels should produce more reliable and valid peer ratings.

There has been little direct research on the role of a rater’s content knowledge in the assessment domain in the reliability and validity of peer assessment results. Research that compares different types of raters has yielded indirect evidence. Kaufman, Gentile, and Baer (2005) conducted a study to examine novice versus expert ratings in creative writing. In this study, the novice raters were gifted high school creative writing students who were very knowledgeable and interested in the domain being assessed. The expert raters comprised three groups: cognitive psychologists, creative writers, and teachers. Both gifted novice raters and expert raters were asked to independently rate a set of 27 short stories and 28 poems for creativity using a scale of 1 to 6. These gifted novices produced ratings with nearly as high an interrater reliability as did the experts. Additionally, the gifted novices’ ratings were highly correlated with the creative rating of experts ($r = .78$ for poetry and $r = .77$ for short stories, $p < .001$). The result provides supporting evidence for the quality of the gifted novices’ aesthetic judgment, and
empirical evidence that when it comes to overall ratings of the creativity of student writings, gifted student writers are a decent substitute for experts.

However, the gifted novice raters in Kaufman, Gentile, and Baer (2005)’s study are not the same as true novices; they are on their way to becoming experts. It may be inappropriate to generalize the results of this study to all levels of student raters. In order to understand the rating quality of typical novice raters, one may also need to look at the evidence from a comparison of experts and non-gifted novices. The same research team conducted another study that examined judgments on the creativity of 205 poems by experts and non-gifted novices comprised of typical college students (Kaufman, Baer, Cole, & Sexton, 2008). The results showed that the college student raters had a lower amount of agreement among themselves (i.e., lower interrater reliability) than the expert raters. Also, the college students’ ratings did not match those of the expert ratings, yielding a correlation of only .21.

Instead of comparing only gifted novice ratings and expert ratings (Kaufman, Gentile, & Baer, 2005) or true novice ratings and expert ratings (Kaufman et al., 2008), Plucker, Kaufman, Temple and Qian (2010) investigated the ratings of true novices, amateurs and experts on movies released from 2001 to 2005. They found that experts (i.e., professional movie critics) gave lower ratings than amateurs (i.e., self-reported novices who have a greater interest in movie than most novices), who in turn provided lower ratings than true novices (i.e., college students). In other words, compared to true novice ratings, amateur ratings are closer to expert ratings. Further, true novices who had seen relatively fewer movies appeared to have especially high ratings and greater variation in their ratings compared to other groups. True novices with less content knowledge or relevant experience had the largest rating discrepancy from experts, and the lowest interrater reliability.

Based on indirect empirical evidence, differences in content knowledge, and differences in the cognition processes between experts and novices, it is reasonable to hypothesize that peer
raters with more content knowledge in assessment domain will produce assessment scores with higher validity using teacher scores as an external criterion, and higher interrater reliability. In order to give accurate judgment of peers’ work, peer raters need to have good content knowledge related to the work under assessment, otherwise they cannot tell what is correct and what is wrong, let alone identifying the origin of any mistakes and providing valuable feedback. In other words, good content knowledge is the most important and basic requirement for making a reliable and valid peer assessment.

Although there is little literature about the effects of raters’ content knowledge or expertise in assessment domain on intra-rater reliability of rating outcomes, raters who have enough content knowledge can make judgments by applying their knowledge. Raters with little to no content knowledge, however, need to make judgments by guessing, which may result in very inconsistent result (i.e., poor intra-rater reliability). Additionally, experts have more judgments at the level of automaticity, which would produce consistency from one rating to the next. Novices would not have such automaticity, and would need to process the information from rating to rating, leading to potential sources of inconsistency in these areas.

3.2. Effect of peer rater’s training experience

Aside from content knowledge in assessment domain, peer raters’ assessing training experience may be another determinant of the reliability and validity of peer assessment results.
One survey of 280 college students showed that 70% of undergraduates had never been involved in peer assessment activities (Wen & Tsai, 2006). According to another survey study of 1740 college students (Liu, 2005), roughly 63% of the students reported that they had never, or had rarely, conducted peer assessments. These results indicated that most students lack peer assessment skills before participating in peer assessment, which is another important reason why the reliability and validity of peer assessment may be compromised (Svinicki, 2001).

It is vital to a reliable and valid peer assessment that raters clearly understand the scoring rubric and how their peers’ performance will be rated (Falchikov, 1995). However, students may have difficulty understanding the scoring rubric, or they may interpret terms used in the rubric differently than do the teachers (Orsmond, Merry, & Reiling, 1996). It is possible that some students with good content knowledge in the assessment domain who make responsible judgments of their peers’ work can still fail to provide reliable and valid rating results due to a lack of rating skills, and a misunderstanding of the rubric. Therefore, it is crucial that assessing criteria and performance expectations are transparent and clearly communicated prior to peer assessment (Liu & Li, 2014).

Although providing training in assessment skills to peer raters has been attempted by many researchers (e.g. Brand-Gruwel, van Merriënboer, & Sluijsmans, 2002; Falchikov, 1995b; Mike & Tim, 1997; Orsmond, 1996; Saito, 2008; Liou & Peng, 2009; Topping, 1998; Xiao & Lucking, 2008; Yurdabakan, 2012), the effectiveness of training has yet to be fully examined, and studies of its effects on the reliability and validity of peer assessments are sparse. In existing literature, the content and format of peer assessment skill training varies depending on the specific purpose of the training and the focus of the particular assessing skills. Sluijsmans, Brand-gruwel, and van Merrienboer (2002) identified three main peer assessing skills: (1) defining assessment criteria and referring to the product or process, (2) judging the performance of a peer and identifying their strengths and weaknesses, and (3) providing feedback about a peer’s
performance for future learning. Based on these identified skills, Xiao (2008) defined peer-assessment skill training as a “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” (Xiao, 2008, p. 45).

To improve reliability and validity, peer assessment training content usually includes two parts: discussing the rubric to reach a general consensus; and practice, by rating sample performances or assignments. Such training has been found to attenuate extreme differences between peer ratings and instructor rating (Liu & Li, 2014), to increase the self-consistency of peer raters (Weigle, 1998; Wiggleworth, 1993), to enhance students’ attitudes toward peer assessment (Xiao, 2008), to improve students’ task performance (Saito, 2008; Sluijsmans, Brand-Gruwel, & van Merriënboer 2002) and to reduce rater biases (Weigle, 1998).

In Sluijsmans, Brand-Gruwel, and Merrienboer (2002)’s study, for example, 93 student teachers were randomly assigned to control groups and experimental groups. Before engaging in assessing their peers’ video recorded teaching performances, the experimental groups received a four-hour training with a focus on defining performance criteria, giving feedback, and writing assessment reports. Instead of completing the training tasks, the control groups attended four extra hours in the pedagogic domain to elaborate on certain aspects of creative learning. As a result of the training, students in the experimental groups surpassed students from the control groups not only in quality of peer assessment skills, but also in all of the end products of the course which include a collection of assignments, a group report, an individual report, and a creative project. The findings indicate that training may lead to an improved performance in the content domain of designing creative lessons. According to the responses of a self-reported questionnaire, students in the experimental groups were more satisfied with the peer assessment activity and course design. The result also indicates that the experimental groups showed better
assessment skills than the control groups in terms of using the scoring criteria more effectively, giving more constructive comments, and using fewer simple words.

If training can help peer raters gain better assessing skills, then the acquired skills should lead to more reliable and valid peer rating outcomes. To investigate the effects of training on peer assessment validity, Liu and Li (2014) conducted a study in a technology-facilitated peer assessment with 78 college students. During the training, peer raters watched a video that summarized the key elements of the rubric, and then discussed the conceptual terms that were used in the rubric with the instructor to make sure that the terms conveyed the same meaning to the students and the instructor. Peer raters also graded two sample projects composed by former students, and then compared their own ratings with teacher ratings in group discussion and class discussion formats. Data were collected in the form of students’ ratings of the sample projects both before and after the training. The results showed that the assessment training led to a significant decrease in the discrepancy between student ratings and instructor ratings of the sample projects across all eight rating criteria. They also found that the students who received training provided more accurate and helpful written feedback to their peers. Students with less discrepancies between their ratings and the instructor ratings tended to perform better on the final, revised project, indicating that training may enable students to become better assessors and better at receiving assessment.

Previous studies also found that training may lead to higher inter-rater reliability of peer assessment. In a quasi-experimental study of 473 undergraduate students that investigated the effects of peer assessment skill training on students’ writing performance and peer assessment quality, Xiao (2008) reported that students who were in experimental groups and received either target-criteria-based peer-assessment training or principle-based peer assessment training generated scores with higher inter-rater reliability than those who did not receive training but had an 80-minute logical peer assessment introduction and practice. In addition, the findings
confirmed that peer assessment training has positive effects upon students’ writing performance, student assessment skills, and students’ attitude toward peer assessment.

Despite the empirical evidence on the positive effects of training on reliability and validity of peer assessment, most of these training methods are collaborative and instructor-orientated and consume precious lecturing time in the classroom. Instructors may have difficulty allotting enough time to training activities in their classes (Satio & Fujita, 2004). Self-paced training seems more versatile because it involves raters working through the same training content at their own pace, it can be done outside of the classroom, and it generally requires less time (Johnson, Penny, & Gordon, 2009). Yet, there is little research examining the effectiveness of self-paced training on the reliability and validity of peer assessment. A piece of indirect evidence has been found in a study that compares the effectiveness of self-paced and collaborative training on the rating accuracy of experienced raters in a large-scale writing assessment (Raczynski, Cohen, Engelhard, & Lu, 2015). The results suggest that the self-paced method was equivalent in effectiveness to the more time-intensive and costly collaborative method.

Based on the existing literature, it is reasonable to propose that self-paced training might also lead to higher validity and higher inter-rater reliability of peer assessment. Although there is little to no research that examines training effects on the intra-rater reliability of peer assessments, several studies with non-student raters have shown that rater training is successful in helping inexperienced raters to give more predictable scores and increase the intra-rater reliability (Weigle, 1998; Weigle, 1994). Student raters in peer assessments are also inexperienced raters, therefore it is more likely that peer assessment training may increase intra-rater reliability of scores generated by peer raters
3.3. Effect of peer rater’s rating motivation

In order to improve the reliability and validity of peer assessment, most studies have focused on topics such as peer assessment procedure, rating format modification, and peer raters’ cognitive processes. These approaches merely assume that peer raters are inherently motivated to rate seriously. In reality, however, peer raters may have little motivation to provide serious and meaningful ratings, because instructors often use peer rating to determine peer ratees’ grades without providing peer raters any feedback about their rating quality, or holding peer raters accountable (Friedman, Cox, & Maher, 2008). Using the framework of expectancy-value theory, Chen and Lu (2004) argued that if peer evaluations are used for purposes that students do not value, or if they (peer raters) see no visible result from their participatory efforts, they will cease to give meaningful input.

From the perspective of expectancy-value theory, a person’s motivation is chiefly determined by their belief about how well they will do on an activity (i.e. expectancies) and the extent to which they value the activity (i.e. values). Theorists in this tradition attempt to use a person’s “expectancies” and “values” to explain their choice, persistence, and performance (Atkinson 1957; Vroom, 1964; Wigfield & Eccles, 1992). According to expectancy-value theory, motivation consists of three components: ability beliefs, expectancies for success, and achievement values. Ability beliefs refers to the estimations of one’s current competence at a given activity (Wigfield & Eccles, 2000). Thus, ability beliefs focus on present ability, whereas expectancies of success focus on future performance and outcome. Achievement values are defined in terms of four different components: attainment value (i.e. the importance of doing well on a given task), intrinsic value (i.e. enjoyment one gains from doing the task), utility value (i.e. how a task fits into an individual’s future plan), and cost (i.e. perceived cost of engaging in the activity, such as time taken away from other activities, amount of effort needed to accomplish the
activity, and emotional cost associated with the activity). People feel motivated when they believe that they have the ability needed to complete a given task, that their efforts will lead to some expected outcomes, and that the outcomes are personally valued (Isaac, Wilfred, & Douglas, 2001). Individuals may be confident about their ability and expect to succeed, but if they do not value the task, they might not choose to engage in the task. Similarly, people may believe that a task is valuable to them, but if they feel that they are unable to do the task, they may not choose to engage in the task, either.

Under the framework of expectancy-value theory, the most commonly used approach to promote task motivation is to provide an intervention that targets an individual’s perceived value of a given task (Brophy, 1999; Gaspard et al., 2015). Several studies have shown that such relevant interventions are effective tools to enhance motivation. For example, Gaspard et al. (2015) conducted an experiment on 1,916 ninth graders from 82 classrooms to examine the effects of an intervention that demonstrated the relevance of mathematics on the value beliefs of mathematics achievement. The experimental groups received an intervention consisting of a 90-minute lesson on the relevance of mathematics, which included a psychoeducational presentation for the whole class and relevance-inducing tasks for individual students. Results showed that students in the experimental groups reported higher utility value, attainment value, and intrinsic value than those in the control groups.

As previously stated, peer assessment activity is beneficial to students in many aspects such as learning outcomes, higher order thinking skills, self-awareness, social skills, assessment skills and self-efficacy. Although the various benefits of peer assessment are well known to researchers and educators, they are not known to students. Many students have negative attitudes toward peer assessment and perceive doing peer assessment merely as completing an assessment task rather than as a beneficial learning experience (Cho, Schunn, & Wilson, 2006). Not surprisingly, students may not have the motivation to assess their peers’ work if they cannot
perceive any benefit or value from the peer assessment activities. Therefore, it is reasonable to hypothesize that verbally persuading students about the values and benefits of engaging in peer assessments may increase students’ motivation, which could in turn improve the reliability and validity of peer assessment outcomes.

Another potential way to motivate peer raters to invest more effort into peer assessment is to provide them with external incentives which have high perceived value. Although there is little, if any, research investigating what kind of incentives are most valued in peer assessment, indirect evidence has been collected in the context of low-stakes tests, where there are little to no consequences or stakes attached to performance on the tests. To determine students’ preferred incentives for taking a low-stakes test in a college setting, Steedle (2010) did a survey with 2,242 freshmen from 24 colleges and universities. Results revealed that cash (i.e. being paid in money) and prizes (i.e. a chance to win an iPod) have the highest perceived value among all 11 listed incentives (e.g. recognition, feedback about strength and weakness, comparison to other students, helping school assess student learning, building resume, and being required to participate by a school administer or faculty member).

As one of the most desirable incentives, money has been frequently used to enhance task motivation and to gain cooperation in everyday life. Employers provide higher salary for better job performance; schools offer merit-based scholarships to reward and encourage excellent academic achievement; researchers pay potential participants to increase respond rate of their surveys; and even parents give children money for some specific achievements or progress. Can monetary incentives indeed increase task motivation? This issue has been under debate for decades. Some scholars believe that monetary incentives for performing a task, as well as other forms of external rewards, may be effective in getting people to perform tasks by increasing extrinsic motivation, but will undermine individuals’ long term intrinsic motivation or interest in the same or similar tasks (Deci and Ryan, 1985; Deci, Koestner, & Ryan, 1999; Danner and
Lonky, 1981). Others claim that negative effects of such external incentives is limited, and that monetary incentives can be used to increase motivation (Bandura, 1977; Cameron & Pierce, 1994; Eisnbenger & Cameron, 1996; Flora & Flora, 1999; Shunck, 1984).

Even those opponents of monetary incentives acknowledge that such a tangible reward can increase motivation to complete the current task. For example, in a laboratory experiment by Deci (1971), 24 college students was randomly assigned into one of two groups to work on a puzzle in three separated sessions. In the first session, both the experimental group and control group performed the same task in the same context (i.e. using the puzzle pieces to reproduce configurations). In the second session, each participant in the experimental group was paid $1 for each configuration he or she reproduced. During the third session, both groups were asked to reproduce more configurations, but neither group received pay. Each participant’s intrinsic motivation of solving puzzles was measured by the amount of time they spent on the activity during the free choice period. Results showed that there was no significant difference in time spent during the first session between two groups, which indicated that the two groups initially had similar intrinsic motivation in the puzzle game. However, the experimental group spent much more time in the second session than in the first session, whereas, such an increase was not observed in the control group. Thus, one can see strong evidence that money is an effective incentive for increasing intrinsic motivation. In the third session, the experimental group spent significantly less time than in the first session, whereas such decrease did not occur in the control group. Based on these findings, Deci (1971) concluded that when money is used as an external reward for an activity, the subjects lose intrinsic motivation for the activity. He also admitted, however, that money may work to “buy off” one’s intrinsic motivation for an activity.

Similar results have been reported by many other researchers in early studies (Deci, et al., 1999). These early studies appeared to offer some support to the claim that although monetary incentives are effective in increasing people’s motivation in current tasks, once such incentives
are no longer available, people’s intrinsic motivation to engage in similar activities will be undermined. A noteworthy common feature among these early experiments is that the selected tasks a subject could participate in were inherently interesting, such as word games, drawing pictures, reading magazines, and playing puzzles. People have strong intrinsic motivation to do these activities even without any incentives, because just engaging in the activities themselves is a kind of reward. However, in educational settings, most activities (e.g., taking assessments, preparing for exams, taking notes, doing homework, and completing peer assessment) are not that attractive to students. It is very likely that the effect of monetary incentives on motivation may vary depending on initial interest level of the task.

Today, researchers have recognized that the effects of monetary incentives depend on the reward contingency, how they are allocated, and the context in which they are administered (Cameron, Pierce, Banko, & Gear, 2005). As a result, the focus of the debate has moved beyond whether monetary incentives are inherently harmful or beneficial to motivation in general, and onto identifying the facilitating conditions and factors that need to be considered when designing a successful incentive program (Cameron & Pierce, 2002). Based on Cameron, Banko, & Pierce’s (2001) meta-analytic review of studies about reward and motivation, monetary incentives have been found to increase motivation on tasks that are of low initial interest. Regarding a task with high initial interest, monetary incentives can increase motivation when they are presented beforehand rather than presented unexpectedly after the activity, and when they are linked to the level of performance rather than loosely tied to participating behavior.

As an example of high initial interest tasks, in Eisenberger, Rhoades, and Cameron’s (1999) study, participants were asked to find detailed differences between cartoon drawings. One experimental group was offered $3 if they found more differences than 80% of their classmates (normative standard). The same amount of money was given to the other experimental group if they found more than four differences (absolute standard). The monetary incentive information
was delivered to all experimental participants before they began the task. The two experimental
groups were compared with two non-rewarded control groups on task enjoyment, and free time
spent on task following the withdrawal of monetary incentives. The result showed that
experimental participants had a higher level of task enjoyment and spent more free time on the
following task than did those in control groups. This indicates that when monetary incentives are
introduced beforehand, and are given for achieving a performance criterion or norm, they can
enhance people’s intrinsic motivation.

On low initial interest tasks, expected monetary incentives have been found to have a
positive effect on motivation. An experiment by O’Neil, Abedi, Miyoshi, and Mastergeorge
(2005) examined the effect of monetary incentives on twelfth graders’ motivation to take low-
stakes tests. Using 20 mathematics items released from the Third International Mathematics and
Science Study (TIMSS), the experimental group was offered $10 per correct item. The
experimental group was compared to a non-incentive control group on effort, self-efficacy, and
anxiety level on taking the low-stakes test. The result indicated that the incentive group put more
effort into the test, had higher self-efficacy for doing the test well, and experienced less anxiety
when taking the test than did those in the non-incentive group, suggesting that money is an
effective incentive for increasing people’s motivation to complete tasks that are of low initial
interest.

There is a large body of empirical evidence that shows the effectiveness of monetary
incentives on task motivation (Baumert & Demmrich, 2001; O’Neil, Sugrue, & Baker, 1996).
What is the theoretical explanation for this phenomenon? There are several ways to interpret the
effect. Based on expectancy-value theory, most people highly value money as an outcome of
performance. As long as people believe that they have the ability needed to complete a given task
and that their effort will lead to a highly valued outcome (i.e. money), they are likely to exert
more effort in task performance. From the perspective of Bandura’s self-efficacy theory, when
monetary incentives are given as rewards for achievement, they result in high personal standards. In turn, attainment of high personal standards leads to better task involvement, positive evaluations of performance, and increased self-efficacy (Bandura, 1986, 1997). Using the framework of attribution theory (Weiner, 1985), monetary incentives can directly lead to greater task involvement and interest. As task interest increases, people may see themselves as motivated by personal (internal) causes.

Based on empirical studies, existing literature, and common sense, it is reasonable to believe that offering monetary incentives to peer raters for a high quality peer rating task would lead to more motivated raters. A piece of evidence that directly supports this has been found in Salvemini, Reilly, and Smither’s (1993) study investigating the influence of rater motivation on assimilation effect and accuracy in performance ratings. In this study, 108 student-volunteers were invited to rate customer sales representatives’ job performance, as recorded in videotapes. Some raters were offered monetary incentives for accurate ratings, either before or after observing ratees’ behavior. Other raters were offered no incentives. Some raters were informed that their ratees had received either above or below average ratings in prior positions, while other raters were given no prior performance information. Raters in incentive groups were also informed that the monetary rewards would be allocated depending upon how close their rating came to the rating of experts: $200 for first place; $150 for second place; $75 for third place; $ 50 for fourth place; and $25 for fifth place. The result showed that in the absence of monetary incentives, information of the ratees’ prior performance resulted in an assimilation effect and relatively low rating accuracy. When incentives were available, information on prior performance did not influence the overall average rating or evaluation accuracy. This indicated that monetary incentives did motivate raters to exert more effort on rating more relevant aspect of ratees’ behavior, rather than to simply rely on peripheral aspects. The results also showed that raters who were informed about incentives before observing the ratees’ job performance provided the most
accurate ratings, suggesting that expected monetary incentives have more positive effects on motivation than unexpected incentives.

While monetary reward is an effective incentive to raise people’s motivation for completing a task, the elements that motivate people may be beyond the amount of money, per se. Take Amazon’s Mechanical Turk (MTurk) for example, a novel open online crowdsourcing marketplace for getting tasks done by others. Individuals are able to register as “requester” (i.e. employer) to create tasks (post jobs) or as a “turker” (i.e. worker) to browse among existing jobs and complete them in exchange for a small monetary payment set by the employer. In January 2011, there were reported to be more than 500,000 workers from over 190 countries who completed tens of thousands of tasks daily (Wikipedia). Although compensation in MTurk is monetary, the amount awarded is typically so small (e.g., 10 cents for 10-20 minute tasks) that it cannot explain such a high level of participation. In a psychological study conducted by Buhrmester, Kwang, and Gosling (2011), evidence showed that the participation of workers is affected by compensation rate and task length, but workers can still be recruited rapidly and inexpensively because many of the workers completed tasks for enjoyment rather than monetary compensation. According to Self-Determination theory (Deci & Ryan, 1985, 2000), monetary incentives based on performance level increase perceived competence; when these incentives convey freedom of action to the recipients rather than control over performance, they can enhance perceived self-determination (autonomy). Satisfying the psychological needs of competence and autonomy can result in higher intrinsic motivation.

As evidenced by the existing literature, effective incentives for motivating peer raters to seriously rate their peers’ work include providing peer raters with verbal persuasion about the values and benefits of engaging in peer assessment, and offering them expected monetary incentives before conducting peer assessments. Motivated peer raters are likely to generate more reliable and valid scores.
Chapter 4
Methodology

This chapter presents a description of the methodology and research design that was employed in the present study. The following issues will be addressed: research question and hypotheses, participants, procedure, instrumentation and measures, and data analysis methods.

4.1. Research questions and hypotheses

The overarching aim of the present study was to investigate the effect of peer raters’ quality characteristics on rater reliability and rating validity of peer assessment, and provide empirical evidence for possible interactions between these rater quality characteristics. Specifically, this study examined how peer raters’ content knowledge, training experience, and motivation to make serious and responsible judgments influence the inter-rater reliability, intra-rater reliability, and the rating validity of peer assessments of written essays. It also examined if there are interaction effects among peer raters’ content knowledge, training experience, and rating motivation. The following research questions were addressed:

1. Does peer raters’ content knowledge in the assessment domain affect the rater reliability and rating validity of peer assessment?

2. Does peer raters’ training experience lead to more reliable and valid peer assessment scores?

3. Does peer raters’ rating motivation influence the reliability and validity of peer assessment?

4. Is there an interaction effect between peer raters’ rating motivation and training experience on reliability and validity of peer assessment?
5. Is there an interaction effect between peer raters’ rating motivation and content knowledge on reliability and validity of peer assessment?

6. Is there an interaction effect between peer raters’ training experience and content knowledge on reliability and validity of peer assessment?

Unlike the commonly used approach, both reliability and validity in this study are measured for each individual peer rater rather than for the overall peer assessment. Instead of using the reliability coefficient and validity coefficient directly, the reliability and validity in this study are measured by their inverses, namely the different types of measurement errors. There are two types of reliability that are of interest in this study: inter-rater reliability and intra-rater reliability. In this study, inter-rater reliability examines the agreement between the score given by a peer rater with the mean of scores given by all peer raters across all assignments. It is measured by the average discrepancy between that peer rater’s score and the grand mean of all peer raters’ scores. This discrepancy is named “inter-rater error”. A high value of the inter-rater error corresponds with low inter-rater reliability. Intra-rater reliability refers to the consistency within an individual peer rater, and examines the extent to which a peer rater consistently applies a rating rubric to assess different assignments. It is measured by the average discrepancy between the deviation scores and the average deviation of a peer rater’s scores from the true scores. The true scores, in the present study, are the average of scores given by the two teachers. This discrepancy is called the “intra-rater error”. A high value of the intra-rater error indicates poor intra-rater reliability. Validity of peer assessment refers to the degree of similarity between the scores given by a peer rater and those given by the teacher. A peer rater’s rating validity can be measured as the average discrepancy between the peer rater’s scores and the average of teachers’ scores. If the average teachers’ score is treated as the external criterion (i.e. the ground truth), a larger discrepancy value indicates less accuracy of a peer rater. This average discrepancy is called the “criterion error” for the purpose of this study.
The most obvious difference between teacher raters and peer raters is their content knowledge level in the assessment domain. Compared to teachers (i.e. the expert raters), peer raters are novices in the knowledge domain. The knowledge level varies dramatically among peer raters, and this difference is extremely difficult, if not impossible, to eliminate. In order to provide valid and reliable peer assessment results, a peer rater needs to master the content knowledge covered in the assessment domain to understand, and then evaluate, their peers’ work. It is reasonable to believe that peer raters with a high content knowledge level should produce more reliable and valid peer ratings. Therefore, the following hypotheses are posed about the effect of peer raters’ content knowledge in the assessment domain on the reliability and validity of their peer assessment:

Hypothesis 1(a). There will be a negative correlation between peer raters’ content knowledge in assessment domain and their inter-rater error in peer assessment.

Hypothesis 1(b). There will be a negative correlation between peer raters’ content knowledge in assessment domain and their intra-rater error in peer assessment.

Hypothesis 1(c). There will be a negative correlation between peer raters’ content knowledge in assessment domain and their criterion error in peer assessment.

Another difference between peer rater and teacher/expert rater is rating experience. Students are believed to have little to no experience in rating. Providing peer raters with some training is another possible way to improve the reliability and validity of peer rating. Through a training process, peer raters may have the opportunity to gain a deeper understanding of the scoring rubric, read examples of expert ratings, practice peer rating, and compare the scores they assigned with those assigned by teachers/experts and adjusted their own standard before real peer assessments. Therefore, regarding the effect of training on the reliability and validity of peer assessment, the following research hypotheses are posed:
Hypothesis 2(a). Peer raters who received training have less inter-rater errors than those who did not receive training.

Hypothesis 2(b). Peer raters who received training have less intra-rater error than those who did not receive training.

Hypothesis 2(c). Peer raters who received training have less criterion error than those who did not receive training.

Moreover, the difference between peer raters and teachers may not only be within their cognition or competence, but also in their levels of motivation. Hanranhan and Issacs (2001) noted that peer raters may lack effort or seriousness when performing the assessment. Peer raters’ motivation affects their rating reliability and validity because it influences the willingness of the rater to make responsible judgments of their peers’ work. If peer raters have low or even no motivation, the whole process, as well as the result of peer assessment, may become meaningless. Therefore, it is reasonable to propose the following hypotheses:

Hypothesis 3(a). There is a negative correlation between peer raters’ rating motivation and their inter-rater error in peer assessment.

Hypothesis 3(b). There is a negative correlation between peer raters’ rating motivation and their intra-rater error in peer assessment.

Hypothesis 3(c). There is a negative correlation between peer raters’ rating motivation and their criterion error in peer assessment.

Because peer raters are individual persons with different quality characteristics, their rating behaviors and results might not be affected solely by an individual quality characteristic, but possibly by a combination of different quality characteristics. When peer raters lack rating motivation, they may not apply their content knowledge to make judgments of their peers’ work, which in turn might undermine the effect of content knowledge on the reliability and validity of peer assessment. Similarly, when peer raters have little rating motivation, they may not use their
rating skills obtained from training to do peer assessment, and thus may impair the effect of training on the reliability and validity of peer assessment. Accordingly, the following hypotheses are posed with regard to the possible interaction effects of peer rater’s content knowledge and training experience with their rating motivation:

Hypothesis 4(a). There is a significant interaction effect of peer raters’ rating motivation and content knowledge on inter-rater reliability of the scores they given in peer assessment. Specifically, the expected negative correlation between content knowledge and inter-rater error for peer raters who have high rating motivation is stronger in magnitude than it is for peer raters who have low rating motivation.

Hypothesis 4(b). There is a significant interaction effect of peer raters’ rating motivation and content knowledge on intra-rater reliability of the scores they given in peer assessment. Specifically, the expected negative correlation between content knowledge and intra-rater error for peer raters who have high rating motivation is stronger in magnitude than it is for peer raters who have low rating motivation.

Hypothesis 4(c). There is a significant interaction effect of peer raters’ rating motivation and content knowledge on the validity of the scores they give in peer assessment. Specifically, the expected negative correlation between content knowledge and criterion error for peer raters who have high rating motivation is stronger in magnitude than it is for peer raters who have low rating motivation.

Hypothesis 5(a). There is a significant interaction effect of peer raters’ rating motivation and training experience on inter-rater reliability of the scores they give in peer assessment. Specifically, the training effect on reducing inter-rater error for peer raters who have high rating motivation is more salient than it is for those who have low rating motivation.

Hypothesis 5(b). There is a significant interaction effect of peer raters’ rating motivation and training experience on intra-rater reliability of the scores they given in peer assessment.
Specifically, the training effect on reducing intra-rater error for peer raters who have high rating motivation is more salient than it is for those who have low rating motivation.

Hypothesis 5(c). There is a significant interaction effect of peer raters’ rating motivation and training experience on the validity of the scores they give in peer assessment. Specifically, the training effect on reducing criterion error for peer raters who have high rating motivation is more salient than it is for those who have low rating motivation.

Hypothesis 6(a). There is a significant interaction effect of peer raters’ content knowledge and training experience on inter-rater reliability of the scores they give in peer assessment. Specifically, the expected negative correlation between content knowledge and inter-rater error for peer raters who received training is stronger in magnitude than it is for peer raters who did not.

Hypothesis 6(b). There is a significant interaction effect of peer raters’ rating motivation and training experience on intra-rater reliability of the scores they give in peer assessment. Specifically, the expected negative correlation between content knowledge and intra-rater error for peer raters who received training is stronger in magnitude than it is for peer raters who did not.

Hypothesis 6(c). There is a significant interaction effect of peer raters’ rating motivation and training experience on validity of the scores they give in peer assessment. Specifically, the expected negative correlation between content knowledge and criterion error for peer raters who received training is stronger in magnitude than it is for peer raters who did not.
4.2. Research design

A 2 × 2 × 2 completely crossed factorial design was employed with two levels of training intervention and two levels for each of the two motivation interventions. Participants from 24 classes were expected to rate four Chinese essays written by their peers in a mid-term examination. Training intervention and motivation incentives were given at class level before peer assessment. In order to control trainer effect as well as to increase the external validity of the findings in this study, a self-paced training method was preferred to a collaborative training method. Motivation intervention includes two methods: (1) teachers’ verbal persuasion about the academic value of seriously completing peer assessment and (2) monetary incentives for accurate peer assessment result (More details are described later). The assignment of experimental condition was summarized in Table 4.1.

Because the control group would not receive any motivation incentives, if they knew such incentives were provided to the experimental group it would be possible for them to experience feelings of resentment and demoralization; this could lead to an abnormally low performance. In order to eliminate this undesirable resentful demoralization effect, the control group performed peer assessment one day earlier than the experimental group. Each peer rater’s Chinese mid-term examination score was used to measure content knowledge in the assessment domain of Chinese composition.
Table 4-1. Assignment of Experimental Condition

<table>
<thead>
<tr>
<th>Group</th>
<th>Class</th>
<th>Self-paced Training</th>
<th>Verbal Persuasion</th>
<th>Monetary incentives</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1001</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>1</td>
<td>1002</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>2</td>
<td>1003</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>3</td>
<td>1004</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>4</td>
<td>1005</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>4</td>
<td>1006</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>5</td>
<td>1007</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>5</td>
<td>1008</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>6</td>
<td>1009</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>7</td>
<td>1010</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>8</td>
<td>1011</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>8</td>
<td>1012</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>2</td>
<td>1101</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>2</td>
<td>1102</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>1</td>
<td>1103</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>4</td>
<td>1104</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>3</td>
<td>1105</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>3</td>
<td>1106</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>6</td>
<td>1107</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>6</td>
<td>1108</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>5</td>
<td>1109</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>8</td>
<td>1110</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>7</td>
<td>1111</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>7</td>
<td>1112</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
</tbody>
</table>
4.3. Participants

Participants were 838 students (426 males, 406 females, and 6 not specified) recruited from a total of 24 classes in a typical high school in south China. The typical class size is around 45 students. The participants were from tenth and eleventh grades, aged from 13 to 20, with an average of 16.7, a standard deviation of 1.0, and mode of 16.5. A histogram of age distribution was provided in Figure 4-1. According to the report of Sixth National Population Census of the People's Republic of China (2010), there are 56 ethnic groups in China, and the majority is Han People, which account for 91.51% of the population (Xinhua, 2011). The participants in this study are from 11 different ethnic groups: 91.53% Han, 3.63% Uyghur, 1.09% Hui, 0.97% Hani, and other 2.78%.

Figure 4-1. Histogram of Participants’ Age Distribution
4.4. Procedure

This procedure section will describe a typical Chinese language high school mid-term examination, and the process of teacher rating, the selection standard and process of essays used as training materials, as well as the final peer assessment, training method, motivation incentives, and peer assessment procedure.

4.4.1. High school Chinese language mid-term examination and teacher rating

In a typical high school in China, all students are expected to take a mid-term examination every semester. For the subject of the Chinese language, the testing format and rating rubric used is the same across all three different grades (i.e., tenth grade, eleventh grade, and twelfth grade), mirroring the testing format of the Chinese language subject in the Gaokao (National College Entrance Examination). There are typically four different parts in this Chinese language exam: application of language and character (语言文字运用), reading comprehension of classical Chinese (文言文阅读), reading comprehension of modern Chinese (现代文阅读), and essay composition (作文). Because essay composition accounts for a large proportion of the Chinese language subject examination score (60 out of 150) in the Gaokao, which is a very high-stakes examination in China, teachers use the Gaokao Chinese writing rubric (scoring standard) to guide their teaching of writing, and to evaluate their students’ essay composition from the first year of senior high school. The writing rubric is developed by a panel of experts. It includes three dimensions: Content (内容), Expression (表达), and Feature (特征). Each dimension has a maximum score of 20, and is classified into four different levels (等级). Explanatory description of corresponding writing quality and standard is provided for each level. An example of the writing rubric is attached in Appendix A.
After students turn in their mid-term exam papers, ratings are provided immediately. All Chinese teachers are involved in rating the exams within each grade, using the standard scoring rubric. To control for teachers’ prior knowledge of students’ academic performance in Chinese classes, all examinee identification information is withheld from teacher-raters. Each Chinese language teacher is assigned to rate a specific part or particular question(s) of the exam across all students and classes within one grade. Because the evaluation of essays involves subjective judgment, each essay is rated by two Chinese language teachers independently using the Gaokao Chinese writing rubric. If the difference between scores given by two Chinese language teachers for a particular essay is greater than five points, that essay is re-rated by the two teachers after a brief discussion. If the difference still exceeds five points, the essay will be presented to a third experienced Chinese language teacher to rate. The average of the two most similar scores is then the final score for each essay.

4.4.2. Selection of training materials and final peer assessment essays

This section focuses on the process used in this study to assemble not only the training materials, but also the final peer assessment essays that all peer raters rate to determine their rater reliability and rating validity. Because some essays have proved to be more difficult to rate accurately than others (Engelhard, 1996), it is important to choose essays carefully. Both training essays and peer assessments in this study were selected from Chinese language mid-term examination essays. The essays were written by tenth and elevenths grader in spring 2016. The topic of the tenth graders’ essay was about writing a letter to address one’s opinion on whether to keep the old learning and teaching building. The prompt was given as follow:

Please read the following material and write an essay with at least 800 words based on the requirement (60 point)
One school has a learning and teaching building built in late 1950s. Numerous eminent peoples, which include famous scholars, writers, and hosts, had studied in that building. Now this small crude old building cannot meet the development need of modernization. Due to its particularity, the school wants to keep this building. However, some people emphasize development and suggest knocking it down because of the lack of usable land as well as the cost of repairing the old building. Therefore, the school hesitates. This problem has attracted the attention from many teachers, students and alumni. A famous writer who had studied in this building came back to take photographs with this building and to strive for keep the building.

What is your opinion on this issue? Please write a letter to the principal of this school, the famous writer, or other related people to express your attitude to and view of this issue. Please synthesis the content of the above material, select an appropriate perspective, make an arguable claim, and complete this writing task. Please specify the receiver of the letter, and use “Xiaolin” as the sender to exclude your personal identification information.

The topic of the eleventh graders’ essay was about writing a letter to the principal to address one’s opinion on whether to keep the old learning and teaching building. The prompt was given as follow:

Please read the following material and write an essay with at least 800 words based on the requirement (60 point)

In Southern Sichuan Bamboo Forest Scenic Region, there is 1.65 acre state-owned forest being classified as carving character forest. Tourists who have need of “carving” are guided to carve in this special area. The purpose of this classification is to
prevent tourists’ graffiti behavior. As a result, about five thousand bamboos are ravaged by tourists every year.

Berlin Wall in German is covered by all kinds of strange pictures and characters. However, the various graffiti on the Berlin Wall not only chronicles a period of history, but also adds another culture. What is your opinion on these two issues? Please write an essay to express your attitude and view. Please synthesis the content and implication of the above material, select an appropriate perspective, make an arguable claim, specify a style, and name a title yourself. Do not copy. Do not plagiarize.

The original Chinese version of these two topics and sample essays are attached in Appendix B. As mentioned earlier, all essays in a mid-term examination have been rated by two Chinese language teachers independently using the Gaokao Chinese writing rubric. In order to be selected as training essays or final peer assessment essays in the study, an essay needed to be assigned the same or very similar dimension scores as well as total scores (i.e., the score difference of each dimension should not exceed two points out of 20, and the difference of total score should not exceed two points out of 60 either) by two teachers independently. If the difference of either dimension scores or total scores assigned by the two teachers on any given essay was larger than two points, the essay was not used as a training essay or as a final peer assessment essay. The average score between the two Chinese teachers for each qualified essay was used as the correct (final) score of that essay, and as the ground truth to evaluate rater reliability and rating validity of peer assessment.

After the two Chinese teachers had completed rating all of the essays, the teachers’ scores were compiled, and selections were made for training essays and final peer assessment essays using the aforementioned criteria. Ultimately, one set of eight essays was selected as training benchmarks, including four essays from tenth grade and four essays from eleventh grade. Four
training essays (i.e., two essays from each grade level) were used as training benchmarks, while the other four training essays (i.e., two essays from each grade level) were used as practicing benchmarks. Another set of four essays comprised of two essays from each grade were randomly selected as final peer assessment tasks. Because all of the essays used in this study were written by participants, four students ended up rating their own essays in the final peer assessment. The data related to these four students were excluded from this study.

For each of the eight training essays, there was a teacher assigned holistic score and analytic score for each dimension, as well as detailed explanations about why such a score was assigned. Such scoring information and explanation was presented after each corresponding training-benchmark essay; this provided a separate key sheet for the four practicing-benchmark essays, but was not available for the four essays used for the final peer assessment task. In addition to the scoring information and explanation, an instruction page about recommended procedure on completing this training was also included in each training package.

### 4.4.3. Peer rater training method

A random half of the classes at each grade level received a training package which included five parts: an introduction page, a rating rubric, four training-benchmark essays followed by scores given by experienced teachers and corresponding explanations, four practicing-benchmark essays, and a practice scoring sheet with eight multiple choice questions designed to determine if a student completed the self-paced training (See Appendix D). The participants who received training packages were asked to complete a self-paced rater training. The self-paced training proceeded through the following stages:

**Step1.** Students read the suggestion to complete training and the writing prompt on the instruction page.
Step 2. Students studied the rubric of each domain.

Step 3. Students explored the first four training-benchmark essays and the annotations/explanation of each score given to the benchmark essays in each domain. They then answered four multiple-choice questions which were designed to determine whether a peer rater had read the essays and the explanations.

Step 4. Students practiced rating the other four additional practice essays, and wrote down their rating scores on the practicing scoring sheet (See Appendix B).

Step 5. After having completed the rating of the four practice essays in Step 4, students were expected to show the rating result to their Chinese teacher to indicate that they had finished their ratings. The Chinese teacher would not hand out the key sheet (i.e. the expert’s scores and explanation) to a peer rater until the peer rater showed that she or he had completed rating the four practice essays.

Step 6. Students compared their scores assigned to the three dimensions and the total score of each practice essay with those given by teachers and read corresponding explanations. To further ensure that a peer rater in a training class had actually undergone the training, each student in training classes was asked to indicate on the practicing scoring sheet (See Appendix D) whether they overscored, underscored, or accurately scored each practicing-benchmark essay directly after comparing their scores to assigned practice essays with the expert’s scores. It took around one hour for most peer raters to complete the entire self-paced training regimen.

Why use a self-paced training method? First, previous research has shown that self-paced training methods help raters score as accurately as collaborative training methods (Raczynski, Cohen, Engelhard, & Lu, 2015). Second, because no trainers will be involved in this method, there is no worry about the possible threat of undesirable trainer effect on internal validity. Third, in real educational settings, some instructors do not include training in peer assessment due to the concern that the training process is unduly effort-demanding. Also, it is impractical for them to
allocate precious classroom instruction time to complete peer rater training. Self-paced training does not need to be completed using class time, and it allows peer raters to complete it anywhere according to their own schedule availability. This flexibility may increase the external validity of results in the present study. Finally, self-paced training is more versatile. In large size classes or online courses, the logistics of self-paced training are less cumbersome than other types of collaborative training.

4.4.4. Motivation intervention

In order to manipulate peer raters’ rating motivation level, two types of intervention were provided: Verbal persuasion and monetary incentives. About half of the classes, randomly selected, were given a short lecture (around two minutes in length) about the value of rating peers’ essays, and what academic benefit they could get by seriously performing this peer assessment. The content of the verbal persuasion lecture, translated from Chinese to English, was the following. A copy of the original lecture in Chinese could be found in Appendix C:

Dear Class,

As you know, essay composition accounts for the largest proportion of the Chinese language subject examination score in the Gaokao. Its importance is self-evidenced. I am so glad that you decided to participate in this peer assessment study, not only because by doing so you help the researchers and contribute toward building knowledge of peer assessment activities, but also because you can benefit a lot from your own engagement. First of all, you will study the Gaokao Chinese writing rubric, know what aspects the raters will look for in your essay, and understand what criteria you will need to meet in order to gain a high score. Secondly, when reading your peers’ essays, you can compare your own essay with your peers’ essays, thus become more aware of your own strengths and weaknesses. Thirdly, you can not only learn from your peers’ strengths, such as creative ideas, unique insights, well-written sentences, and new vocabulary, but can also learn from their mistakes and thus can avoid them in your own essay. Fourthly, rating your peers’ essays enable you to think of effective ways of writing high quality essays from the perspective of a rater, thus you may have a deeper understanding about how to better organize your essay structure, and how to express your ideas impressively. Finally, engaging in peer
assess your evaluation skills which are important to your later academic and career development, but are usually overlooked in our formal education. Existing research has already shown that better evaluation skills are associated with better performance of similar tasks later. Therefore, I really hope that you can put forth great effort to complete this peer assessment activity and benefit more from your serious engagement.

Similar to the intervention of verbal persuasion of task value, roughly another half of the classes at each grade level were randomly selected to be offered monetary incentives. At the beginning of this study, the motivation-incentives groups were orally informed that each essay selected for peer assessment had already been rated by experts who were very familiar with the Gaokao Chinese writing, and those incentive participants would be rewarded monetarily depending upon how close their rating scores come to the rating scores of these experts. Those participants were told that rewards would be allocated as follows: one person for the first place (i.e., the participant whose ratings are closest to the expert ratings), thirty Yuan RMB (6.87 RMB = $1); two people for the second place at twenty Yuan RMB each; and five people for the third place, at ten Yuan RMB each. To ensure the credibility of the monetary incentive manipulation, all incentive participants were given copies of a written memo stating that the money had already been given to their Chinese teachers by the researchers, and the winners would get their monetary reward from their Chinese teacher directly after all data have been collected and analyzed, which would take no more than three weeks.

4.4.5. Peer assessment

All participants received a peer assessment package, which included a copy of the Gaokao Chinese writing rubric, four selected essays, a scoring sheet (See Appendix E), and a questionnaire designed to measure rating motivation. They were asked to complete the peer
assessment in their Chinese language class, which took about 40 minutes. For those in the training groups, each participant also received a training package, and they were strongly recommended to complete training before conducting real peer assessment. In order to eliminate an undesirable, resentful demoralization effect, this peer assessment activity was introduced to participants in non-verbal persuasion classes and non-motivation-incentives classes first. After these participants handed in their peer assessment scoring sheets, those who were in monetary-incentives classes began to do their peer assessment the following day.

4.5. Instrumentation and measures

This section will focus on instrument and measure of the three independent variables (i.e., peer raters’ content knowledge in assessment domain, training experience, and rating motivation) and three dependent variables (i.e., peer rating’s inter-rater reliability, intra-rater reliability, and validity)

4.5.1. Content knowledge in assessment domain

Peer raters’ content knowledge in assessment domain was measured by the total score of their mid-term examination in the Chinese language subject, which has a score ranging from 0 to 150. The Chinese examination includes four parts: application of Chinese language and character (30 points), reading comprehension of classical Chinese (30 points), reading comprehension of modern Chinese (30 points), and essay composition (60 points). The testing format and rating rubric are the same for both tenth graders and eleventh graders.
4.5.2. Training experience

Self-paced training was provided to roughly half of the classes at each grade level. Each participant in these classes received a training package and completed the training process by himself or herself. In order to ensure that those who received training package actually completed the training, extra mini-tests and supervision were included in the training process as described earlier: (1) Four multiple-choice questions designed for testing the main ideas of the first four training-benchmark essays and explanation of each score given to these benchmark essays, (2) Chinese teachers checked each student’s practice scoring sheet to make sure that only those who had completed rating the four practice-benchmark essays could get the key sheet (i.e., the expert’s scores and explanations for the four practice essays), and (3) another four multiple choice questions asked students to indicate whether they overscored, underscored, or accurately scored each of the four practice-benchmark essays compared with the experts’ scores. The answers for these multiple choice questions were used to determine whether a student in training classes had indeed undergone the training. All of the training fidelity data were on the practicing scoring sheets (See Appendix D), which were collected immediately after completion of training by Chinese language teachers.

Peer raters’ training experience at class level was measured by whether they received the training package and indeed complete training. For the purpose of data analyses, those who were in training classes were coded as “1” for the independent variable of Self-paced Training, and those who were in non-training classes (i.e. not receiving training package) were coded as “0”. At the individual student level, the independent variable of Previous Rating Experience was measured by how many times they had rated peers’ essays before.
4.5.3. Rating motivation

Two types of motivation intervention were randomly given at the class level: Verbal persuasion about the academic value of performing peer assessment seriously and monetary incentives for accurate peer rating results. Both of them aimed to increase peer raters’ rating motivation and were used as measures of motivation at the class level. For the purpose of data analyses, the two types of motivation were scored with separate dummy codes: classes that did not receive verbal persuasion were coded as “0”, and those that received verbal persuasion as “1” on the Verbal Motivation variable. Similarly, classes that did not receive monetary incentives were coded as “0”, and those that received monetary incentive as “1” on the Monetary Motivation variable.

In order to validate the effectiveness of the two incentives (i.e., value-persuasion and monetary incentives) on students rating motivation, a 6-item motivation questionnaire based on expectancy-value theory was given to all participants (See Appendix F). All items were answered on Likert-type scales with seven ordered options, ranging from 1 to 7. Table 1 provides a summary of the three groups of independent variables and their measurements.

Table 4-2. Independent Variables and Their Measures

<table>
<thead>
<tr>
<th>Student-level Measures</th>
<th>Independent Variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rating Motivation</td>
<td>N/A</td>
</tr>
<tr>
<td>Training experience</td>
<td>The number of previous participated peer assessments</td>
</tr>
<tr>
<td>Content Knowledge</td>
<td>Mid-term Chinese language examination score (continuous scale: ranging from 0 to 150)</td>
</tr>
</tbody>
</table>
### 4.5.4. Dependent variables

The present study aims to investigate the effect of peer raters’ characteristics (i.e., content knowledge, training experience, and rating motivation) on rater reliability and rating validity of each peer rater. The dependent variables include inter-rater reliability, intra-rater reliability, and validity of each rater, which are measures by inter-rater error, intra-rater error, and criterion-error.

#### 4.5.4.1. Inter-rater error and intra-rater error

Both inter-rater reliability index and intra-rater reliability index of peer assessment in previous studies have been calculated for the scores of all raters rather than for the scores assigned by an individual peer rater. Although existing literature does not provide a formula to calculate a reliability index for an individual peer rater, several measures of rater accuracy may be modified to calculate inter-rater reliability or intra-rater reliability for each individual peer rater in peer assessment. Instead of measuring the similarity between one peer-given score and the mean of the peer scores of one assignment directly, inter-rater reliability for each peer rater can be examined from an opposite or indirect approach by measuring the difference between an
individual peer score and the mean of all peer scores; this is referred to as “inter-rater error” in the present study.

One possible formula can be developed based on the Leniency/Severity index described by McIntrye, Smith and Hassett (1984). The original purpose of the Leniency/Severity index was to measure rater leniency of an individual rater. It is:

$$\text{Leniency}_k = \sum_{j=1}^{r} \left( \sum_{i=1}^{d} \frac{R_{ij} - R_{ijk}}{d} \right),$$  \hspace{1cm} (4)

where originally, $d$ is the number of items (or domains); $r$ is the number of ratees (or assignments), $k$ is the subscript referring to the $k$th rater, $R$ refers to the observed rating, and $T$ refers to the true score, which is defined as the mean rating across all raters from the same class.

When modifying this formula to measure inter-rater error of each rater in peer assessment, $R_{ijk}$ is the score given by peer rater $k$ to domain $i$ of assignment $j$, and $T_{ij}$ is the group mean of all peer raters’ score given to domain $i$ of assignment $j$. According to this modified formula, an inter-rater error (i.e., leniency in the original formula) with an index value of zero indicates perfect accuracy, meaning that the scores given by an individual peer rater for each assignment are the same as the total mean of all peer scores for these assignments. A positive index value means that a peer rater’s assessment is severe relative to the average of other peer raters, while a negative value means the rater is lenient (Raczynski, Cohen, Engelhard, & Lu, 2015). This index is sensitive to peer raters whose severity fluctuated because the positive deviation score and the negative deviation score will cancel each other; that is, if a peer rater is severe on half the assignments, and lenient on the other half with the same degree, the leniency index for this peer rater would be zero which indicates that this peer rater appears to be neither severe or lenient, but perfectly agrees with the means of all raters on these assignments.
One solution to this potential problem is to take the absolute value of the difference between an individual peer score and the mean of all peer scores, yielding a measure called “Distance accuracy” by McIntrye, Smith and Hassett (1984) as expressed in Equation 5.

\[
\text{Distance Accuracy} = \sum_{j=1}^{r} \frac{\sum_{i=1}^{d} |T_{ij} - R_{ijk}|}{d},
\]

where \( r, d, k, R, T \), have been defined above in Equation 4. This distance accuracy index does not tell whether a peer rater is severe or lenient relative to the mean score, but neither is it subject to the problem of fluctuating severity.

An alternative solution to the issue of fluctuating severity is to use the root-mean-squared distance (RMSD) between the peer ratings for a given assignment and the means of all peer scores for that assignment (Cho, Schunn, & Wilson, 2006). The square root of the sum of the squared differences divided by the number of peer ratees/assignments minus one produces an unbiased estimate. It is computed in the following formula:

\[
\text{RMSD} = \sqrt{\frac{\sum_{j=1}^{r} \frac{\sum_{i=1}^{d} (T_{ij} - R_{ijk})^2}{d}}{r-1}},
\]

where \( r, d, k, R, T \), have been defined above in Equations 4 and 5. Similar to distance accuracy, a value of RMSD of zero denotes perfect agreement, indicating that scores given by a peer rater are exactly the same as the mean of all peer scores for these assignments. A larger value indicates less accuracy or less valid peer rating. One limitation of RMSD is that it may change original distribution of distance values by first squaring each difference value, adding up all squared distance, getting the average, and then taking the square root of it. Also, compared with the Distance Accuracy index, the meaning of the RMSD is more difficult to interpret. Therefore, the Distance Accuracy formula is preferred in the present study, and is modified to calculate inter-rater error, intra-rater error, and criterion error.
If distance accuracy is calculated for the rating score of each dimension, rather than for an overall index, a simplified version of Equation 5 can be used, as expressed in Equation 7.

\[ \text{Distance Accuracy}_{ik} = \frac{\sum_{j=1}^{r} |T_{ij} - R_{ijk}|}{r} , \]  
(7)

where Distance Accuracy\(_{ik}\) is an index of peer rater \(k\)'s inter-rater error index in dimension \(i\); \(T_{ij}\) is the mean of all peer raters’ score given to dimension \(i\) of assignment \(j\); \(R_{ijk}\) is the score given by peer rater \(k\) to dimension \(i\) of assignment \(j\).

If each assignment is graded holistically with only one global score, this formula can be further simplified as the following:

\[ \text{Distance Accuracy}_{k} = \frac{\sum_{j=1}^{r} |T_{j} - R_{jk}|}{r} , \]  
(8)

where \(T_{j}\) is the mean of all peer raters’ score given to assignment \(j\); \(R_{jk}\) is the score given on assignment \(j\) by peer rater \(k\).

Because the scores used to determine rater reliability and rating validity in the present study is the total score of each essay, rather than the scores from each dimension, Equation (8) was modified to measure the lack of inter-rater reliability of the scores given by each individual peer rater. Inter-rater reliability is concerned with the extent to which peer raters’ scores agree with each essay. The measure of a lack of inter-rater reliability may best be named an “inter-rater error”, which can be expressed as

\[ \text{Inter-rater error}_{j} = \frac{\sum_{i=1}^{n} |S_{ij} - M_{i}|}{n} , \]  
(9)

where \(n\) is the number of essays a peer rater graded. In this study, \(n=4\); \(S_{ij}\) is the score awarded on assignment \(i\) by peer rater \(j\); \(M_{i}\) is the grand mean of scores awarded by all peer raters on assignment \(i\). The inter-rater error index obtained in this formula tells us, on average, how far a peer rater’s score deviates from the group mean of all peer raters’ scores. It provides a measure of the degree to which a rater is lacking inter-rater reliability. A smaller value indicates better
agreement. Inter-rater error of zero means that each of the score given by peer rater $j$ is exactly the same as the group mean of all peer raters’ scores across all $n$ assignments.

As intra-rater reliability examines the extent to which assessment results are consistent within a peer rater on different assignments or exams. It is the consistency from rating to rating within the same rater, and can be measured inversely by the degree of inconsistency indirectly. This measure can be named as “intra-rater error”, which can be expressed as

$$\text{Intra-rater error} = \frac{\sum_{i=1}^{n} |(S_{ij} - T_i)| - \frac{1}{n} \left[ \sum_{i=1}^{n} (S_{ij} - T_i) \right]}{n},$$

(10)

where $n$ and $S_{ij}$ have been defined in above equation of Inter-rater error $j$ (Equation 9). $T_i$ is the score awarded by the teacher on essay $i$; $(S_{ij} - T_i)$ is the deviation score between peer rater $j$ and teacher on assignment $i$; $\frac{1}{n} \left[ \sum_{i=1}^{n} (S_{ij} - T_i) \right]$ is the average amount of deviation for peer rater $j$’s scores from teacher scores. The intra-rater error index obtained from this formula is the average deviation of the deviation scores for peer rater $j$ from the teacher score. It informs us of the extent to which peer rater $j$ consistently applies a rating rubric to assess their peers’ work when teacher scores are used as the ground truth. A smaller value indicates high levels of consistency. An intra-rater error of zero means that the deviation scores between peer rater $j$ and the instructor are the same across all $n$ assignments, which indicates that peer rater $j$ is perfectly consistent with the teacher when applying the rubric to assess their peer’s work.

4.5.4.2. Criterion error

In peer assessment within a classroom setting, one definition of rating validity is the degree of agreement or similarity between the scores assigned by peers and those assigned by teachers (Falchikov & Goldfinch, 2000). Because the similarity of peer-given scores and teacher-
given scores can provide evidence of convergent validity and criterion-related validity, the lack of such validity evidence can be calculated by the difference in scores given by the peer rater and the teacher on the same essay indirectly. This measure may best be named a “criterion error”, which may be expressed as

\[
\text{Criterion error}_j = \frac{\sum_{i=1}^{n} |S_{ij} - T_i|}{n},
\]

where \( n \), \( S_{ij} \), and \( T_i \) have been defined above in equation (10). The criterion error index obtained in this formula tells us, on average, how inaccurate a peer raters’ score is when using teacher’s score as the external criterion (i.e. ground truth). A smaller value indicates less discrepancy between a peer raters’ score and the teacher’s score, and thus suggests higher validity of peer assessment. A criterion error index value of zero means that each of the score given by peer rater \( j \) is exactly the same as the teacher’s score.

### 4.6. Data analysis

Because the data have the following hierarchical structure, peer raters nested within class, a series of 2-level hierarchical linear models (HLM) were fitted to peer raters’ inter-rater error, intra-rater error, and criterion error respectively. The level -1 model in the HLM consists of three parts: a sampling model, a link function, and a structural model (Raudenbush & Bryk, 2002). The sampling model for continuous data like the dependent variables in this study is a normal sampling model, which can be written as

\[
Y_{ij} | \mu_{ij} \sim NID(\mu_{ij}, \sigma^2),
\]

Meaning that level-1 dependent variable \( Y_{ij} \), given the predicted value, \( \mu_{ij} \), is normally and independently distributed with an expected value of \( \mu_{ij} \) and a constant variance, \( \sigma^2 \).
(Raudenbush & Bryk, 2002). The link function varies based on different types of distributions of the dependent variable. The link function is the identity link.

The structural model can be written as

\[
Y_{ij} = \beta_0 + \beta_1 X_{1ij} + \beta_2 X_{2ij} + \cdots + \beta_p X_{pij},
\]

(13)

where,

- \(\beta_0\) = the intercept;
- \(\beta_p (p = 0, 1, 2, \ldots, P)\) = the level-1 coefficients;
- \(X_{pij} (p = 0, 1, 2, \ldots, P)\) = the level-1 predictor \(p\) for peer rater in class \(j\).

First, a series of unconditional 2-level HLM were fitted to examine whether there is inter-class variation in inter-rater error index, intra-rater error index, and criterion error index values.

The level-1 (student-level) unconditional model is

\[
Y_{ij} = \beta_0 + e_{ij},
\]

(14)

where the intercept \(\beta_0\) is the mean inter-rater error/ intra-rater error/ criterion error index value of peer assessment for class \(j\); \(e_{ij}\) is assumed \(\sim\) independently \(N (0, \sigma^2)\) for \(i = 1, \ldots, n_j\) students in class \(j\), and \(j = 1, \ldots, J\) classes; \(\sigma^2\) is the student-level variance.

The level 2 (class-level) model is

\[
\beta_0 = \gamma_0 + u_{0j},
\]

(15)

where \(\beta_{0j}\), each class’ mean rater reliability/rating validity of peer assessment, is a function of grand mean, \(\gamma_0\), plus a random error, \(u_{0j}\); \(u_{0j} \sim\) independently \(N (0, \tau_{00})\), \(\tau_{00}\) is the class-level variance. This yields the composite model with fixed effect \(\gamma_0\) and random effects \(u_{0j}\) and \(e_{ij}\):

\[
Y_{ij} = \gamma_0 + u_{0j} + e_{ij}.
\]

(16)

The random \(u_{0j}\) provides an estimate of how classes vary in overall inter-rater error/intra-rater error/criterion error index value of peer assessment.
Next, one comprehensive model was created to address all research questions in this study. Three separate analyses were run to examine inter-rater reliability (inter-rater error), intra-rater reliability (intra-rater error), and validity (criterion error) respectively, with the same sets of independent variables: (1) As measures of peer rater’s content knowledge in assessment domain, a continuous independent variable \( K_{ij} \) was added at level-1. The independent variable \( K_{ij} \) is the peer rater’s Chinese language mid-term examination score. For interpretation, \( K_{ij} \) is centered around the grand mean \( (K_{ij} - \bar{K}_{..}) \) and is denoted as \( cK_{ij} \). (2) As measure of peer rater’s training experience, one continuous independent variable \( T_j \) was included at level-1, which reflects the amount of Previous Peer Rating Experience of the rater, and one categorical independent \( T_j \) was included at level-2, which reflects whether the class received Self-paced Training in this study and takes a value of “0” for no and a value of “1” for yes. (3) As measures of a peer rater’s rating motivation, two categorical independent variables \( (V_j) \) and \( (M_j) \) were added at level-2. The Independent variable \( V_j \) denotes whether a class verbal persuasion or not, and \( M_j \) denotes whether a class receive monetary incentives or not. Both of them take a value of “0” for no and a value of “1” for yes.

The level-1 model of this comprehensive model is

\[
Y_{ij} = \beta_{0j} + \beta_{1j} cK_{ij} + \beta_{1j} N_{ij} + e_{ij}. \tag{17}
\]

Level 2 models include

\[
\beta_{0j} = \gamma_{00} + \gamma_{01} T_j + \gamma_{02} V_j + \gamma_{03} M_j + \gamma_{04} T_j \ast V_j + \gamma_{05} T_j \ast M_j + \gamma_{06} M_j \ast V_j + u_{0j}. \tag{18}
\]

\[
\beta_{1j} = \gamma_{10} + \gamma_{11} T_j + \gamma_{12} V_j + \gamma_{13} M_j + u_{1j}. \tag{19}
\]

\[
\beta_{2j} = \gamma_{20} + \gamma_{21} T_j + \gamma_{22} V_j + \gamma_{23} M_j + u_{2j}.
\]

Combining level-1 and level-2 models yields
\[ Y_{ij} = \gamma_{00} + \gamma_{01} T_j + \gamma_{02} V_j + \gamma_{03} M_j + \gamma_{04} T_j \cdot V_j + \gamma_{05} T_j \cdot M_j + \gamma_{06} M_j \cdot V_j + \gamma_{10} cK_{ij} + \gamma_{11} V_j \cdot cK_{ij} + \gamma_{12} V_j \cdot cN_{ij} + \gamma_{13} M_j \cdot cK_{ij} + \gamma_{20} N_{ij} + \gamma_{21} T_j \cdot cN_{ij} + \gamma_{22} V_j \cdot cN_{ij} + \gamma_{23} M_j \cdot cN_{ij} + u_{0j} + u_{1j} \cdot cK_{ij} + u_{2j} \cdot cN_{ij} + e_{ij} \]  

(21)

where,

\( Y_{ij} \) = the inter-rater error/intra-rater error/criterion error index value of peer rater \( i \) in class \( j \);

\( \gamma_{00} \) = the overall intercept;

\( \gamma_{01} \) = the main effect of training;

\( \gamma_{02} \) = the main effect of motivation intervention of verbal persuasion;

\( \gamma_{03} \) = the main effect of motivation intervention of monetary incentives;

\( \gamma_{04} \) = the class level interaction between training and motivation intervention of monetary incentives;

\( \gamma_{05} \) = the class level interaction between training and motivation intervention of verbal persuasion;

\( \gamma_{06} \) = the class level interaction between effect of verbal persuasion and monetary incentives;

\( \gamma_{10} \) = the main effect of Chinese language mid-term examination score;

\( \gamma_{11} \) = the cross level interaction involving training and Chinese language mid-term examination score;

\( \gamma_{12} \) = the cross level interaction involving motivation intervention of verbal persuasion and Chinese language mid-term examination score;

\( \gamma_{13} \) = the cross level interaction involving motivation intervention of monetary incentives and Chinese language mid-term examination score;

\( u_{0j} \) = the random effect of class \( j \) on the intercept;
\[ u_{1j} \text{= the random effect of class } j \text{ on the slope } \beta_{1j}; \]

\[ u_{2j} \text{= the random effect of class } j \text{ on the slope } \beta_{2j}; \]

\[ e_{ij} \text{= the level-1 error.} \]

The fixed effects of \( \gamma_{10} \) provide evidence pertinent to the first research question: Will peer raters’ content knowledge in assessment domain affect the rater reliability and rating validity of their peer assessment? A significant estimate for \( \gamma_{10} \) indicates a correlation between peer raters’ Chinese language mid-term examination scores and their inter-rater error/intra-rater error/criterion error.

Hypothesis testing related to the fixed effect \( \gamma_{01} \) and \( \gamma_{20} \) provides information pertinent to research questions 2: Will peer raters’ training experiences lead to more reliable and valid peer assessment scores? If the estimates for \( \gamma_{01} \) is negatively significant, there will be evidence to prove that the peer raters whose class received self-paced training have smaller inter-rater error/ intra-rater error/ criterion error in peer assessment than those whose class did not receive training. If the estimates for \( \gamma_{20} \) is negatively significant, there would be evidence to prove that peer raters who had done more peer assessments before have smaller inter-rater error/ intra-rater error/ criterion error in peer assessment than those whose class did not receive training.

The fixed effects of \( \gamma_{02} \) and \( \gamma_{03} \) provide evidence pertinent to the third research question: Will peer raters’ rating motivation influence the reliability and validity of peer assessment? A significant estimates of \( \gamma_{02} \) indicates a difference in inter-rater error/ intra-rater error/ criterion error index value of peer assessment between those who are in verbal persuasion classes and those in non-verbal persuasion classes. A significant estimate of \( \gamma_{03} \) indicates a difference in inter-rater error / intra-rater error / criterion error index value of peer assessment between those who are in monetary incentives classes and those in non-monetary incentives classes.
This model can also be used to address research question 4, 5 and 6 about whether there are interaction effects among peer raters’ content knowledge, training experience, and rating motivation on the reliability and validity of peer assessment. Of particular interest are the fixed effects of $\gamma_{04}$, $\gamma_{05}$, $\gamma_{11}$, $\gamma_{12}$, $\gamma_{13}$, $\gamma_{22}$, and $\gamma_{23}$. Hypothesis testing related to the fixed effects of $\gamma_{12}$ and $\gamma_{13}$ provide information pertinent to research question 4: Is there an interaction effect between peer raters’ rating motivation, and their training experience on rater reliability and rating validity of peer assessment? If $\gamma_{04}$ or $\gamma_{05}$ is significant, there would be evidence of interaction effects between rating motivation measured by receiving verbal persuasion or receiving monetary incentives, and peer raters’ previous peer training experience. If $\gamma_{22}$ or $\gamma_{23}$ is significant, there would be evidence of interaction effects between rating motivation measured by receiving verbal persuasion or receiving monetary incentives, and peer raters’ previous peer rating experience. Hypothesis testing related to the fixed effects of $\gamma_{12}$ and $\gamma_{13}$ provide information pertinent to research question 5: Is there an interaction effect between peer raters’ rating motivation and their content knowledge on rater reliability and rating validity of peer assessment? If $\gamma_{12}$ or $\gamma_{13}$ is significant, there would be evidence of interaction effect between rating motivation measured by whether receiving verbal persuasion or whether receiving monetary incentives and peer raters’ content knowledge measured by Chinses mid-term examination scores. Hypothesis testing related to the fixed effect of $\gamma_{11}$ provide information pertinent to research question 4: Is there an interaction effect between peer raters’ content knowledge and their training experience on rater reliability and rating validity of peer assessment? If $\gamma_{11}$ is significant, there would be evidence of interaction effect between training experience measured by number of times of performing peer assessments previously and peer raters’ content knowledge measured Chinses mid-term examination score.
Chapter 5

Results

Estimation of the fixed and random effect in Equation 16 and 21 was done using the HLM software, version 7 (Raudenbush, Bryk, Cheong, Congdon, & du Toit, 2011). The descriptive statistics and graphs were generated using SPSS version 21. There are several outliers of dependent variables identified in the process of exploratory data analysis (EDA), (see the Boxplots in Figure 4-1, Figure 4-2, and Figure 4-3). From the scatter plot of two continuous independent variables (i.e., Content knowledge in the assessed domain and previous rating experience) and three dependent variables (i.e., inter-rater error, intra-rater error, and criterion error), there is no clear pattern other than linear, which means the linearity assumption holds. Descriptive statistics of both dependent and independent variables are provided in Table 4-1. There are a few missing values that measure content knowledge and previous peer rating experience. Because the missing pattern is completely random (Little’s MCAR test: Chi-Square = 11.870, DF = 14, p-value = 0.617) and the percent of missing values is less than 2%, list-wise deletion approach was adopted. Equations 16 and 21, the 2-level hierarchical linear models (HLM) were fitted to peer raters’ inter-rater error, intra-rater error, and criterion error respectively. The results are reported in three sections: one for each of the three dependent variables.
<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Mean</th>
<th>Std. Deviation</th>
<th>Skewness</th>
<th>Kurtosis</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Statistic</td>
<td>Statistic</td>
<td>Statistic</td>
<td>Statistic</td>
<td>Statistic</td>
<td>Statistic</td>
<td>Std. Error</td>
</tr>
<tr>
<td>Inter-Rater Error</td>
<td>838</td>
<td>.6</td>
<td>31.6</td>
<td>7.38</td>
<td>3.72</td>
<td>1.56</td>
<td>.084</td>
</tr>
<tr>
<td>Intra-Rater Error</td>
<td>838</td>
<td>.75</td>
<td>24.50</td>
<td>9.73</td>
<td>3.44</td>
<td>.30</td>
<td>.084</td>
</tr>
<tr>
<td>Criterion Error</td>
<td>838</td>
<td>3.25</td>
<td>34.00</td>
<td>12.01</td>
<td>4.16</td>
<td>1.00</td>
<td>.084</td>
</tr>
<tr>
<td>Content Knowledge</td>
<td>828</td>
<td>55</td>
<td>132</td>
<td>100.27</td>
<td>12.86</td>
<td>-.39</td>
<td>.085</td>
</tr>
<tr>
<td>Pre-Rating Experience</td>
<td>824</td>
<td>1</td>
<td>7</td>
<td>2.75</td>
<td>1.85</td>
<td>1.09</td>
<td>.085</td>
</tr>
<tr>
<td>Self-paced Training</td>
<td>838</td>
<td>0</td>
<td>1</td>
<td>.45</td>
<td>.50</td>
<td>.20</td>
<td>.084</td>
</tr>
<tr>
<td>Verbal Persuasion</td>
<td>838</td>
<td>0</td>
<td>1</td>
<td>.50</td>
<td>.50</td>
<td>-.01</td>
<td>.084</td>
</tr>
<tr>
<td>Monetary Incentives</td>
<td>838</td>
<td>0</td>
<td>1</td>
<td>.51</td>
<td>.50</td>
<td>-.04</td>
<td>.084</td>
</tr>
</tbody>
</table>
Figure 5-1. Boxplot of Inter-Rater Error across 24 Classes
Figure 5-2. Boxplot of Intra-rater Error across 24 Classes
Figure 5-3. Boxplot of Criterion Error across 24 Classes
5.1. Effect of peer rater’s quality characteristics on inter-rater reliability of peer assessment

The results of fitting the unconditional model (i.e., Equation 16) with the inter-rater error as the dependent variable are presented in Table 5-2. The value of intraclass correlation coefficient (ICC) is 10.74%, $\tau_{00} = 1.483$, $\chi^2_{(23)} = 125.049$, $p < 0.001$, which indicates that classes vary in overall Inter-rater error index value of peer assessment, and thus a 2-level model is appropriate and needed.
Table 5-2. Results from the Unconditional Model (Inter-rater Error)

<table>
<thead>
<tr>
<th>Fixed Effect</th>
<th>Coefficient</th>
<th>Standard error</th>
<th>t-ratio</th>
<th>d.f.</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>For INTRCPT1, $\beta_0$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>INTRCPT2, $\gamma_{00}$</td>
<td>7.409190</td>
<td>0.277644</td>
<td>26.686</td>
<td>23</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Random Effect</th>
<th>Standard Deviation</th>
<th>Variance Component</th>
<th>d.f.</th>
<th>$\chi^2$</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>INTRCPT1, $u_0$</td>
<td>1.21792</td>
<td>1.48333</td>
<td>23</td>
<td>125.04874</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>level-1, $r$</td>
<td>3.51094</td>
<td>12.32668</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The hypothesized model with the Level-1 predictors and Level 2-predictors which were specified in Equation 21 was fitted with the inter-rater error as the dependent variable. With regard to specification of Level-1 coefficients as random or non-random, the following four models were compared: Model 1 with randomly varying intercepts, Model 2 with randomly varying intercepts and a slope of $cK_{ij}$, Model 3 with randomly varying intercepts and a slope of $cN_{ij}$, and Model 4 with randomly varying intercepts and two slopes. The model comparison results indicated that Model 3 had the best model fit. More details were provided in Table 5-3.

Table 5-3. Model Comparison Results from the Hypothesized Model (Inter-rater Error)

<table>
<thead>
<tr>
<th>Model Comparison</th>
<th>$\chi^2$</th>
<th>d.f.</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model 1 vs Model 2</td>
<td>0.790</td>
<td>2</td>
<td>0.674</td>
</tr>
<tr>
<td>Model 1 vs Model 3</td>
<td>12.896</td>
<td>2</td>
<td>0.002</td>
</tr>
<tr>
<td>Model 3 vs Model 4</td>
<td>6.329</td>
<td>3</td>
<td>0.097</td>
</tr>
</tbody>
</table>

Note. Model 1 has randomly varying intercepts only, Model 2 has randomly varying intercepts and a slope of $cK_{ij}$, Model 3 has randomly varying intercepts and a slope of $cN_{ij}$, and Model 4 has randomly varying intercepts and two slopes.
The results of Model 3 were presented in Table 5-4. The fixed effects of $\gamma_{10}$ provide evidence pertinent to part of the first research question: Will peer raters’ content knowledge in the assessed domain affect the inter-rater reliability? From Table 5-4, $\gamma_{10} = -0.011609$, $t = -1.066$, $p = 0.287$, indicating a negative correlation between students’ mid-term examination score and inter-rater error in peer assessment; however, the correlation is not statistically significant in this study.

Hypothesis testing related to the fixed effect $\gamma_{01}$ and $\gamma_{20}$ provide information pertinent to research question 2: Will peer raters’ training experience lead to small inter-rater error? Because $\gamma_{01} = -0.019356$, $t = -0.041$, $p = 0.968$, there is not enough statistical evidence to show that peer raters whose class received training have smaller inter-rater error in peer assessment than those whose class did not receive training. Also peer raters with more rating experience do not outperform those with less rating experience, $\gamma_{20} = -0.000494$, $t = -0.005$, $p = 0.996$.

The fixed effects of $\gamma_{02}$ and $\gamma_{03}$ provide evidence pertinent to part of the third research question: Will peer raters’ rating motivation influence the inter-rater reliability of peer assessment? The results showed that the students who received verbal persuasion have smaller inter-rater error than those in non-verbal persuasion classes, $\gamma_{02} = -0.829513$, $t = -1.925$, $p = 0.071$, but the difference is not significant. However, students in monetary incentives classes have statistically significant smaller inter-rater error than those in non-monetary incentives classes, controlling for the effect of verbal persuasion and training, $\gamma_{03} = -1.318284$, $t = -3.335$, $p = 0.004$. The Cohen’s $d$ is 0.28, which indicates a small effect size.

For interaction effects between peer raters’ content knowledge and rating motivation on inter-rater reliability (research question 4), the results showed that students in monetary incentive classes have stronger slope than those in non-monetary incentives classes, but the difference is not statistically significant, $\gamma_{13} = 0.040025$, $t = 1.770$, $p = 0.077$. Similarly, no statistically
significant difference in slope strength was found between students in verbal persuasion class and nonverbal persuasion class.

Table 5-4 Results from the Intercepts-and Slopes-as-Outcome Model (Inter-rater Error)

<table>
<thead>
<tr>
<th>Fixed Effect</th>
<th>Coefficient</th>
<th>Standard error</th>
<th>t-ratio</th>
<th>Approx. d.f.</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>For INTRCPT1, $\beta_0$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>INTRCPT2, $\gamma_{00}$</td>
<td>7.479712</td>
<td>0.188474</td>
<td>39.686</td>
<td>17</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>TRAIN, $\gamma_{01}$</td>
<td>0.011519</td>
<td>0.407100</td>
<td>0.28</td>
<td>17</td>
<td>0.978</td>
</tr>
<tr>
<td>LANGUAGE, $\gamma_{02}$</td>
<td>-0.829513</td>
<td>0.430946</td>
<td>-1.925</td>
<td>17</td>
<td>0.071</td>
</tr>
<tr>
<td>MONEY, $\gamma_{03}$</td>
<td>-1.318284</td>
<td>0.395250</td>
<td>-3.335</td>
<td>17</td>
<td><strong>0.004</strong></td>
</tr>
<tr>
<td>T*L, $\gamma_{04}$</td>
<td>1.524019</td>
<td>0.738812</td>
<td>2.063</td>
<td>17</td>
<td>0.055</td>
</tr>
<tr>
<td>T*M, $\gamma_{05}$</td>
<td>1.697676</td>
<td>0.672412</td>
<td>2.525</td>
<td>17</td>
<td><strong>0.022</strong></td>
</tr>
<tr>
<td>M*L, $\gamma_{06}$</td>
<td>-0.022900</td>
<td>0.714193</td>
<td>-0.032</td>
<td>17</td>
<td>0.975</td>
</tr>
<tr>
<td>For TOTALK slope, $\beta_1$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>INTRCPT2, $\gamma_{10}$</td>
<td>-0.011609</td>
<td>0.010891</td>
<td>-1.066</td>
<td>762</td>
<td>0.287</td>
</tr>
<tr>
<td>TRAIN, $\gamma_{11}$</td>
<td>0.008730</td>
<td>0.022418</td>
<td>0.389</td>
<td>762</td>
<td>0.697</td>
</tr>
<tr>
<td>LANGUAGE, $\gamma_{12}$</td>
<td>-0.028512</td>
<td>0.023077</td>
<td>-1.236</td>
<td>762</td>
<td>0.217</td>
</tr>
<tr>
<td>MONEY, $\gamma_{13}$</td>
<td>0.040025</td>
<td>0.022607</td>
<td>1.770</td>
<td>762</td>
<td>0.077</td>
</tr>
<tr>
<td>For QEXPERIE slope, $\beta_2$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>INTRCPT2, $\gamma_{20}$</td>
<td>-0.000494</td>
<td>0.099099</td>
<td>-0.005</td>
<td>20</td>
<td>0.996</td>
</tr>
<tr>
<td>TRAIN, $\gamma_{21}$</td>
<td>0.064162</td>
<td>0.198260</td>
<td>0.324</td>
<td>20</td>
<td>0.750</td>
</tr>
<tr>
<td>LANGUAGE, $\gamma_{22}$</td>
<td>0.140970</td>
<td>0.205736</td>
<td>0.685</td>
<td>20</td>
<td>0.501</td>
</tr>
<tr>
<td>MONEY, $\gamma_{23}$</td>
<td>0.420489</td>
<td>0.209011</td>
<td>2.012</td>
<td>20</td>
<td>0.058</td>
</tr>
</tbody>
</table>

Random Effect | Standard Deviation | Variance Component | d.f. | $\chi^2$ | p-value |
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>INTRCPT1, u0</td>
<td>0.94166</td>
<td>0.88672</td>
<td>17</td>
<td>56.85200</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>QEXPERIE slope, u2</td>
<td>0.42956</td>
<td>0.18452</td>
<td>20</td>
<td>51.90049</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>level-1, r</td>
<td>3.42615</td>
<td>11.73848</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Hypothesis testing related to the fixed effects of $\gamma_{04}$, $\gamma_{05}$, $\gamma_{06}$, $\gamma_{22}$ and $\gamma_{23}$ provide information pertinent to research question 5: Is there an interaction effect between peer raters’ rating motivation and rating/training experience on inter-rater reliability of peer assessment?

There is no statistically significant difference in slope strength between students in verbal persuasion class and non-verbal persuasion class, $\gamma_{22} = 0.140970$, $t = 0.685$, $p = 0.501$. Similarly,
no statistically significant difference in slope strength was found between students in monetary incentives classes and those in non-monetary incentives classes, $\gamma_{23} = 0.420489$, $t = 2.012$, $p = 0.058$. However, there is a significant interaction effect on interrater error between self-paced training and monetary incentives at class level, $\gamma_{05} = 0.697676$, $t = 2.525$, $p = 0.022$.

For interaction effects between peer raters’ content knowledge and training experience on inter-rater reliability (research question 6), the results showed that students in training classes have stronger slope than those in non-training classes, but the difference is not statistically significant, $\gamma_{11} = 0.008730$, $t = 0.389$, $p = 0.697$

For model assumption check, The QQ plots for residuals at both Level 1 and Level 2 are presented in Figure 5-5. The plots are approximately linear, suggesting there is not serious departure from a normal distribution and that the normality assumption is tenable. The Scatter plots of residuals versus fitted values are generated at each level (See Figure 5-6). No clear on these scatter plot indicate that homoscedasticity assumption is not violated. The box-plots for residuals at each level are presented in Figure 5-7. There are several outlier of residuals at Level 1 and one outlier at Level 2.

![Figure 5-5. Normal Q-Q Plots of Residuals at Level 1 and Level 2 (Inter-rater Error)](image-url)
Figure 5-6. Scatter plot of Residual vs. Fitted Value at Level 1 and Level 2 (Inter-rater Error)

Figure 5-7. Boxplot of Residuals at Level 1 and Level 2 (Inter-rater Error)
5.2. Effect of peer rater’s quality characteristics on intra-rater reliability of peer assessment

The results of fitting the unconditional model (i.e., Equation 16) with the intra-rater error as the dependent variable are presented in 5-5. The value of intraclass correlation coefficient (ICC) is 11.11%, $\tau_{00} = 1.32064$, $\chi^2_{(23)} = 124.44608$, $p <0.001$, which indicate that classes vary in overall Intra-rater error index value of peer assessment and thus a 2-level model is appropriate and needed.

Table 5-5. Results from the Unconditional Model (Intra-rater Error)

<table>
<thead>
<tr>
<th>Fixed Effect</th>
<th>Coefficient</th>
<th>Standard error</th>
<th>t-ratio</th>
<th>Approx. d.f.</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>For INTRCPT1, $\beta_0$</td>
<td>9.702430</td>
<td>0.260967</td>
<td>37.179</td>
<td>23</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>INTRCPT2, $\gamma_{00}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Random Effect</th>
<th>Standard Deviation</th>
<th>Variance Component</th>
<th>d.f.</th>
<th>$\chi^2$</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>INTRCPT1, $u_0$</td>
<td>1.14919</td>
<td>1.32064</td>
<td>23</td>
<td>124.44608</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>level-1, $r$</td>
<td>3.25029</td>
<td>10.56435</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The hypothesized model with the Level-1 predictors and Level 2-predicators which were specified in Equation 21 was fitted with the intra-rater error as the dependent variable. With regard to specification of Level-1 coefficients as random or non-random, the following four models were compared: Model 1 with randomly varying intercepts, Model 2 with randomly varying intercepts and a slope of $cK_{ij}$, Model 3 with randomly varying intercepts and a slope of $cN_{ij}$, and Model 4 with randomly varying intercepts and two slopes. The model comparison results indicated that more complicated models did not have significant better fit (See Table 5-6), therefore Model 1 was preferred for parsimony concern.
Table 5-6. Model Comparison Results from the Hypothesized Model (Intra-rater Error)

<table>
<thead>
<tr>
<th>Model Comparison</th>
<th>$\chi^2$</th>
<th>d.f.</th>
<th>$p$-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model 1 vs Model 2</td>
<td>1.691</td>
<td>2</td>
<td>0.429</td>
</tr>
<tr>
<td>Model 1 vs Model 3</td>
<td>1.018</td>
<td>2</td>
<td>0.601</td>
</tr>
<tr>
<td>Model 3 vs Model 4</td>
<td>1769</td>
<td>3</td>
<td>0.622</td>
</tr>
</tbody>
</table>

Note. Model 1 has randomly varying intercepts only, Model 2 has randomly varying intercepts and a slope of $cK_{ij}$, Model 3 has randomly varying intercepts and a slope of $cN_{ij}$, and Model 4 has randomly varying intercepts and two slopes.

The results of Model 1 were presented in Table 5-7. The fixed effects of $\gamma_{10}$ provide evidence pertinent to part of the first research question: Will peer raters’ content knowledge in assessment domain affect the intra-rater reliability? From Table 5-7, $\gamma_{10} = -0.108999$, $t = -4.391$, $p < 0.001$, indicating a significant negative correlation between students’ mid-term examination score and intra-rater error in peer assessment. The partial standardized regression coefficient of content knowledge in the assessed domain is $-0.024$, which means for every standard deviation unit increase in content knowledge in the assessment domain, the intra-rater error of peer rating decrease by 0.024 standard deviation unites.

Hypothesis testing related to the fixed effect $\gamma_{01}$ provides information pertinent to research question 2: Will peer raters’ training/rating experience lead to small intra-rater error? Because $\gamma_{01} = 0.093853$, $t = 0.179$, $p = 0.860$, there is not enough statistical evidence to prove that peer raters whose class received training have smaller intra-rater error in peer assessment than those whose class did not receive training. However, the negative correlation between students’ previous peer rating experience and intra-rater error was statistically significant, $\gamma_{20} = -0.307415$, $t = -2.125$, $p = 0.034$. 
The fixed effects of $\gamma_{02}$ and $\gamma_{03}$ provide evidence pertinent to part of the third research question: Will peer raters’ rating motivation influence the intra-rater reliability of peer assessment? The results showed that there was no statistically significant difference in intrarater error between the students in verbal-persuasion classes and those in non-verbal persuasion classes, controlling for the effect of monetary incentives and training, $\gamma_{02} = 0.710184$, $t = 1.448$, $p = 0.166$. However, the effect of monetary incentives is statistically significant $\gamma_{03} = -2.811668$, $t = -6.288$, $p < 0.001$, Cohen’s $d = 0.56$; the students in monetary incentives classes have statistically significant smaller intra-rater error than those in non-monetary incentives classes, controlling for the effect of verbal persuasion and training.

For interaction effects between peer raters’ content knowledge and rating motivation (research question 4), rating motivation and training experience (research question 5), and training experience and content knowledge (research question 6) on intra-rater reliability of peer assessment, no statistically significant difference in slope strength was found.

For interaction effects between peer raters’ content knowledge and rating motivation on intra-rater reliability (research question 4), the results showed that students in monetary incentive classes have significantly stronger slope than those in non-monetary incentives classes, $\gamma_{13} = 0.146992$, $t = 3.371$, $p < 0.001$. However, no statistically significant difference in slope strength was found between students in verbal persuasion class and nonverbal persuasion class, $\gamma_{12} = -0.009175$, $t = -0.213$, $p = 0.832$.

Hypothesis testing related to the fixed effects of $\gamma_{04}$, $\gamma_{05}$, $\gamma_{22}$ and $\gamma_{23}$ provide information pertinent to research question 5: Is there an interaction effect between peer raters’ rating motivation and rating/training experience on intra-rater reliability of peer assessment? There is no statistically significant difference in slope strength between students in verbal persuasion class and non-verbal persuasion class, $\gamma_{22} = 0.571411$, $t = 1.873$, $p = 0.061$. Similarly,
no statistically significant difference in slope strength was found between students in monetary incentives classes and those in non-monetary incentives classes, $\gamma_{23} = 0.495995$, $t = 1.474$, $p = 0.141$. However, there is a significant interaction effect on intra-rater error at class level between self-paced training and verbal persuasion $\gamma_{04} = -2.958884$, $t = -2.901$, $p = 0.010$.

For interaction effects between peer raters’ content knowledge and training experience on intra-rater reliability (research question 6), the results showed that students in training classes have stronger slope than those in non-training classes, but the difference is not statistically significant, $\gamma_{11} = 0.003798$, $t = 0.083$, $p = 0.934$.

Table 5-7. Results from the Intercepts-and Slopes-as-Outcome Model (Intra-rater Error)

<table>
<thead>
<tr>
<th>Fixed Effect</th>
<th>Coefficient</th>
<th>Standard error</th>
<th>$t$-ratio</th>
<th>Approx. d.f.</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>For INTRCPT1, $\beta_0$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>INTRCPT2, $\gamma_{00}$</td>
<td>19.378327</td>
<td>0.229268</td>
<td>84.522</td>
<td>17</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>TRAIN, $\gamma_{01}$</td>
<td>0.093853</td>
<td>0.523662</td>
<td>0.179</td>
<td>17</td>
<td>0.860</td>
</tr>
<tr>
<td>LANGUAGE, $\gamma_{02}$</td>
<td>0.710184</td>
<td>0.490583</td>
<td>1.448</td>
<td>17</td>
<td>0.166</td>
</tr>
<tr>
<td>MONEY, $\gamma_{03}$</td>
<td>-2.811668</td>
<td>0.447161</td>
<td>-6.288</td>
<td>17</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>TL, $\gamma_{04}$</td>
<td>-2.958884</td>
<td>1.019885</td>
<td>-2.901</td>
<td>17</td>
<td>0.010</td>
</tr>
<tr>
<td>TM, $\gamma_{05}$</td>
<td>0.998658</td>
<td>0.967430</td>
<td>1.032</td>
<td>17</td>
<td>0.316</td>
</tr>
<tr>
<td>ML, $\gamma_{06}$</td>
<td>-2.197066</td>
<td>0.875794</td>
<td>-2.509</td>
<td>17</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>For TOTALK slope, $\beta_1$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>INTRCPT2, $\gamma_{10}$</td>
<td>-0.108999</td>
<td>0.024826</td>
<td>-4.391</td>
<td>782</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>TRAIN, $\gamma_{11}$</td>
<td>0.003798</td>
<td>0.045961</td>
<td>0.083</td>
<td>782</td>
<td>0.934</td>
</tr>
<tr>
<td>LANGUAGE, $\gamma_{12}$</td>
<td>-0.009175</td>
<td>0.043173</td>
<td>-0.213</td>
<td>782</td>
<td>0.832</td>
</tr>
<tr>
<td>MONEY, $\gamma_{13}$</td>
<td>0.146992</td>
<td>0.043607</td>
<td>3.371</td>
<td>782</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>For QEXPERIE slope, $\beta_2$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>INTRCPT2, $\gamma_{20}$</td>
<td>-0.307415</td>
<td>0.144674</td>
<td>-2.125</td>
<td>782</td>
<td>0.034</td>
</tr>
<tr>
<td>TRAIN, $\gamma_{21}$</td>
<td>0.223148</td>
<td>0.301037</td>
<td>0.741</td>
<td>782</td>
<td>0.459</td>
</tr>
<tr>
<td>LANGUAGE, $\gamma_{22}$</td>
<td>0.571411</td>
<td>0.305101</td>
<td>1.873</td>
<td>782</td>
<td>0.061</td>
</tr>
<tr>
<td>MONEY, $\gamma_{23}$</td>
<td>0.495995</td>
<td>0.336559</td>
<td>1.474</td>
<td>782</td>
<td>0.141</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Random Effect</th>
<th>Standard Deviation</th>
<th>Variance Component</th>
<th>d.f.</th>
<th>$\chi^2$</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>INTRCPT1, u0</td>
<td>0.32935</td>
<td>0.10847</td>
<td>17</td>
<td>14.89275</td>
<td>&gt;0.500</td>
</tr>
<tr>
<td>level-1, r</td>
<td>7.45342</td>
<td>55.55340</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
For model assumption check, the QQ plots for residuals at both Level 1 and Level 2 are presented in Figure 5-8. The plots are approximately linear, suggesting there is not serious departure from a normal distribution and that the normality assumption is tenable. The Scatter plots of residuals versus fitted values are generated at each level (See Figure 5-9). No clear on these scatter plot indicate that homoscedasticity assumption is not violated. The box-plots for residuals at each level are presented in Figure 5-10. There are several outlier of residuals at Level 1, while on outlier is detected at Level 2.

Figure 5-8. Normal Q-Q Plots of Residuals at Level 1 and Level 2 (Intra-rater Error)
Figure 5-9. Scatter plot of Residual vs. Fitted Value at Level 1 and Level 2 (Intra-rater Error)

Figure 5-10. Boxplot of Residuals at Level 1 and Level 2 (Intra-rater Error)
5.3. Effect of peer rater’s quality characteristics on rating validity of peer assessment

The results of fitting the unconditional model (i.e., Equation 16) with the criterion error as the dependent variable are presented in Table 5-8. The value of intraclass correlation coefficient (ICC) is 9.00%, $\tau_{00} = 1.560$, $\chi^2_{(23)} = 104.804$, $p <0.001$, indicating that classes vary in overall criterion error index value of peer assessment and thus a 2-level model is appropriate and needed.

Table 5-8. Results from the Unconditional Model (Criterion Error)

<table>
<thead>
<tr>
<th>Fixed Effect</th>
<th>Coefficient</th>
<th>Standard Error</th>
<th>t-ratio</th>
<th>Approx. d.f.</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>For INTRCPT1, $\beta_0$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>INTRCPT2, $\gamma_{00}$</td>
<td>12.009199</td>
<td>0.284559</td>
<td>42.203</td>
<td>23</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Random Effect</th>
<th>Standard Deviation</th>
<th>Variance Component</th>
<th>d.f.</th>
<th>$\chi^2$</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>INTRCPT1, $u_0$</td>
<td>1.24895</td>
<td>1.55988</td>
<td>23</td>
<td>104.80432</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>level-1, $r$</td>
<td>3.97062</td>
<td>15.76584</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The hypothesized model with the Level-1 predictors and Level 2-predictors which were specified in Equation 21 was fitted with the criterion error as the dependent variable. With regard to specification of Level-1 coefficients as random or non-random, the following four models were compared: Model 1 with randomly varying intercepts, Model 2 with randomly varying intercepts and a slope of $cK_{ij}$, Model 3 with randomly varying intercepts and a slope of $cN_{ij}$, and Model 4 with randomly varying intercepts and two slopes. The model comparison results indicated that more complicated models did not have significant better fit (See Table 5-9), therefore Model 1 was preferred for parsimony concern.
Table 5-9. Model Comparison Results from the Hypothesized Model (Criterion Error)

<table>
<thead>
<tr>
<th>Model Comparison</th>
<th>$\chi^2$</th>
<th>d.f.</th>
<th>$p$-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model 1 vs Model 2</td>
<td>1.454</td>
<td>2</td>
<td>0.483</td>
</tr>
<tr>
<td>Model 1 vs Model 3</td>
<td>3.41</td>
<td>2</td>
<td>0.182</td>
</tr>
<tr>
<td>Model 3 vs Model 4</td>
<td>1.800</td>
<td>3</td>
<td>0.614</td>
</tr>
</tbody>
</table>

Note. Model 1 has randomly varying intercepts only, Model 2 has randomly varying intercepts and a slope of $cK_{ij}$, Model 3 has randomly varying intercepts and a slope of $cN_{ij}$, and Model 4 has randomly varying intercepts and two slopes.

The results of Model 1 are found in Table 5-10. The fixed effects of $\gamma_{10}$ provide evidence pertinent to part of the first research question: Will peer raters’ content knowledge in assessment domain affect the validity of peer assessment? From Table 5-10, $\gamma_{10} = -0.074207, t = -5.887, p < 0.001$, indicates a statistically significant negative correlation between students’ mid-term examination score and criterion error in peer assessment. The partial standardized regression coefficient of content knowledge in the assessed domain is -0.029, which means for every standard deviation unit increase in content knowledge in the assessment domain, the criterion rater error of peer rating decrease by 0.029 standard deviation unites.

Hypothesis testing related to the fixed effect $\gamma_{01}$ provides information pertinent to research question 2: Will peer raters’ training experience lead to small criterion error? No statistically significant effect of training experience on criterion error was found in this study. A statistically significant negative correlation was found between students’ previous rating experience and criterion error in peer assessment, $\gamma_{20} = -0.197985, t = -2.518, p = 0.012$. The partial standardized regression coefficient of previous rating experience is -0.12, which means for every standard deviation unit increase previous rating experience, the criterion rater error of peer rating decrease by 0.12 standard deviation unites.
The fixed effects of $\gamma_{02}$ and $\gamma_{03}$ provide evidence pertinent to part of the third research question: Will peer raters’ rating motivation influence the validity of peer assessment? The results showed that the students in monetary incentives classes have statistically significant smaller criterion error than those in non-monetary incentives classes, controlling for the effect of verbal persuasion and training, $\gamma_{03} = -1.864317, t = -8.028, p < 0.001$, Cohen’s $d = 0.68$. The effect of verbal persuasion is not statistically significant $\gamma_{02} = -0.013909, t = -0.058, p = 0.954$.

For interaction effects between peer raters’ content knowledge and rating motivation on criterion error (research question 4), the results showed that students in monetary incentive classes have stronger slope than those in non-monetary incentives classes, $\gamma_{13} = 0.077012, t = 3.268, p = 0.001$. No statistically significant difference in slope strength was found between students in verbal persuasion classes and nonverbal persuasion classes, $\gamma_{12} = -0.006419, t = -0.288, p = 0.773$.

Hypothesis testing related to the fixed effects of $\gamma_{22}$ and $\gamma_{23}$ provide information pertinent to research question 5: Is there an interaction effect between peer raters’ rating motivation and rating experience on validity of peer assessment? The results showed that students in verbal persuasion classes have stronger slope than those in non-verbal persuasion classes, $\gamma_{22} = 0.318510, t = 2.008, p = 0.045$. No statistically significant difference in slope strength was found between students in monetary incentive classes and non-monetary incentives classes, $\gamma_{23} = 0.288281, t = 1.697, p = 0.090$.

Hypothesis testing related to the fixed effects of $\gamma_{04}, \gamma_{05}, \gamma_{22}$ and $\gamma_{23}$ provide information pertinent to research question 5: Is there an interaction effect between peer raters’ rating motivation and rating/training experience on validity of peer assessment? There is no statistically significant difference in slope strength between students in verbal persuasion class and non-verbal persuasion class, $\gamma_{22} = 0.309991, t = 1.921, p = 0.055$. Similarly, no statistically significant difference in slope strength was found between students in monetary incentives classes
and those in non-monetary incentives classes, $\gamma_{23} = 0.252994$, $t = 1.376$, $p = 0.169$. However, there is a significant interaction effect on criterion error at class level between self-paced training and verbal persuasion $\gamma_{04} = -1.304229$, $t = -2.436$, $p = 0.026$.

For interaction effects between peer raters’ content knowledge and training experience on validity (research question 6), the results showed that students in training classes have stronger slope than those in non-training classes, but the difference is not statistically significant, $\gamma_{11} = 0.009711$, $t = 0.392$, $p = 0.695$.

Table 5-10. Results from the Intercepts-and Slopes-as-Outcome Model (Criterion Error)

<table>
<thead>
<tr>
<th>Fixed Effect</th>
<th>Coefficient</th>
<th>Standard error</th>
<th>t-ratio</th>
<th>Approx. d.f.</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>For INTRCPT1, $\beta_0$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>INTRCPT2, $\gamma_{00}$</td>
<td>12.097744</td>
<td>0.120994</td>
<td>99.986</td>
<td>17</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>TRAIN, $\gamma_{01}$</td>
<td>0.005980</td>
<td>0.269528</td>
<td>0.022</td>
<td>17</td>
<td>0.983</td>
</tr>
<tr>
<td>LANGUAGE, $\gamma_{02}$</td>
<td>-0.013909</td>
<td>0.238226</td>
<td>-0.058</td>
<td>17</td>
<td>0.954</td>
</tr>
<tr>
<td>MONEY, $\gamma_{03}$</td>
<td>-1.864317</td>
<td>0.232238</td>
<td>-8.028</td>
<td>17</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>T*L, $\gamma_{04}$</td>
<td>-1.304229</td>
<td>0.535382</td>
<td>-2.436</td>
<td>17</td>
<td>0.026</td>
</tr>
<tr>
<td>T*M, $\gamma_{05}$</td>
<td>0.024311</td>
<td>0.494342</td>
<td>0.049</td>
<td>17</td>
<td>0.961</td>
</tr>
<tr>
<td>M*L, $\gamma_{06}$</td>
<td>-1.129255</td>
<td>0.468522</td>
<td>-2.410</td>
<td>17</td>
<td>0.028</td>
</tr>
<tr>
<td>For TOTALK slope, $\beta_1$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>INTRCPT2, $\gamma_{10}$</td>
<td>-0.074207</td>
<td>0.012604</td>
<td>-5.887</td>
<td>782</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>TRAIN, $\gamma_{11}$</td>
<td>0.009711</td>
<td>0.024801</td>
<td>0.392</td>
<td>782</td>
<td>0.695</td>
</tr>
<tr>
<td>LANGUAGE, $\gamma_{12}$</td>
<td>-0.006419</td>
<td>0.022270</td>
<td>-0.288</td>
<td>782</td>
<td>0.773</td>
</tr>
<tr>
<td>MONEY, $\gamma_{13}$</td>
<td>0.077012</td>
<td>0.023567</td>
<td>3.268</td>
<td>782</td>
<td>0.001</td>
</tr>
<tr>
<td>For QEXPERIE slope, $\beta_2$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>INTRCPT2, $\gamma_{20}$</td>
<td>-0.197985</td>
<td>0.078637</td>
<td>-2.518</td>
<td>782</td>
<td>0.012</td>
</tr>
<tr>
<td>TRAIN, $\gamma_{21}$</td>
<td>0.154193</td>
<td>0.172687</td>
<td>0.893</td>
<td>782</td>
<td>0.372</td>
</tr>
<tr>
<td>LANGUAGE, $\gamma_{22}$</td>
<td>0.309991</td>
<td>0.161392</td>
<td>1.921</td>
<td>782</td>
<td>0.055</td>
</tr>
<tr>
<td>MONEY, $\gamma_{23}$</td>
<td>0.252994</td>
<td>0.183844</td>
<td>1.376</td>
<td>782</td>
<td>0.169</td>
</tr>
</tbody>
</table>

Random Effect

<table>
<thead>
<tr>
<th>Fixed Effect</th>
<th>Standard Deviation</th>
<th>Variance Component</th>
<th>d.f.</th>
<th>$\chi^2$</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>INTRCPT1, u0</td>
<td>0.10799</td>
<td>0.01166</td>
<td>17</td>
<td>15.36115</td>
<td>&gt;0.500</td>
</tr>
<tr>
<td>level-1, r</td>
<td>3.85198</td>
<td>14.83772</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
For model assumption check, The QQ plots for residuals at both Level 1 and Level 2 are presented in Figure 5-11. The plots are approximately linear, suggesting there is not serious departure from a normal distribution and that the normality assumption is tenable. The Scatter plots of residuals versus fitted values are generated at each level (See Figure 5-12). No clear on these scatter plot indicate that homoscedasticity assumption is not violated. The box-plots for residuals at each level are presented in Figure 5-13. There are several outlier of residuals at Level 1, and one outlier at Level 2.

Figure 5-11. Normal Q-Q Plots of Residuals at Level 1 and Level 2 (Criterion Error)
Figure 5-12. Scatter plot of Residual vs. Fitted Value at Level 1 and Level 2 (Criterion Error)

Figure 5-13. Boxplot of Residuals at Level 1 and Level 2 (Criterion Error)
Chapter 6
Discussion, Implications, Limitations and Future Research

This dissertation used randomized cluster trials on peer assessments of written compositions of Chinese high school students with the aim of improving peer assessment reliability and validity by providing empirical evidence and rationale to support the idea of giving different weights to scores assigned by different peer raters based on their content knowledge in the assessment domain, training experience, and rating motivation. Each participant was expected to rate four well-selected essays written by peers for a Chinese language mid-term examination using the standard Gaokao (National College Entrance Examination) Chinese Writing Rubric. Before performing peer assessment, half of the classes were randomly selected to receive a self-paced training of peer rating. Similarly, to manipulate students’ rating motivation, randomly chosen classes of students were offered motivation incentives which included monetary rewards for accurate peer rating results and the teacher’s verbal persuasion about the value of performing peer assessment. This 2 × 2 × 2 completely crossed factorial design was employed with two levels of training intervention (i.e., self-paced training vs. no training), two levels of monetary incentives (i.e., monetary reward vs. no monetary reward), and two levels of verbal persuasion (i.e., verbal persuasion vs. no verbal persuasion). Peer rater’s mid-term Chinese examination score and the number of peer assessments they had previously participated were used to individually measure peer rater’s content knowledge in assessment domain and training experience respectively.

This chapter begins with the overall findings. It then discusses the implications of the findings on in-person and online peer assessment activities. Finally, limitations of the study and important areas for future research are addressed.
6.1. Discussion of the overall findings

The following briefly discusses the overall findings of the HLM analysis. The first research question is about the effect of peer raters’ content knowledge on rater reliability and rating validity of peer assessment. Hypothesis 1 postulated that there is a negative correlation between peer raters’ content knowledge in the assessed domain and their (a) inter-rater error, (b) intra-rater error, and (c) criterion error in peer assessment. Part (b) and part (c) were supported. The result showed that peer raters with higher scores of content knowledge in the assessed domain tend to have higher intra-rater reliability and rating validity in peer assessments. This result is also well supported in the literature and theory about novice raters and expert raters. Compared to novices, experts have more comprehensive knowledge in a specific domain and a better hierarchical organization of the knowledge, which leads to more effective cognitive processes and better task performance (Ross, 2006; Schunk, 2008). In assessment situations, peer raters are novices in the content knowledge domain, but the amount of their content knowledge in the assessed domain varies considerably. Therefore, peer raters with more content knowledge in the assessed domain, as hypothesized, produce assessment scores closer to the expert’s scores and more consistent across different assignments.

The effect of peer rater’s content knowledge in the assessed domain on inter-rater reliability was not found. That is, there is no significant relationship between peer raters’ content knowledge in the assessed domain and their inter-rater error. The inter-rater error tells us, on average, how far a peer rater’s score deviates from the grand mean of all peer raters’ scores. One possible explanation for this insignificant result may be that the formula of inter-rater error, which is derived from McIntrye, Smith and Hassett (1984)’s “Leniency/Severity index”, may not be applicable in this peer assessment situation. The true score in the Leniency/Severity index formula is defined as the mean rating across all raters involved in assessment activity. Severe
raters are more likely to assign lower scores for all the assignments, while lenient raters tend to give higher scores across all the assignments than the server raters. This type of error is systematic and follows a constant pattern, therefore the Leniency/Severity index formula using the average score as the true score is based on the rationale that the measurement errors due to rater leniency and rater severity will cancel each other. However, in peer assessment, most rater errors are not systematic, but random, which do not seem to follow any particular discernible pattern. Some peer raters do not have enough content knowledge in the assessment domain to make reasonable judgements about their peers’ work, while some peer raters may lack rating skills or even rating motivation. Therefore, the average score of all peer raters may not be the true score where peer raters’ quality cannot be guaranteed. In hind sight, it makes sense that knowledge is not related to agreement with the rest of the raters. The mean of all other raters may represent the score given by a rater in the mid-range of knowledge. Both raters with high content knowledge and those with low content knowledge would depart from those with mid-level knowledge. That is, both high level knowledge and low level knowledge lead to high interrater errors while those with mid-level knowledge would have the least amount of interrater errors.

The second research question is about the effect of peer raters’ training experience on reliability and validity of peer assessment. Hypothesis 2 postulated that there is a negative correlation between peer raters’ training experience and their (a) inter-rater error, (b) intra-rater error, and (c) criterion error in peer assessment. Training experience includes two aspects: previous peer rating experience, and the one time self-paced training offered in this study. For previous peer rating experience, part (b) and part (c) of Hypothesis 2 was supported. The results showed that the more peer assessments a peer rater have done before, the higher intra-rater reliability and rating validity his or her peer assessment result will be. Similar to the effect of peer raters’ content knowledge in assessment domain on inter-rater reliability, no statistically
significant result was found concerning the relationship between peer raters’ previous peer rating experience and their inter-rater reliability.

As to the effect of one time self-paced training on the reliability and validity of peer assessment, no statistically significant result was found. That is, there was no statistically significant difference in inter-rater reliability, intra-rater reliability, or rating validity between those who received self-paced training and those who did not. These results are not consistent with the findings reported in existent literature. The most commonly used training method in peer assessment is collaborative training, in which instructors serve as trainers and lead discussions among students (Raczynski, Cohen, Engelhard, & Lu, 2015; Saito, 2008). The collaborative training of peer assessment can decrease the discrepancy between student ratings and teacher ratings (Liu & Lu, 2014), increase the self-consistency of peer raters (Weigle, 1998; Wiggleworth, 1993), and improve the inter-rater reliability (Xiao, 2008). The peer assessment training method used in this study is not collaborative training, but self-paced training. Self-paced training has been found to be equivalent to the more time-intensive and costly collaborative method in effectiveness on rating accuracy of experienced raters (Raczynski, Cohen, Engelhard, & Lu, 2015). Therefore, I hypothesized that peer raters who received self-paced training would generate more reliable and more valid peer assessment results than who did not. However, no significant result was found in the present study.

The nonsignificant result in this study may due to several reasons. First, the amount of training content and the length of training period might be insufficient. In this study, peer raters in self-paced training groups only receive one training package containing eight essays in total. There were two essay-topics, and for each topic, peer raters in training groups could only see four sample essays. Given the large variation of essay scores, four training essays for each topic might not be enough for peer rating to have a clear understanding of experts’ rating standard. As peer assessment consists of a set of complex skills (Sluijsmans et al., 2002), more training time might
be required to obtain the skills. The peer raters in Xiao (2008)’s study received either target-criteria-based peer-assessment training or principle-based peer assessment training through a whole semester, while the peer raters in the present study only received self-paced training once, which took one hour to complete.

Second, it is possible that the finding of equivalence in effectiveness on rating accuracy between self-paced training and collaborative training methods could not be generalized from experienced raters to novice peer raters. In a study by Raczynski et al., (2015), the equivalent effect on two training methods on rating accuracy was found among experienced raters, who were very familiar with the content knowledge in the assessment domain and had rated the previous year’s essays in the same exam. They might have already had some of the necessary assessment skills before they received either collaborative or self-paced training. In contrast, most peer raters in the present study are novice raters with only a little rating experience, and they might have lacked basic rating skills. Due to the large difference in rater quality, the equivalent result reported by Raczynski et al., (2015) may not be generalized to novice peer raters.

Finally, the nonsignificant finding may result from low fidelity implementation of self-paced training. In the present study, peer raters in training groups completed the training by themselves. After studying the rubric and the first four training-benchmark essays, as well as annotations/explanation of each score given to the benchmark essays in each domain, they were expected to answer five multiple-choice questions which were designed to determine if a peer rater had read the essays and explanations. Then those peer raters completed rating of the four practice essays. After that, they received the key sheet of the expert’s scores and explanation for the practice essay comparison. To further ensure that a peer rater in a training class had actually undergone the training, each student in training classes were asked to indicate on the practice scoring sheet whether they had over-scored, under-scored, or accurately scored each practicing-benchmark in multiple-choice format. Among 337 peer raters who were assigned to a training
class, 23 (6.1%) of them did not do the training at all, and 41 (10.88%) of them only completed part of the training task. The nine multiple choice questions in the training package were designed to examine whether a peer rater had completed the self-paced training seriously or not. As long as they had read the material carefully and compared their rating with the expert rating for each of the four practice essays as suggested in the instruction, they should have had no trouble to answering any of the multiple choice questions. However, it turns out 169 peer raters missed more than two questions. If missing more than two questions is used as a cutoff, almost half (44.56%) of the peer raters who were assigned in the self-paced training groups did not complete the training seriously. This low fidelity may be the main reason for the nonsignificant result of the training effect.

The third research question is about the effect of peer raters’ rating motivation on the rater reliability and rating validity of peer assessment. Hypothesis 3 postulated that there is a negative correlation between peer raters’ rating motivation in assessment domain and their (a) inter-rater error, (b) intra-rater error, and (c) criterion error in peer assessment. A peer rater’s rating motivation was measured by (1) whether or not the peer rater was offered monetary incentives, and (2) whether the peer rater received verbal persuasion about the value of seriously completing this assessment. For monetary incentives, all three parts of Hypothesis 3 were supported. The results showed that peer raters who were offered monetary rewards directly before conducting peer assessment tasks had smaller inter-rater error, intra-rater error, and criterion error in their peer rating than those who were not offered monetary incentive. In other words, providing monetary incentives to peer raters can increase inter-rater reliability, intra-rater reliability, and rating validity of their peer assessment result. For verbal persuasion about the value of completing peer assessment seriously, no statistically significant result was found.

The results indicate that monetary incentives can effectively increase peer raters’ rating motivation, and in turn improve the reliability and validity of peer assessment. Verbal persuasion
did not have a statistically significant effect on peer raters’ rating motivation in this study. In order to validate and compare the effectiveness of the two incentives (i.e., value-persuasion and monetary incentives) on students rating motivation, a 6-item motivation questionnaire based on expectancy-value theory was given to all participants. Peer raters who were offered monetary incentives had higher motivation scores (Mean = 27.666, SE= 0.246, 95% CI [27.183, 28.149]) than those who were not offered monetary incentives (Mean = 25.808, SE= 0.252, 95% CI [25.313, 26.302]), \( t =5.286, \text{df} = 806, p < 0.001 \). No statistically significant difference in motivation score was found between peer raters who received verbal persuasion and those who did not (\( t = 1.858, \text{df} = 806, p = 0.064 \)). These results confirmed the findings of Steedle’s (2010) survey which showed that being paid in money and prizes were the most valued incentives in enhancing students’ motivation in low-stake tasks. The effectiveness of monetary incentives on improving task motivation was found not only among college students (Deci, 1971; Eisenberger, Rhoades, & Cameron, 1999), but also in high schoolers (O’Neil, Abedi, Miyoshi, & Mastergeorge, 2005).

In this study, there is not enough statistical evidence to support the effects of verbal persuasion about the value of completing peer assessment seriously on peer rater’s rating motivation. This is not consistent with the existing literature on verbal persuasion and motivation (e.g., Gaspard et al., 2015; Brophy, 1999; Gaspard et al., 2015; Hullenman & Harackiewicz, 2009; Hulleman, Godes, Hendricks, & Harackiewicz, 2010, Hidi & Harackiewicz, 2000). One possible reason is value of the format and intensity in the verbal persuasion. The format of verbal persuasion in this study was a 5-minute presentation given by Chinese teachers. Compared to 90-minutes lessons for the whole class and relevant activities for each student about the value belief of mathematics achievement in the study by Gaspard et al. (2015), the format of verbal persuasion in this study seems too simple, and the 5-minutes presentation may not be enough to enhance a peer rater’s value belief on the benefits related to seriously assessing peers’ essays.
The fourth, fifth and sixth research questions are about interaction effects of peer raters’ rating motivation and content knowledge, rating motivation and training experience, and training experience and rating motivation on rater reliability and rating validity of peer assessment respectively. Hypothesis 4 postulated that there is a significant interaction effect of peer raters’ rating motivation and content knowledge on their (a) inter-rater error, (b) intra-rater error, and (c) criterion error in peer assessment. Part (b) and part (c) were supported. The result showed that the correlation between content knowledge and intra rater reliability for peer raters who received monetary incentives is stronger than it is for those not, and the same pattern exists in the correlation between content knowledge and rating validity. In other words, a peer rater’s rating motivation can influence the effect of their content knowledge on the intra-rater reliability and validity of their peer assessment result. The higher the motivation of a peer rater, the more likely it is for him or her to apply their content knowledge and make meaningful and accurate judgements of peers’ essays; thus, the validity of their peer assessment result will be higher. When peer raters’ rating motivation is low, they are less likely to rate their peers’ essay seriously, and they may not be willing to put in effort or apply their content knowledge to a peer assessment, which weakens the negative relationship between peer raters’ content knowledge and intra-rater error as well as criterion error of peer assessment. Such an interaction effect was not found on inter-rater error. One possible reason, as previously mentioned, is that the average score of all peer raters may not be the true score when peer raters’ quality cannot be guaranteed.

Hypothesis 5 postulated that there is a significant interaction effect of peer raters’ rating motivation and training experience on their (a) inter-rater error, (b) intra-rater error, and (c) criterion error in peer assessment. All the three parts were supported at class level, but none at the student level. The results showed that the correlation between training experience and inter-rater reliability for monetary incentives classes is stronger than it is for non-monetary incentives classes. In other words, providing monetary incentives can influence the effects of training
experience on the inter-rater reliability of peer assessment result at class level. For intra-rater reliability, the results indicate the correlation between training experience and intra-rater reliability for verbal persuasion classes is stronger than it is for non-verbal persuasion classes. That is teacher’s verbal persuasion of the value or benefit of doing peer assessment can influence the effects of training experience on the inter-rater reliability of peer assessment result at class level. This verbal persuasion intervention also influence the training effect on rating validity. Both monetary incentives and verbal persuasion affect training effect by increase students’ motivation. The higher the motivation of a peer rater, the more likely it is for him or her to apply the rating skills they obtained in a previous training experience to make meaningful and accurate judgements of peers’ essays. When peer raters’ rating motivation is low, they are less likely to rate their peers’ essays seriously, and they may not be willing to put in effort or apply their rating skills to a peer assessment. This weakens the relationship between peer raters’ training experience and the reliability and validity of peer assessment.

Hypothesis 6 postulated that there is a significant interaction effect from peer raters’ content knowledge and training experience on their (a) inter-rater error, (b) intra-rater error, and (c) criterion error in peer assessment. This hypothesis was not supported in this study. There is not enough statistical evidence to support the claim that a peer rater’s training experience can influence the effect of their content knowledge on the inter-rater reliability and validity of their peer assessment result. The low fidelity implementation of self-paced training may also result in the nonsignificant interaction effect of peer rater’s training experience and content knowledge in assessment domain on the reliability and validity of peer assessment.
6.2. Implication for educational practice

The findings in the present study indicate that the reliability and validity of peer assessment can be affected by peer raters’ quality characteristics, including content knowledge in the assessed domain, training experience, and rating motivation. Knowing how peer rater’s quality characteristics influence the reliability and validity of peer assessment may help teachers to identify high quality peer raters, to have a better peer assessment design, and to improve the reliability and validity of peer assessment. When peer assessment is used as a source of summative assessment, high reliability and validity are not only desired, but are also required out of concern for fairness. Based on the findings of the this study, this section provides suggestions for future educational practice, which includes (1) weighting scores assigned by different peer raters based their content knowledge in the assessed domain, (2) weighting scores assigned by different raters based on their previous peer rating experience, and (3) promoting peer raters’ rating motivation.

As shown in this study, a peer rater’s content knowledge in assessment domain is positively correlated with intra-rater reliability and rating validity. Peer raters with more content knowledge tend to have a better understanding of their peers’ assignments, provide more consistent ratings within themselves, and generate more accurate rating scores. Content knowledge in the assessed domain is the most basic requirement of making a meaningful judgement. However, peer raters content knowledge in the assessed domain varies dramatically. Some gifted students may master everything covered in the course and might even spend extra time developing their own expertise in the specific domain, while mediocre or poor performing students may feel confused about much of the content knowledge covered in the course. This variation is even larger in MOOCs, which have no prerequisites before a student can register. In order to have more reliable and more accurate peer assessment results, instructors should ascribe different weights to the rating scores provided by peer raters based their content knowledge.
The rating score given by a peer rater with more content knowledge should have a heavier weight than a score given by a peer rater with less content knowledge. When a peer rater has little to no content knowledge in the assessed domain, their judgements amount to no more than a guess, therefore it may be reasonable to exclude the rating score given by such a peer rater when calculating the final score of the assignment. When conducting real peer assessment, since the raters have yet to evaluate their peers’ work, the content knowledge used to determine the weight of an assignment score is the specific content of the work being rated, but some related knowledge. The related prior knowledge can be obtained from various approaches, which include, but not limit to, pretest score in the same area, previous homework grade, related project performance, and quiz scores. Different methods can be used to determine the weight for each peer rater’s score. For example, the weight for a particular rater on a particular assignment can be calculated by dividing that rater’s prior content knowledge score by average of the prior content knowledge scores of all the raters who are assigned to rate the same specific assignment.

Aside from differently weighting the scores of peer raters based on their content knowledge, instructors can also assign more weight to scores from raters with more previous peer rating experience. Raters’ previous peer rating experience can be easily obtained from a simple survey question which asks directly how many times a rater has previously participated in a peer assessment. Similar to determining the weight related to content knowledge in the assessed domain, the weight related to previous peer rating experience for a particular rater on a particular assignment can be calculated by dividing the number of previous participated peer assessments for that rater by average of the number of previous participated peer assessments all the raters who are assigned to rate the same specific assignment.

Having enough content knowledge in assessment domain and receiving regular training in assessment skills are two basic quality characteristics of peer raters that are necessary to achieve peer assessment results with high reliability and validity. A highly reliable and valid peer
assessment requires not only peer raters’ content knowledge and assessment skills, but also their motivation to rate. As shown in the present study, peer raters who were offered motivation incentives for rating accuracy generated more reliable and more valid peer ratings than those who were not provided with any motivation incentives. In real educational settings, instructors often use peer rating scores to determine grades without providing feedback or holding peer raters accountable for the quality of the ratings they provide (Friedman, Cox, & Maher, 2008). As a result, students may have little motivation to apply their content knowledge and assessment skills to provide meaningful ratings. If peer raters are not willing to take the rating process seriously, all peer assessment data will be meaningless, and high reliability and validity will be an unachievable goal for peer assessment. Therefore, teachers should give students some intervention to stimulate and enhance their motivation to complete peer assessment seriously.

From the perspective of expectancy-value theory, people feel motivated when they believe that they have the ability needed to complete a given task, that their effort will lead to some expected outcomes, and that the outcomes are personally valued. In order to encourage peer raters’ rating motivation, teachers need to provide intervention that targets students’ perceived value of completing a peer assessment seriously. In the present study, since teacher’s verbal persuasion did not seem to lead to a substantial effect, while money did. The result suggests that possibly appeals to extrinsic motivation works better than appeals to intrinsic motivation. This confirmed the Cameron, Banko, and Peirce (2001)’s finding that rewards given for low-interest tasks enhance free-choice motivation. Monetary reward was found to be an effective motivation incentive in the present study, but it is not the only effective way to promote rating motivation. In real educational settings, peer raters’ rating motivation can be stimulated and enhanced through various approaches, such as holding peer raters accountable for assessment quality (Cho, Schunn, & Wilson, 2006), informing peer raters about the usage of their ratings (Chen & Lou, 2004), and providing peer raters frequent feedback on their rating quality (Friedman, Cox, & Maher, 2008). Teachers can do a
survey to figure out what outcomes students value most, and provide corresponding incentives to increase their motivation to complete peer assessment seriously.

6.3. Limitations and future research

In this section, limitation and recommendations for future study are discussed. Although the present study was carefully designed to minimize possible threats to internal and external validity, there were still some limitations. Therefore, the results of this study should be interpreted with some caution, due to some methodological limitations. The present study can also inspire many areas for future study.

6.3.1. Sampling and generalizability of findings in this study

Sampling could be a limitation of generalizability of the findings in this study. Although the sample size (n=838) is very large among studies on peer assessment (Cho, Schunn, & Wilson, 2006), and the design is that of a cogent, evidence-based randomized cluster trial at the classroom level, all participants are from one typical high school in southern China. The characteristics of participants might have limited generalizability in terms of age, culture, language, ethnicity, and socioeconomic status. Also, the peer assessment task in this study is rating Chinese composition under the standardized Gaokao rubric, and the entire peer assessment procedure was completed in traditional classrooms. It is possible that the findings of this study are limited to this particular content domain. Future study could recruit students from several different schools from different regions. Rating essays often involves some subjective judgement of the rater and is very different from rating science or mathematics problems with objective answers. Future researchers can
extend their scope, look for different effects in other assessment tasks, and investigate the effects of different assessment tasks on the reliability and validity of peer assessment.

6.3.2. Treatment fidelity of peer rater training

The unexpectedly low treatment fidelity of the self-paced training in the present study might have impaired its effectiveness, albeit it is a time-saving and maneuverable training method. Given instructors common concerns that the traditional, collaborative training process is unduly effort-demanding and time-consuming, the present study introduced a self-paced training method to peer assessment based on Raczynski, Cohen, Engelhard, & Lu’s (2015) finding on the equivalent effect between self-paced training and collaborative training on rating accuracy. The entire self-paced training process took most peer raters one hour to complete. Researchers, in general, agree that peer raters may have difficulty understanding the rubric, and may interpret the terms used in the rubric differently than experts (Orsmond and Merry, 1996). Even a one-time, short term collaborative training should improve peer rating quality (Saito, 2008). Therefore, it was reasonable to believe that the self-paced training in the present study should have improved the reliability and validity of peer assessment. However, this training effect was not found in the present study, most probably due to an unexpectedly low level of treatment fidelity in actual implementation. Among 337 peer raters who were assigned a training class, 6.1% of them did not do the training at all, 10.88% of them only completed part of the training task, and around half (44.56%) of them did not complete the training seriously. Because of the high potential value of self-paced training methods in educational settings, future studies should examine the effect of this training method on peer rating quality after controlling the fidelity of implementation.
6.3.3. Motivation incentive methods in peer assessment

Peer raters’ motivation to rate was measured indirectly by whether or not they were offered motivation incentives due to the fact that there is no well-developed instrument for measuring peer raters’ rating motivation. The absence of such an instrument is surprising given the large body of evidence demonstrating the significant effect of rater motivation on rating quality (Salvenimi & Reilly, 1993) and the wide application of peer assessment (Falchikov & Goldfinch, 2000, Li et al., 2015). Therefore, it would be very helpful if researchers in the future can develop and validate an instrument to measure peer raters’ motivation to rate.

Two motivation interventions used in the present study are monetary rewards for accuracy and verbal persuasion about the benefit of completing peer assessment seriously. Offering monetary reward does lead to more reliable and more valid peer assessment results, however, this motivation intervention could not be widely and frequently adopted as a regular approach in real educational settings. In contrast, verbal persuasion is very affordable and practical. Previous research has shown that verbal persuasion can increase students’ task motivation (Brophy, 1999; Gaspard et al., 2015; Hullenman & Harackiewicz, 2009). As discussed, the non-significant effects of verbal persuasion on peer rating reliability and validity in this study may be due to the format and intensity of the verbal persuasion. One potential direction for future study is to investigate the effects of verbal persuasion, using different formats and intensities, on peer rating quality. Another direction for future study is to explore and examine the effects of other practical motivation incentive methods, such as offering extra credits, providing academic awards, and enforcing accountability for rating quality.
6.3.4. Peer rater’s demographic and psychological characteristics

Although peer raters’ characteristics include many aspects, the present study focused on the effects of quality characteristics (i.e., peer raters’ content knowledge in assessment domain, training experience, and motivation to rate) on the reliability and validity of peer assessment. Given that these quality characteristics are found as significant influential factors for some aspects of reliability and validity, other aspects of peer raters’ characteristics, such as demographic characteristics and psychological characteristics, may also affect the reliability and validity of peer assessment. With respect to demographic characteristics, gender and race or ethnicity bias in assessment have been examined for decades (O’Neill, 1985; Falchikov & Magin, 1997; Ghorpade & Lackritz, 2001; Tucker, 2014, Watson, Barnier & Pavur, 2010), while there is little to no research examining the effect of gender on accuracy of peer assessment. Additionally, age is another demographic variable that might influence the reliability and validity of peer assessment. Some people believe that age can be an indicator of maturity and experience (Watson, Barnier & Pavur, 2010), including rating experience, which can in turn can affect peer rater’s rating ability. Little research has investigated the direct effect of peer rater’s age on rating accuracy. Future studies can investigate how rater’s demographic characteristics, including but not limited to gender, ethnicity, and age, influence the reliability and validity of peer assessment.

When peer assessment is conducted in MOOCs with students from all over the world, there should be a large variation in peer rater’s English language proficiency level, geographic origin, and culture orientation. The effects of these demographic characteristics can also be examined in future research.

As to peer rater’s psychological characteristics, only a few studies have investigated the possible effects on the reliability and validity of peer assessment. AlFallay (2004), selected five psychological and personality traits and examined how these selected traits of a peer rater can
influence the accuracy of peer assessment in a second language classroom. Future research can expand the types and range of psychological characteristics and examine their effects on the reliability and validity of peer assessment in various disciplines.

6.3.5. Other types of rater error

The accuracy of peer assessment can be evaluated by a variety of peer rater errors. The present study mainly focused on three categories of peer rater error: inter-rater error, intra-rater error, and criterion error. To have a deep understanding of the pattern and nature of peer rater errors, future studies can investigate different types of rater errors in peer assessment, such as the halo effect, central tendency, rater severity, and the restriction of the rater, and examine how these rater efforts can be detected in peer assessment. In alignment with this research, future researchers can also investigate how a specific type of peer rater errors is affected by a peer rater’s quality characteristics, and what strategies or training can be used to minimize the effects of different types of peer rater error.

6.3.6. Identify high quality peer raters and improve the accuracy of peer assessment

Due to common concerns about the accuracy of peer assessment results, several approaches have been developed and applied to peer assessment in both traditional classrooms and MOOC environments (Suen, 2014). For example, University of California – Los Angeles developed a Calibrated Peer Review system to improve peer rating accuracy by assigning weights to each peer rater’s rating based on their relative degree of accuracy (Reynolds & Moskovitz, 2008). To refine and modify the Calibrated Peer Reviews method, Xiong, Goins, Suen, Pun, and Zang (2014) proposed the Credibility Index approach, which takes into account peer rating errors
from three sources: inaccuracy, inconsistency, and intransferability. This model might be further modified and refined by including peer rater’s quality characteristics, and/or demographic, as well as psychological characteristics. Aside from refining existent approaches, future researchers can also develop new models using peer rater’s characteristics to improve the reliability and validity of peer assessment.
References


Xiao, Y. (2010). *The Effects of Training in Peer Assessment on University Students' Writing Performance and Peer Assessment Quality in an Online Environment*. ProQuest LLC.

789 East Eisenhower Parkway, PO Box 1346, Ann Arbor, MI 48106.


# Appendix A

## The Gaokao Chinese Writing Rubric

<table>
<thead>
<tr>
<th>RUDIMENTS LEVEL</th>
<th>Level 1 Exceptional (20-16)</th>
<th>Level 2 Skilled (15-11)</th>
<th>Level 3 Developing (10-6)</th>
<th>Level 4 Inadequate (5-1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Content (20)</td>
<td>Effectively addresses the task</td>
<td>Effectively addresses the task</td>
<td>Addresses the task well, though some points may not be fully elaborated</td>
<td>Limited development in response to the task</td>
</tr>
<tr>
<td></td>
<td>Effectively and consistently focuses on a clearly identified purpose and topic throughout</td>
<td>Adequately focuses on an identified purpose and topic</td>
<td>Somewhat focuses on an identified purpose and topic</td>
<td>Shows a minimal focus on purpose and topic</td>
</tr>
<tr>
<td></td>
<td>Provides substantial content</td>
<td>Provides somewhat substantial in content</td>
<td>Provides insufficient content</td>
<td>Provides unsuitable content</td>
</tr>
<tr>
<td></td>
<td>Delivers healthy central idea</td>
<td>Delivers healthy central idea</td>
<td>Delivers somewhat healthy central idea</td>
<td>Delivers unhealthy central idea</td>
</tr>
<tr>
<td></td>
<td>Conveys true and deep emotion</td>
<td>Conveys true emotion</td>
<td>Conveys somewhat true emotion</td>
<td>Conveys deceptive emotion</td>
</tr>
<tr>
<td>Expression (20)</td>
<td>Meets the stylistic requirements</td>
<td>Meets the stylistic requirements</td>
<td>Somewhat meets the stylistic requirements</td>
<td>Fails to meet the stylistic requirements</td>
</tr>
<tr>
<td></td>
<td>Uses a clear and effective organizational structure</td>
<td>Uses a complete organizational structure</td>
<td>Uses a somewhat complete organizational structure</td>
<td>Uses a chaotic organizational structure</td>
</tr>
<tr>
<td></td>
<td>Uses sophisticated and fluent language</td>
<td>Uses precise language</td>
<td>Uses somewhat precise language</td>
<td>Use imprecise language, and has frequent grammatical errors</td>
</tr>
</tbody>
</table>
|                  | Expresses in neat handwriting | Expresses in clear handwriting | Expresses in somewhat clear handwriting | Expression in sprawling and
<table>
<thead>
<tr>
<th>Development Level</th>
<th>Feature (20)</th>
<th>Development Level</th>
<th>Development Level</th>
<th>Development Level</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Develops the whole idea profoundly</td>
<td>Develops the whole idea somewhat profoundly</td>
<td>Develops the whole idea a little profoundly</td>
<td>Develops profoundly in single sentence(s)</td>
</tr>
<tr>
<td></td>
<td>Supplies thorough and convincing examples /evidence</td>
<td>Supplies adequate relevant examples / evidence</td>
<td>Supplies some relevant examples / evidence</td>
<td>Supplies single piece(s) of relevant evidence</td>
</tr>
<tr>
<td></td>
<td>Thoroughly shows literary grace</td>
<td>Adequately shows literary grace</td>
<td>Shows somewhat literary grace</td>
<td>Partly shows literary grace only in single sentence(s)</td>
</tr>
<tr>
<td></td>
<td>Thoroughly shows creativity</td>
<td>Adequately shows creativity</td>
<td>Shows somewhat creativity</td>
<td>Partly shows insight only in single sentence(s)</td>
</tr>
</tbody>
</table>

Notes:

(1) The rating of Rudiment Level should mainly focus on topic, content, language, and style using a holistic view. Meeting the stylistic requirements means meeting the requirement of the article type selected by the examinees.

(2) The rating of Development Level should be based on the four criteria. Instead of covering all the four criteria, as long as an essay achieve one of the four criteria, corresponding scores can be awarded based on the it quality.

(3) Two points deducted for no title of the essay composition. No point deduction for 1~2 wrong characters (typos); one point deduction for 3 wrong characters; two points deduction for 4 wrong characters; three points deduction for 5 wrong characters, four points deduction for six and more than six wrong characters. Repeating wrong characters do not count. For essays that did not meet the required minimum number of characters, there was a one point deduction for every 50 missing characters.
(4) Plagiarized essays can only have a score within the score range of Level 4 for the Rudiment Level, and have zero point for the Development Level.

(5) Essays do not connect to the topic and task can only have a score within the score range of Level 4 for the Rudiment Level, and have zero point for the Development Level.

(6) Essays with less than 400 characters can only have a score within the score range of Level 4 for the Rudiment Level, and have zero point for the Development Level.

高考作文评分标准表

<table>
<thead>
<tr>
<th>基础等级</th>
<th>发展等级</th>
</tr>
</thead>
<tbody>
<tr>
<td>内容</td>
<td>表达</td>
</tr>
<tr>
<td>20分</td>
<td>20分</td>
</tr>
<tr>
<td>一等 (20—16)</td>
<td>二等 (15—11)</td>
</tr>
<tr>
<td>符合题意</td>
<td>符合题意</td>
</tr>
<tr>
<td>中心突出</td>
<td>中心明确</td>
</tr>
<tr>
<td>内容充实</td>
<td>内容较充实</td>
</tr>
<tr>
<td>思想健康</td>
<td>思想健康</td>
</tr>
<tr>
<td>感情真挚</td>
<td>感情真实</td>
</tr>
<tr>
<td>符合文体要求</td>
<td>符合文体要求</td>
</tr>
<tr>
<td>结构严谨</td>
<td>结构完整</td>
</tr>
<tr>
<td>语言流畅</td>
<td>语言通顺</td>
</tr>
<tr>
<td>字迹工整</td>
<td>字迹清楚</td>
</tr>
<tr>
<td>深刻</td>
<td>较深刻</td>
</tr>
<tr>
<td>丰富</td>
<td>较丰富</td>
</tr>
<tr>
<td>有文采</td>
<td>较有文采</td>
</tr>
<tr>
<td>有创意</td>
<td>较有创意</td>
</tr>
</tbody>
</table>

说明：
①基础等级评分以题意、内容、语言、文体为重点，全面衡量。符合文体要求，指符合考生所选文体的要求。

②发展等级评分，依据4个评分点，不求全面，只需一点突出，即可按等评分，直至满分。

③未拟题目扣2分。出现错别字，1～2个不扣分，3个扣1分，4个扣2分，5个扣3分，6个以上(含6个)扣4分，重复不计。不足字数者，每少50字扣1分。

④抄袭的文章，“基础等级”在四等之内评分;“发展等级”不给分。

⑤判偏题的，只能评四等，但“发展等级分”一分不给。

⑥不足400字的，只能评四等，但“发展等级分”一分不给。
写作题目要求 1：阅读下面材料，根据要求写一篇不少于 800 字的文章。（60 分）

某校有一座建于上世纪 50 年代末的教学楼，从这里走出去一批批优秀人才，其中不乏知名学者、作家、主持人。如今，这座矮小、简陋、老旧的教学楼已经不能满足学校现代化发展的需要。考虑到它的特殊性，学校想保留，但有人认为学校用地有限，修缮、养护还要花钱，建议着眼发展，将其拆除。为此，学校很犹豫。不少师生和校友表示非常关注，曾在该校就读过的某知名作家还专程赶回母校拍照留念，并积极争取留住它。

对于以上事情，你怎么看？请给该校校长、该知名作家或其他相关方写一封信，表明你的态度，阐述你的看法。

要求综合材料内容及含义，选好角度，确定立意，完成写作任务。明确收信人，统一以“小林”为写信人，不得泄露个人信息。
1号作文：

尊敬的校长，

您好，我知道贵校最近正为老教学楼的事情烦心，我恰恰对此事有些看法。

这座老楼历史悠久，战功赫赫，培养了不少有用之材。对于这些人来说，这座楼即使在他们的整个生命中，都占有着不小的地位。这座楼象征着他们的回忆，青春。而留下这座楼的理由就很充分了，情怀二字足矣。

但情怀太泛滥了，看电影看情怀，买手机买情怀，连喝个奶茶都喝情怀。那样的情怀，真的还有太多意义吗？或许相比这件事，我所说的，情怀都显得太小了，但我只是想说明一点，社会普遍的现象若带动的是精神层面的泛滥，那这种精神势必会改变性质，丧失价值。

从这些贵校走出的人中，那些渴望保留教学楼的人，不妨想一想，我们为什么不能办一个送别仪式，好好的送它走，把它封存在珍贵的影视资料中，更好的纪念它呢？历史的长河向我们证明，时间会逐渐带走一切，这座楼迟早要被拆毁，与其在众人漠视目光中离开，倒不如在感动中得到新生！我们常常认为自己能把握一切东西，永远保持在自己的控制范围内。
内，但实际上皮炎的居民甚至来不及为自己\n的塑像摆一个好看的位置就死去了。把事情控\n制在可控范围内更好的去完成。就像科比即使\n可以奇迹般地奇迹再打两年，但坏是要求了好\n的谢幕，这样的选择难道不明智至极吗？

从学校的职能来说，学校毕竟是需要承担\n教育的责任的，不断加强教育水平是学校的本\n分，而要做到这样，就要发展，这座教学楼显\n然已经阻碍了学校的发展，拆除它对教育水平\n提高有帮助，那更应该毫不犹疑了。换个角度\n讲，如果教学水平提高，那种新的新人才，不\n是就会把新教学楼当作纪念了吗？

前段时间去电影院看了《老炮儿》和《大\n小欢喜》。这是两部充满了情怀的电影，看完\n了我久久不能忘怀，现在仔细一想：“老炮儿\n之所以那么感动观众，其实正是因为那份”守\n不住的“倔”，再怎么变也要变老，但那股子\n劲儿却也传承下去了，就像一个神话般的圣\n诞，支撑着几百年的世人，情怀，其实就是\n自己的事，传承靠的是精神，好的东西传下去，\n旧的载体没了，新的载体出来了，精神恒久远，\n这情怀。\n
所以说呀，对校方您认为怎么样我不了\n解，但留下来的理由除了情怀我想是别无其他。\n若是情怀吧，我希望您坚持，情怀很流行，也\n很好。但希望您记着别用情怀绑架未来了！

小林
2号作文：

敬爱的校长：

您好。

通过网上的介绍我了解到了关于你们学校目前的状况。一所历史悠久、培养了这么多优秀人才的学校，可以说是屈指可数了。我认为既然这所学校被创办了，那么它就有继续存在的必要。无论如何，这所学校不应该消失。

上世纪60年代末至今已有将近50年的时间，可以说这栋建筑物陪伴了许多老师和学生们走过了大半个人生历程。无数的人在这里追求和放飞理想，梦想的种子在此生根发芽，这所教学楼陪伴了老师们学生们度过一个又一个春夏秋冬。

那些曾经无法磨灭的改变和回忆，这栋教学楼已经搭建起了漫长岁月桥梁，因此，它不可轻易地被毁坏拆掉。

知名学者、作家、主持人以及各种优秀人才才在此诞生。在这样人才济济的学校，师资水平肯定也是顶尖的，肯定也不缺乏良好的学习环境。如此良好的校风和口碑可以带动更多有梦想的学生来到此学习，以学长学姐们为目标不断努力，以此为国家培养出更多优秀的学生。就如中国最著名的清华大学，它们就是以优秀人
太多而闻名，吸引很多的学生以此为目标不断奋斗。这栋教学楼也是如此，所以它就有存在的价值。

许多老师也将他们的青春和心血投入到这所学校当中，有些甚至将自己的人生都投入到其中。如果学校拆除了，他们的大半生的心血可能就毁于一旦。或许学校的消失对于绝大多数人来说只是一个建筑物的消失，但对这些老师们来讲，他们失去了青春，结束了教师生涯。对于一个老师来说，草子是一次次生灭的渴望啊。

一所学校的消失究竟意味着什么，这我

说不清楚也道不明的。只知道，既然一个东西

被创造了，它就被赋予了存在的意义。’’简陋

简陋，老旧’’’’不能满足现代化发展”“用她有限

修缮，养护要花钱’’这些都不足以成为这个

学校必须被拆除的理由。这些想法都太过与表

面。对于我来说，这所学校就如同名胜古迹一

样需要去维护。不是那样的话，这所学校既然被

创造，它就有存在的意义。”

小林
写作题目要求 2：阅读下面的材料，根据要求写一篇不少于 800 字的文章。(60 分)

蜀南竹海风景区的 10 亩国有林，被划定为“刻字林”，引导有“刻字”需求的游客前往刻字专区，以此杜绝游客乱刻乱画的行为。因此，每年有近五千棵竹子惨遭游客蹂躏。

德国柏林墙上满是千奇百怪的图画和文字，这满墙的涂鸦反而让柏林墙除了承载历史外，又多承载了一种文化。

对以上事情，你怎么看？请写一篇文章，表明你的态度，阐述你的看法。

要求综合材料内容及含意，选好角度，确定立意，明确文体，自拟标题，不要套作，不得抄袭。
3号作文

<table>
<thead>
<tr>
<th>题目</th>
<th>粗鄙年轻扯即文化</th>
</tr>
</thead>
<tbody>
<tr>
<td>故宫藏品被胡乱涂画已不是新闻，遍布世界各地的“到此一游”给中国人的头顶带来一片片阴霾。甚至在黄帝海南，有一片国有林专供游客刻字。每年有近五千棵树木惨遭涂鸦蹂躏，这仿佛在控诉游客的不文明暴行，揭露刻字涂鸦的罪恶。但也有人说，柏林墙上的涂鸦是一种文化，它装点修缮了柏林墙。</td>
<td></td>
</tr>
<tr>
<td>涂鸦也好，刻字也罢，都是平川流水，引人往左，则藤田苗，导之以右，则抑房屋屋。涂鸦本没有错，错在它出现的位置不对，如果加以引导，而不是一味责难，涂鸦不文明的涂鸦就能成为一种文化，粗鄙年轻扯即文化。</td>
<td></td>
</tr>
</tbody>
</table>
| 涂鸦、刻字，其实就是微型的画作和雕刻，是一种文化形式。中国人似乎喜欢留下自己的痕迹，大诗人那酒吟诗，吟罢便留在墙上，收藏者曾玩书画，看到喜欢的就画上印题诗，国人的涂刻天赋似乎是天生的，流淌在民族血液中的。不只是局限于中国，国外街头涂鸦更是流行，在繁华紧促的纽约，一位艺术家在十九年里在街头画上小动物，一条平淡无奇的街道变成了梦的国度，急忙的纽约人曾经过时也会慢下来。柏林墙上的图画文字，尽管千奇百怪，但的确让这面背负着沉重历史的老墙
轻松起来，将文化与历史相交融。

涂鸦刻画，自古就有，而且并不只是中国
人如此，外国人甚至更加。

以其便捷的优势，让人们的得以更好地发挥个性。
一个名字可能不足为奇，但千万个名字集成的
民族精神就可以歌颂官员的政绩，一个抽象化
的涂鸦或许有些可笑，但如果加以引导，它能
汇聚，涂鸦就能展现出民族的创造力。

涂鸦的却很好，可乱涂乱刻就截然不同了。
在故宫藏品上乱刻画，不仅仅是留给工作人员
一个难题，更是对藏品的极大破坏与不尊重。
不令时有地聚集的涂刻就背起之痛，当止于
始发。不止要阻止这类不文明的行为，更要禁
止，并加以引导。

对于涂鸦刻画，我们应加以鼓励，但不能
不去规范，不去引导，林海中的刻字林不一定
只是一种浪费，也许有一天，它们会被写在
你上的古书一样珍贵，对于不文明的粗暴加以
导引导，或许是一种美好的文化。会成为
４号作文：

题目论“刻字林”

在翠袖竹海风景区，专门设立“刻字林”，是为了满足游客们刻字的需求，林木墙也成为了游客表达需求的工具。

我不赞成设立“刻字林”。我认为这并非真正杜绝游客乱刻乱画的目的。

首先，我认为设立“刻字林”治标不治本。看上去好像游客已经不再在这乱涂乱画了，似乎

除非有专门的刻字专区，他们才不会乱写乱画，不再随他选放“刻字”的需求，但是，这根本起

不到不让游客乱刻乱画的作用。政府相当于巧妙地偏架换柱，将原本游客该刻的字从墙壁

转移到了刻字林上。游客内心对刻字需求并没有减少。原因是我们没有找到问题所在。游

客为什么要乱刻乱画？因为他们缺乏规范意识，缺乏正确的价值观，不知道在景区刻字是一种

不道德的，应该被严厉禁止的行为。他们不知道这种行为会破坏景点和冲击正确的价值观。

一切都是因为“无知”二字。这种无知绝对不

是政府设个刻字林就能转变的，就算在这里，

在专设刻字林的地方，游客们不刻景点了，但

在别处，他们还是会刻。他们的刻字需求一直

存在着。所以我认为设立“刻字林”是一种无用功。

其次，我认为设立“刻字林”是一种潜在
识上的纵容。好像在说：你们在这片林子随便刻，尽管发泄你们的欲望，但是在那片大林子就不要刻了。那个地方，就像幼儿园老师发糖果给小朋友，当发到最后一个小朋友时，糖果没了。老师摸摸他的脑袋说：

“没有糖果，就给你个面包吧！”政府相当于搂着游客的脑袋安抚他们说：“那片林子不要刻，给你这片林子刻吧！”

这根本不会让游客感受到半点“杜绝”的意思在里面，反而是一种对他们刻字欲望的纵容，让他们以为，只要换个替代品，你刻字的行为就没有错。但是，如果到了没有刻字林的地方，他们想刻字的欲望只会更强烈。甚至还会有抱怨：景区自己不设置个刻字专区，乱涂乱画就不要怪我啦！

再者说，没刻字林，让那些没有刻字经验的人也养成这种恶习。看着别人都去了刻字林，自己不去好像也不好。去了之后发现那些人“纵欲”地这么痛快，自己也来试一下。久而久之，在，也养成了刻字的习惯。

所以，政府应该做的不是设刻字专区，而是从根本上入手，教育人们这种做法的错误之处，给他们的价值观补上一课，甚至可以重金罚款作为辅助手段，而不是偷换概念，让他们刻在刻字林上就完事了。

综上所述，我不支持刻字林”的设立。
Appendix C

Original Verbal Persuasion Lecture in Chinese

“同学们，在高考语文卷中，作文占分比重最大，为了提高你们的作文分数，我们进行一项同学互评作文的练习。通过这项练习，你们首先可以明白高考作文的评分标准，知道要想写出高分作文需要达到哪些要求。其次，在批改同伴作文的过程中可以学习其他同学写的好词好句和好的观点，也可以从别人容易犯的错误中吸取教训。最后，批改同伴的作文，有利于你从阅卷人的角度去思考作文的写法，从而更深入地了解和体会如何布局谋篇、遣词造句。所以，希望大家认真批改同伴的作文，相信你能从中获益匪浅。”
Appendix D

Practicing Scoring Sheet

Class: ___________________  Name: ___________________

Gender: ___________________  Age: ___________________  Ethnicity: ___________________

** Four multiple choice questions (focusing on the content of the four training-benchmark essays)

<table>
<thead>
<tr>
<th>Essay Number</th>
<th>Content Level</th>
<th>Content Score</th>
<th>Expression Level</th>
<th>Expression Score</th>
<th>Feature Level</th>
<th>Feature Score</th>
<th>Total score Level</th>
<th>Total score Score</th>
<th>Hand Writing</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

** Four multiple choice questions (focusing on the comparison of the peer rater’s score and the teacher’s score of the four practicing-benchmark essays)
练习评分表
姓名:______ 班级:______ 性别:______ 出生年月:______ 民族:______ 籍贯:______．

**四道选择题（针对四篇训练标杆作文的内容）

<table>
<thead>
<tr>
<th>作文号</th>
<th>内容</th>
<th>表达</th>
<th>特征</th>
<th>总分</th>
<th>书写得分</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

***四道选择题（比较四篇练习作文的得分与老师评分差）
## Appendix E

**Peer Assessment Scoring Sheet**

<table>
<thead>
<tr>
<th>Class:</th>
<th>Name:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender:</td>
<td>Age:</td>
</tr>
<tr>
<td>Ethnicity:</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Essay Number</th>
<th>Content Level</th>
<th>Content Score</th>
<th>Expression Level</th>
<th>Expression Score</th>
<th>Feature Level</th>
<th>Feature Score</th>
<th>Total score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

## 同伴互评得分表


<table>
<thead>
<tr>
<th>作文号</th>
<th>内容等级</th>
<th>内容分数</th>
<th>表达等级</th>
<th>表达分数</th>
<th>特征等级</th>
<th>特征分数</th>
<th>总分等级</th>
<th>总分分数</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Appendix F
Motivation Questionnaire

1. How attractive is it to you to provide accurate peer rating on these essay compositions?

- Very Unattractive - Unattractive - Somewhat Unattractive - A Little Attractive - Somewhat Attractive - Attractive - Very Attractive

2. How much effort have you exerted to complete the peer rating? (Little Effort, Great Deal of Effort)

- Very Little Effort - A Little Effort - Some Effort - Average Effort - Much Effort - Considerable Effort - Great Deal of Effort

3. How much time did you spend in this peer rating in total?

- No more than 10 minutes - 11-20 minutes - 21-40 minutes - 41-60 minutes - 61-90 minutes - 91-120 minutes - More than 2 hours

4. Please indicate the confidence level of your ability to provide accurate peer rating.

- Extremely Low - Very Low - Low - Average - High - Very High

- Extremely High

5. Please indicate your seriousness level of rating of these essay compositions.

- Extremely Unserious - Very Unserious - Unserious - Average - Serious - Very Serious

- Extremely Serious

6. How much do you care about whether the scores you give for these essay compositions are accurate?

- Not At All - Little - A Little - Average - Some - Much - Very Much
动机态度问卷

1. 你对认真准确评这些作文有多大兴趣？
   ○1 毫无兴趣 ○2 几乎没有兴趣 ○3 不太有兴趣 ○4 有一点兴趣 ○5 比较有兴趣
   ○6 有较大的兴趣 ○7 非常有兴趣

2. 你付出了多少努力来评这些作文？
   ○1 极少的努力 ○2 很少的努力 ○3 较少的努力 ○4 一般的努力 ○5 较大的努力
   ○6 很大的努力 ○7 极大的努力

3. 你一共用了多长时间评这些作文？
   ○1 不到 10 分钟 ○2 16-30 分钟 ○3 21-40 分钟 ○4 41-60 分钟 ○5 61-90 分钟 ○6 91-120 分钟 ○7 2 小时以上

4. 请标明在多大程度上你相信自己有能力给出准确的评分
   ○1 信心极低 ○2 信心很低 ○3 信心较低 ○4 信心一般 ○5 信心较高 ○6 信心很高 ○7 信心极高

5. 请标明你评这些作文的认真程度。
   ○1 极不认真 ○2 很不认真 ○3 较不认真 ○4 一般认真 ○5 比较认真 ○6 很认真 ○7 非常认真

6. 你在多大程度上在乎你对这些作文给出的评分 Yes 准确的？
   ○1 毫不在乎 ○2 很不在乎 ○3 较不在乎 ○4 一般在乎 ○5 比较在乎 ○6 很在乎 ○7 非常在乎
EXPERTISE & SKILLS

- Comprehensive knowledge in applied statistics, educational statistics, and psychological measurement.
- Expertise in quantitative methods, such as structural equation modeling, hierarchical linear modeling, experimental design, categorical data analysis, and longitudinal data analysis
- Rich experience in qualitative and quantitative methods for educational research, assessment, evaluation, focus group interview, and instrument development
- Expertise in large data cleaning, storage, and analysis: Proficiency in MySQL, R, SAS and data analysis software including SPSS, HLM, Mplus, Flexmirt, MS Access, and MS Excel
- Extensive experience in teaching and counseling as well as working with faculty from multiple scholarly domains

EDUCATIONAL BACKGROUND

- **Ph.D.,** Educational Psychology, The Pennsylvania State University, University Park, PA, Expected: August 2017
- **Master of Applied Statistics,** The Pennsylvania State University, University Park, PA, Expected: August 2017
- **Master of Art** in Modern and Contemporary Chinese History, Dalian University of Technology, Dalian, China, 2008
- **Bachelor,** Chinese Language and Literature, Liaoning University, Shenyang, China, 2004

WORKING EXPERIENCE

- **Graduate Assistant,** Department of Teaching and Learning with Technology (TLT), The Pennsylvania State University, University Park, PA, Aug. 2015- Present
- **University Graduate Fellow,** The Pennsylvania State University, University Part, PA, Aug./2013- Aug./2015
- **Teacher and Counselor,** Dalian Maple Leaf International School, Dalian, China, Aug./2008-Oct./2009