UNDERSTANDING ESSAY RATING AS A SOCIALLY MEDIATED ACTIVITY: THE CASE OF A HIGH-STAKES ENGLISH TEST IN CHINA

by

Yi Mei

A thesis submitted to the Graduate Program in Education in conformity with the requirements for the Degree of Doctor of Philosophy

Queen’s University
Kingston, Ontario, Canada
September, 2016

Copyright © Yi Mei, 2016
Abstract

Most essay rating research in language assessment has examined human raters’ essay rating as a cognitive process, thus overlooking or oversimplifying the interaction between raters and sociocultural contexts. Given that raters are social beings, their practices have social meanings and consequences. Hence it is important to situate essay rating within its sociocultural context for a more meaningful understanding. Drawing on Engeström’s (1987, 2001) cultural-historical activity theory (CHAT) framework with a sociocultural perspective, this study reconceptualized essay rating as a socially mediated activity with both cognitive (individual raters’ goal-directed decision-making actions) and social layers (raters’ collective object-oriented essay rating activity at related settings). In particular, this study explored raters’ essay rating at one provincial rating centre in China within the context of a high-stakes university entrance examination, the National Matriculation English Test (NMET).

This study adopted a multiple-method multiple-perspective qualitative case study design. Think-aloud protocols, stimulated recalls, interviews, and documents served as the data sources. This investigation involved 25 participants at two settings (rating centre and high schools), including rating centre directors, team leaders, NMET essay raters who were high school teachers, and school principals and teaching colleagues of these essay raters. Data were analyzed using Strauss and Corbin’s (1990) open and axial coding techniques, and CHAT for data integration.

The findings revealed the interaction between raters and the NMET sociocultural context. Such interaction can be understood through a surface structure (cognitive layer) and a deep structure (social layer) concerning how raters assessed NMET essays, where the surface structure reflected the “what” and the deep structure explained the “how” and “why” in raters’ decision-
making. This study highlighted the roles of goals and rules in rater decision-making, rating tensions and raters’ solutions, and the relationship between essay rating and teaching. This study highlights the value of a sociocultural view to essay rating research, demonstrates CHAT as a sociocultural approach to investigate essay rating, and proposes a direction for future washback research on the effect of essay rating. This study also provides support for NMET rating practices that can potentially bring positive washback to English teaching in Chinese high schools.
Acknowledgements

“No boat makes a journey alone.”

This dissertation could not be possible without the help and support of many.

First of all, I would like to thank my supervisor/mentor and committee members, Drs. Liying Cheng, Christopher DeLuca, and Khaled Barkaoui for taking this journey with me. They challenged me to pursue this new perspective. Their ideas have pushed my thinking and I am grateful for having the opportunity to learn from and work with them. I look forward to our future conversations and collaborations.

Thank you to my chair and examining committee members, Drs. Terry Krupa, Ben Bolden, Jan Mennell, and Guillaume Gentil, for posing challenging and thoughtful questions that will extend my understanding about how to apply a sociocultural perspective to promote better research and practices in writing assessment.

Thank you to my study participants, for generously sharing their time and experiences with NMET essay rating that helped me gain a deeper understanding of their work.

I have met many wonderful people at Queen’s who have made this journey a life-changing experience. To faculty, staff, and colleagues within Assessment and Evaluation Group (AEG), at Faculty of Education, Education Library, and School of Graduate Studies, thank you for supporting my growth in so many ways. To my friends, who discussed and edited numerous drafts of my research proposals, who went on conference trips with me, and who joined me in the “Thesis Persistence” writing programs, thank you for being there for me.

Thank you to my colleagues at Shanxi University and Beijing Normal University, who never hesitated to share their expertise and provided help whenever I needed it.
My special acknowledgements go to Dr. Zunmin Wu, who encouraged me to “aim high and walk every step toward the goal” of becoming a researcher; to Dr. Don Klinger, who gave me invaluable advice on academic and nonacademic matters; to Dr. Ulemu Luhanga, who gave me fresh ideas when discussing CHAT; and to Wei Yan, who generously worked as a back translator for my data collection instruments.

This research is supported by grants from The International Research Foundation for English Language Education (2015 Doctoral Dissertation Grant), Educational Testing Service (2015 TOEFL Small Grant for Doctoral Research in Second or Foreign Language Assessment), and Queen’s University (2014 Graduate Dean's Travel Grant for Doctoral Field Research).

Finally, I would like to thank and dedicate this dissertation to my loving family, who always firmly stood by me, in good times or bad; and most of all, to King, the first reader of my work, for his unconditional love and support, and never doubting that I could get here.
# Table of Contents

Abstract ............................................................................................................................................................. ii

Acknowledgements ........................................................................................................................................ iv

Table of Contents ............................................................................................................................................. vi

List of Tables ................................................................................................................................................ vi

List of Figures ................................................................................................................................................ xi

List of Abbreviations ....................................................................................................................................... xii

Chapter 1 Introduction ................................................................................................................................... 1

  Research Context: Gaokao ............................................................................................................................. 3

  Research Purpose and Questions .................................................................................................................. 5

  Dissertation Overview ................................................................................................................................... 6

Chapter 2 Literature Review ........................................................................................................................... 7

  Current Research on Essay Rating and the Sociocultural Considerations .................................................. 7

  CHAT Theoretical Framework ...................................................................................................................... 16

    Mediation .................................................................................................................................................. 17

    Operation, Action, and Activity .................................................................................................................. 18

    Activity Systems ......................................................................................................................................... 19

  Applications of CHAT in Education Research ............................................................................................ 21

  Operationalizing CHAT to Reconceptualize Essay Rating ......................................................................... 23

  Reinterpreting Previous Research on Essay Rating from a CHAT Perspective .......................................... 26

    Different Communities, Different Activities ............................................................................................. 26

    Same Community, Different Activities ....................................................................................................... 31

    Systemic Contradictions and Tensions ....................................................................................................... 34

  Summary .................................................................................................................................................... 49
Investigating Essay Rating through the CHAT Approach ........................................................ 49
Research Design Considerations .......................................................................................... 50
Data Collection Methods .................................................................................................. 52
Data Analysis and Integration Strategies ........................................................................... 61
Chapter 3 Method .............................................................................................................. 63
Study Settings .................................................................................................................... 63
Rating Centre ..................................................................................................................... 63
Schools ............................................................................................................................... 65
Qualitative Case Study Approach ....................................................................................... 67
Role of the Researcher ........................................................................................................ 69
Overall Research Design .................................................................................................... 71
Research Questions ............................................................................................................ 71
Research Design ................................................................................................................ 74
Participants ......................................................................................................................... 75
Directors .............................................................................................................................. 77
Team Leaders ..................................................................................................................... 77
Raters .................................................................................................................................. 78
Principals ............................................................................................................................. 78
Teaching Colleagues .......................................................................................................... 79
Materials and Instruments .................................................................................................. 79
NMET Essay Writing Tasks ............................................................................................... 79
The NMET Rating Scale ..................................................................................................... 80
NMET Sample Essays ....................................................................................................... 80
Think-Aloud Training Guide .............................................................................................. 81
Stimulated Recall Interview Guide ..................................................................................... 82
List of Tables

Table 1. Essay Rating Approaches: Cognitive and CHAT .......................................................... 23
Table 2. Summary of Research Design .......................................................................................... 74
Table 3. Participants’ Involvement in Relation to Investigation of Activity/Practices in the Two Settings ........................................................................................................... 76
Table 4. Information about Participating Essay Raters ................................................................. 78
Table 5. Data Collection Timeline ............................................................................................... 85
Table 6. Summary of Collected Documents ............................................................................... 89
Table 7. Search Results of “Gaokao Rating” News Reports ....................................................... 92
Table 8. Occurrence Counts of Aspects Raters Attended to ......................................................... 105
Table 9. Translated Example of Think-Aloud Protocol Extract to Illustrate the Coding of Rating Sequence .................................................................................................................. 109
Table 10. Tensions and Raters’ Corresponding Solutions ............................................................ 155
Table 11. Characteristics of Surface and Deep Structures in the NMET ERA ............................. 165
List of Figures

Figure 1. Vygotsky’s model of mediated action (adapted from Engeström, 2001) .......................... 17
Figure 2. Leont’ev’s three levels of activity (adapted from Engeström, 1999; Lantolf, 2006) .... 18
Figure 3. Activity system (adapted from Engeström, 1999, p. 31) ........................................ 20
Figure 4. Activity system of the experienced teachers in Cumming (1990) ................................ 27
Figure 5. Activity system of the novice teachers’ in Cumming (1990) .................................. 29
Figure 6. Activity system in Vaughan (1991) ......................................................................... 32
Figure 7. Activity system in Lumley (2005) ........................................................................... 35
Figure 8. Activity system before training in Weigle (1994, 1998) ............................................. 38
Figure 9. Activity system during training in Weigle (1994, 1998) ........................................... 40
Figure 10. Activity system after training in Weigle (1994, 1998) ............................................ 42
Figure 11. Activity system in operational setting in Baker (2010) ............................................. 45
Figure 12. Activity system in research setting in Baker (2010) ................................................ 46
Figure 13. Coding scheme for TAPs and SRs ....................................................................... 94
Figure 14. Rater’s general rating sequence .......................................................................... 108
Figure 15. Central NMET ERA system at rating centre ......................................................... 149
Figure 16. Relationship between the two activity-related settings (high schools and rating centre) ........................................................................................................................................................................ 162
List of Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AWE</td>
<td>Automated writing evaluation</td>
</tr>
<tr>
<td>CCM</td>
<td>Constant comparison method</td>
</tr>
<tr>
<td>CHAT</td>
<td>Cultural-historical activity theory</td>
</tr>
<tr>
<td>ERA</td>
<td>Essay rating activity</td>
</tr>
<tr>
<td>Gaokao</td>
<td>National University Entrance Examination in China</td>
</tr>
<tr>
<td>EFL</td>
<td>English as a Foreign Language</td>
</tr>
<tr>
<td>ESL</td>
<td>English as a Second Language</td>
</tr>
<tr>
<td>L1</td>
<td>First language</td>
</tr>
<tr>
<td>L2</td>
<td>Second language</td>
</tr>
<tr>
<td>NMET</td>
<td>National Matriculation English Test (the English component of Gaokao)</td>
</tr>
<tr>
<td>OSM</td>
<td>On-screen marking (computerized marking system used in NMET essay rating)</td>
</tr>
<tr>
<td>RQ</td>
<td>Research question</td>
</tr>
<tr>
<td>SR</td>
<td>Stimulated recall</td>
</tr>
<tr>
<td>TAP</td>
<td>Think-aloud protocol</td>
</tr>
<tr>
<td>TESL</td>
<td>Teaching English as a Second Language</td>
</tr>
</tbody>
</table>
Chapter 1 Introduction

Researchers have traditionally considered human essay rating as a cognitive process of information processing (e.g., Bejar, 2012; Freedman & Calfee, 1983) or of problem-solving (e.g., Crisp, 2010; DeRemer, 1998). As a result, essay rating research has focused on how individual raters’ cognitive rating processes and rating results (i.e., scores) are influenced by rater characteristics, artifact features (e.g., writing task, essay, rating scale), and interactions between them (e.g., Barkaoui, 2011a; Elliott, 2013; H. Li & He, 2015; Milanovic, Saville, & Shen, 1996; Weigle, 1994). In this line of research, researchers have conducted extensive studies on the writing aspects raters attend to when assessing essays under different conditions (e.g., Broad, 2003; Brown, 1991; Erdosy, 2005; Hamp-Lyons & Zhang, 2001; Rinnert & Kobayashi, 2001; Shi, 2001; Vaughan, 1991; Wolfe, Kao, & Ranney, 1998). Different models and frameworks have attempted to uncover the rating processes in terms of how raters respond to key artifacts and the rating sequences raters follow to form scoring decisions when assessing essays (e.g., Crisp, 2010; Cumming, Kantor, & Powers, 2002; Homburg, 1984; Lumley, 2005; Milanovic et al., 1996; Sanderson, 2001). While these studies have provided rich information about how individual raters interact with key artifacts when assessing essays, they often assume a decontextualized view that does not address the interactions between raters and the sociocultural contexts where the essay rating takes place. Consequently, their findings fail to fully explain why rating studies conducted in different sociocultural contexts often yield conflicting results about the heeded writing aspects and the rating processes raters go through. These conflicting results suggest that raters not only interact with those key artifacts, but may also interact with the sociocultural environment around them when rating essays. Therefore, essay rating cannot be seen as a solely cognitive process, but rather a socially situated practice with socially constructed
meanings, motives, and consequences (Barkaoui, 2008). A situated understanding of essay rating may yield more meaningful and contextualized interpretations of essay rating in real-life settings, and have important implications for the understanding of scoring reliability and improvement in essay rating practices. In recent years, some researchers (e.g., Barkaoui, 2008; Weigle, 2002) have called for attention to the social aspect of essay rating. Many studies have started to investigate influences of some social factors (e.g., test stakes, rater training, institutional requirements) on rater decision-making (e.g., Baker, 2010; Elder, Knoch, Barkhuizen, & von Randow, 2005; Furneaux & Rignall, 2007; Lumley, 2005; Weigle, 1998). Those studies, however, are confined by the cognitive approach they follow, and hence cannot fully address rater-context interactions, reiterating the need for a comprehensive, situated understanding of essay rating. Therefore, a study on essay rating that situates this activity within its sociocultural context can be a valuable contribution to the literature.

My study drew on cultural-historical activity theory (CHAT; Engeström, 1987, 2001) from sociocultural theory to reconceptualize essay rating as a socially mediated activity and to explore the phenomenon within the context of a high-stakes English test in China. CHAT takes a dialectic view about the relationship between cognition and sociocultural contexts. In this case, all elements external to individual raters are considered part of the sociocultural context. Moreover, key artifacts are no longer viewed as verbal texts as by the cognitive approach, but rather, cultural artifacts with cultural and historical denotations. From a CHAT perspective, essay rating is viewed as a complex activity with cognitive and social layers. The cognitive layer is better understood when it is related to the social layer. To wit, the cognitive layer is individual raters’ short-lived, goal-directed actions, where raters’ decision-making is mediated by different mediators (e.g., tools, rules, community, division of labour) to achieve their immediate goals or
underlying purposes of rating. Meanwhile, the social layer is *raters' durable, collective object-oriented activity*, where essay rating is a summary of raters’ multiple actions driven by a deeply communal motive rooted in the community of which the raters are members. Hence the social layer involves an activity system—consisting of a central rating activity and other interrelated practices—that enriches the meaning of the cognitive layer findings.

From the CHAT perspective, the cognitive approach focuses on individual raters only, irrespective of raters’ goals of rating or assuming raters hold a shared goal (e.g., problem-solving). Such a view tends to ignore the fact that in real-life settings raters may have their own goals when assessing essays, and that their goals may be different than what they are expected to achieve. These goals may subsequently have an impact on their essay rating practices. Therefore, situating individual raters’ decision-making actions within a wider sociocultural context will allow more meaningful interpretations of essay rating.

**Research Context: Gaokao**

Guided by the CHAT framework, this study examined essay rating activity (ERA) within the context of the English component of a high-stakes university entrance examination in China, the National Matriculation English Test (NMET). The NMET is situated in the Chinese testing-driven society (Cheng & Curtis, 2010) and is a national test with one of the highest stakes. China has the longest testing history in the world, originating with the Imperial Examinations during the Han Dynasty (206 BC-220 AD), used by emperors to select government officials (Cheng, 2008). Testing has since played an important role in Chinese social and educational life, enjoying a wide social acceptance and recognition as a fair measurement for selecting talented individuals into the social hierarchy (Cheng, 2008). Such influence remains prominent today, as exemplified by the number of tests Chinese students write throughout their education and at different stages
of their life. Tests are administered to kindergarten students, and to students seeking admission to postgraduate programs. Students take monthly school-based tests, as well as annual nationwide tests, with purposes ranging from determining proficiency to professional certification. Among all tests in China, the national University Entrance Examination (UEE, also known as Gaokao) bears excessively high stakes, as university admission decisions are made based solely on students’ Gaokao results. Gaokao includes three mandatory components (English, Chinese, mathematics) and two optional components (sciences and social sciences). With over 9 million students taking Gaokao every June, Gaokao exerts a huge impact on a large number of stakeholders (e.g., students, teachers, parents, high schools, universities), bringing strong influences on teaching and learning (i.e., washback effect) in secondary education, thereby driving education reforms (Cheng & L. Qi, 2006; Gu, 2013; L. Qi, 2005).

There are a limited number of empirical studies on Gaokao relative to available discussion papers on Gaokao-related topics. Discussion papers have usually provided thoughts and perspectives on topics such as Gaokao policy and education equity issues (H. F. Liu, 2010; Zheng, 2010), the history and impact of Gaokao (J. Liu, 2005; X. Liu, 2011), or comparisons between Gaokao and the Imperial Examinations (X. Liu & H. F. Liu, 2012). In contrast, empirical studies have focused on test item analysis, construct validation, and test fairness (Mei & Cheng, 2014; Wang & Zhan, 2011; Zeng, 2010); influential factors on test-takers’ performance (H. J. Liu & Guo, 2003); and test influences on teaching and learning (X. Huang, 2011; L. Qi, 2005, 2010; Zheng & Chen, 2013).

As mentioned above, Gaokao has three mandatory examination components. The English component (NMET) consists of four sections: listening, language use, reading comprehension, and writing. The writing section has two subsections, error correction and essay writing. The
essay writing subsection accounts for one-sixth of the total NMET score (25 out of 150 points). In the writing task, students are given a prompt (in Chinese) and asked to respond with a purposive letter of approximately 100 words. High school English teachers\(^1\) are usually hired and work centrally at a provincial rating centre after the NMET administration to rate essays by all provincial test-takers. Due to the high-stakes impact of Gaokao and the role of English teachers as NMET raters, it is essential to examine these raters’ NMET ERA as socially mediated.

Previous research on essay rating in Chinese contexts has also followed the cognitive approach (e.g., Gao, 2011; Hamp-Lyons & Zhang, 2001; H. Li & He, 2015; W. Zhang, 1999). Some recent studies have found that immediate rating and teaching communities may exert influences on raters’ decision-making in large-scale high-stakes language tests, such as the College English Test (CET) and the NMET (e.g., Mei & Cheng, 2014; J. Zhang, 2009). Therefore, to obtain a more comprehensive understanding of rating essays in high-stakes tests, studies ought to consider the interaction between raters and sociocultural contexts rather than examining raters’ decision-making as an isolated event. This sociocultural approach that my study employed is especially important within the context of the Chinese NMET.

**Research Purpose and Questions**

The purpose of this study was to understand essay rating activity (ERA) at one provincial rating centre in China as socially mediated by the sociocultural context of the NMET. Considering the high impact and strong washback of the NMET and the dual roles played by high school teacher raters, the sociocultural context of the NMET ERA in this study included the broader (Chinese society) and immediate (rating centre and school) sociocultural contexts. Three research questions were addressed:

\(^1\) In some provinces university English teachers may also be involved.
1. How do raters assess NMET essays to achieve their goals?

2. What are the broader (Chinese society) and immediate (rating centre and school) sociocultural contexts in which the NMET ERA is situated?

3. What is the nature of NMET ERA as an activity system within the above sociocultural contexts?

RQ1 addresses the cognitive layer of NMET essay raters’ decision-making actions; RQ2 addresses the social layer of these raters’ collective ERA in relation to the large Chinese context and the two activity-related settings, the provincial rating centre and the raters’ high school workplaces. RQ3 combines the cognitive and social layers for a more situated, comprehensive understanding of the NMET ERA.

**Dissertation Overview**

This dissertation is organized in five chapters. The current chapter introduced the research context, and my research purpose and questions. The following chapter reviews relevant literature that establishes the theoretical foundations for this study. These foundations include limitations of previous research on essay rating, the guiding CHAT framework, and the application of CHAT to research on essay rating in comparison to the cognitive approach in terms of their conceptual views, research finding interpretations, and methodological considerations. Chapter 3 introduces the study settings, explains the research approach and design involved in this study, and describes the participants, instruments, and procedures involved in data collection, analysis, and integration. Chapter 4 presents research results associated with raters’ goal-directed decision-making actions (cognitive layer) and raters’ object-oriented NMET ERA (social layer). The final chapter presents a discussion of my research findings and implications of this study.
Chapter 2 Literature Review

In this chapter, I review the literature relevant to essay rating research, with the purpose of articulating the significance of integrating sociocultural considerations into essay rating research. I identify the limitations of previous research and subsequently propose theoretical and methodological solutions in this study, namely the use of cultural-historical activity theory (CHAT) as a theoretical framework to guide sociocultural considerations into essay rating research.

This chapter introduces the key concepts and evolution of activity theory, with a brief review of educational scholarship that employs CHAT in research. I then apply CHAT to reconceptualize essay rating and compare this approach with the cognitive approach in terms of their conceptual views. To further illustrate differences between the two approaches and potential values of CHAT, I reinterpret examples from some previous studies on essay rating through CHAT analysis to illustrate how CHAT might enable a better understanding of their findings. Given the theoretical value of adopting CHAT, I explain the considerations and qualifications of investigating essay rating through the CHAT approach vis-à-vis the methodological implications and limitations of the cognitive approach.

Current Research on Essay Rating and the Sociocultural Considerations

Direct writing assessment—usually in the form of timed impromptu essay tests—is widely used in large-scale assessment. Despite the growing prominence of automated scoring in recent years most essay tests are still scored by human raters (Xi, 2010). Following Hamp-Lyons’s (1991) characterization of a direct writing test, Weigle (2002) summarized that essay tests have seven key characteristics. These characteristics are that (a) test-takers are given instructions (i.e., the writing task) and allowed freedom in their responses; (b) the topic is
unknown to test-takers in advance; (c) test-takers have to write at least one piece of continuous text; (d) the text must be written in a preestablished time frame; (e) each text is scored by at least one (typically two or more) trained rater; (f) judgments are based on a common benchmark (e.g., rating scale); and (g) judgments are expressed as numbers, although they may be complemented with verbal descriptions. These characteristics spell out five key elements in an essay test—the test-taker, the rater, the writing task, the produced essay, and the rating scale.

Literature on essay rating has focused on interactions among four of the key elements at the rating stage, namely, how raters respond to writing tasks, essays, and rating scales, and how these interactions affect raters’ decision-making processes and results (e.g., Barkaoui, 2010b; Crisp, 2008; Cumming, 1990; Pula & Huot, 1993; Shi, 2001; Vaughan, 1991b; Weigle, 1994; Wolfe et al., 1998). These studies have examined how various characteristics of the four key elements influence rater decision-making, including essay features raters attend to, sequential procedures they follow, and essay scores they assign.

A variety of essay features (e.g., vocabulary, content, organization, coherence, topic development, errors, length, spelling, writing mode) influence rater decision-making processes (e.g., Broad, 2003; Cumming, 1990; Diederich, French, & Carlton, 1961; Gebril & Plakans, 2014; H. Li & He, 2015; Vaughan, 1991). For example, Diederich, French, and Carlton (1961) found that raters made their evaluation decisions on university student essays based on five general categories of essay features: ideas (e.g., relevance, clarity, development), form (e.g., organization), flavour (e.g., style, interest), mechanics (e.g., errors in grammar and punctuation), and wording (e.g., choice and arrangement of words). In a study conducted with raters assessing student essays at one university, Vaughan (1991) found raters attended to organization, content, grammar, sentence structure, coherence, and handwriting. Broad (2003) also noticed that raters
paid attention to a wide range of textual features (e.g., errors, grammar, punctuation, spelling, legibility, capitalization, content, paragraphing, transitions) when assessing university student essay quality. These studies demonstrated that raters focus on different essay features during decision-making.

Some studies have further explored the relationship between different essay features raters attend to and characteristics of the four key elements (e.g., Barkaoui, 2010a, 2011a; Brown, 1991; Cumming, 1990; Cumming et al., 2005, 2002; Delaruelle, 1997; Shi, 2001).

Researchers have often examined rater characteristics (e.g., rating and teaching experience, language background, professional background) as one variable in such a relationship, but sometimes obtained inconsistent findings. For example, when comparing the influence of rater experience on their rating focuses, Cumming (1990) observed that novice raters tended to rate language use differently than content and rhetoric, while experienced raters showed no significant differences between ratings for those three aspects of writing. However, Barkaoui (2010a) found that novice raters tended to give more importance to argumentation, while experienced raters tended to give more importance to linguistic accuracy. Raters’ language backgrounds are also found to account for their different rating focuses, but findings are mixed in terms of their specific focuses. For example, Shi (2001) compared scoring reasons reported by native English-speaking EFL (English as a Foreign Language) teachers and L1 (first language) Chinese EFL teachers. Her results showed that their rating focuses were slightly different; native-speaker teachers attended more positively to content and language, whereas Chinese teachers attended more negatively to the organization and length of the essays. Moreover, native-speaker teachers focused more on language in their third criteria, whereas Chinese teachers were also more concerned with content and organization in their primary criteria. Connor-Linton (1995)
compared evaluation criteria adopted by American ESL (English as a Second Language) teachers and Japanese EFL teachers when they assessed L1 Japanese EFL student essays. He found, unlike Shi’s findings, that Japanese teachers focused more on matters of accuracy (content, vocabulary, and grammar), while American teachers focused more on both intersentential features of the discourse (e.g., development, organization, cohesion) and specific intrasentential grammatical features (e.g., grammar, vocabulary). Raters’ professional backgrounds are another variable used to explain variability in raters’ focuses. For example, Brown (1991) found when assessing university student essays, English faculty members paid more attention to cohesion and syntax than ESL faculty members as positive essay features; while ESL faculty members considered organization more important. In terms of negative features, both groups of raters seemed to attend to syntax, with mechanics attracting more attention from English faculty and content attracting more attention from ESL faculty. However, when Cumming et al. (2002) compared rating focuses between raters with English L1 and L2 (second language) work experience, they found raters with L2 work experience tended to focus more on language-related issues while raters with L1 work experience tended to focus more on ideas or content. Those studies indicate that the relationship between rater characteristics and their rating focuses is complex and inconclusive.

Variation in writing task characteristics (e.g., genre, input materials, task choice, wording, discourse mode) may also lead raters to focus on different essay features. For example, Delaruelle (1997) found when teachers assessed different genres of writing (interpersonal vs. persuasive) some of the most salient criteria they used for both genres were almost identical—organization, grammar and task completion; while appropriacy and register were more salient for assessing interpersonal writing in contrast with cohesion more salient for assessing persuasive
writing. Cumming et al. (2002) compared raters’ focuses when rating essays elicited from two types of writing tasks (integrated or independent) using different input materials. Raters paid more attention to relevance, originality, and task completion on integrated writing task essays (based on listening and reading materials), but made more comments on language when reading independent writing task essays (based on a given topic).

Rating scale characteristics (e.g., holistic, analytic) were also associated with variability in raters’ focuses, sometimes with inconclusive findings. For example, Barkaoui (2010b, 2011a) compared essay features raters attended to when using holistic and analytic rating scales. He found raters gave more attention to task completion, content, and specific language features (e.g., lexis, error frequency, syntax, spelling) when using the holistic scale, while they gave more attention to text organization, linguistic appropriacy, and language overall using the analytic scale. In contrast to Barkaoui’s findings, H. Li and He (2015) found that raters paid more attention to general quality of language use and non-scale-related language features when using the holistic scale while giving more attention to coherence and grammar when using the analytic scale. These findings suggest that the relationship between rating scale characteristics and raters’ focuses may be rather complex.

Essay characteristics (e.g., proficiency level, writing mode) may lead raters to focus on different essay features as well. Gebril and Plakans (2014) compared raters’ focuses when assessing essays at different proficiency levels elicited from integrated writing tasks. They found that raters attended to specific language features (e.g., spelling, vocabulary, sentence structure) and source citation mechanics (e.g., use of quotation marks and attribution) more frequently when assessing low-proficiency essays. Raters shifted their focuses at higher essay proficiency levels, attending more to topic development, organization, and quality of source integration.
(effectiveness, relevance, and accuracy). However, Cumming et al. (2002) found that when rating essays at different proficiency levels, raters attended more to language features when rating low-proficiency essays but to both rhetoric and language when reading high-proficiency essays. Similar to rating scale characteristics, findings suggest a far from simple relationship between essay characteristics and raters’ focuses.

Given these mixed findings regarding essay features raters attend to, it is not surprising when findings concerning the relationship between essay features and scores are inconclusive (See Huot, 1990, and Barkaoui, 2007a, for detailed reviews of rating L1 and L2 essays, respectively).

Various models and frameworks have attempted to demystify complex interactions between raters and artifacts (writing tasks, rating scales, essays), including Crisp (2010, 2012), Sanderson (2001), and Wolfe (1997) in L1 essay rating; and Cumming et al. (2002), Lumley (2005), and Milanovic et al. (1996) in L2 essay rating. They describe in different contexts the sequence of steps that raters follow and/or the elements raters attend to when making judgments, again often yielding inconsistent findings. Raters tended to employ different approaches or styles (DeRemer, 1998; Lukmani, 1996; Sakyi, 2000, 2003; Vaughan, 1991; Wolfe et al., 1998); or may go through a process that is sequential (Crisp, 2012; Freedman & Calfee, 1983; Homburg, 1984; Milanovic et al., 1996) or complex and interactive (Cumming et al., 2002; Lumley, 2005; Sanderson, 2001).

These empirical studies and conceptual frameworks have undoubtedly enriched the understanding of rater decision-making, with considerable cognitive details that illustrate the complexity of this human activity. It is, however, difficult to determine what influences have contributed to those mixed findings. One possible reason for these inconclusive relationships, as
Barkaoui (2007a) assumed, is the variability in different studies. Such variability can exist in rating task requirements across studies, different ways textual features are defined and measured, non-linearity of relationship between essay features and scores, complex interactions between the key elements, and the possibility that variables associated with the four key elements (e.g., rater background, task type, scale type, essay proficiency level) may mediate the relationship between decision-making processes and results. Additionally, these studies tend to overlook the fact that they were conducted in different sociocultural contexts and the human-context relationship may have mediated rater decision-making processes and results. Recent research (e.g., Lumley, 2005) shows that well-trained, experienced raters show great variation in their behaviours but nonetheless achieve high interrater reliability. Essay rating, then, is a process filled with tensions between institutional requirements and raters’ intuition, which are resolved by an indeterminate process. In this regard, rater decision-making remains (at least partially) a black box as to how and why raters deal with such tensions. The need to understand essay rating continues, and the key lies in understanding interaction between raters and sociocultural contexts.

Researchers (e.g., Barkaoui, 2008; Weigle, 2002) have recently acknowledged the importance of understanding the interplay between raters and sociocultural contexts on rater decision-making, and called for more research on essay rating taking this into consideration. As Barkaoui pointed out, “essay rating is a socially situated process with a social meaning and social consequences” (p. 231); situating essay rating in broader sociocultural contexts can make research findings more understandable. As human activity does not happen in a vacuum, raters’ interaction with sociocultural contexts is inseparable from their ERA, thus potentially contributing to rater variability.

Weigle (2002) first alluded to the influence of sociocultural contexts on essay rating,
which she called the “social aspect” of essay scoring” (p. 74). She explained the sociocultural context was “the particular social milieu in which the assessment takes place (i.e., the school or institutional setting), which determines the goals of the assessment and the broader social and cultural context and relates to cultural norms about writing assessment” (p. 60). More specifically, Barkaoui (2008) explained the context as a sociocultural, institutional, and political environment where essay rating occurs, which “specifies the criteria, purposes, and possibly processes, of reading, interpreting, judging, and applying the three texts to arrive at a rating decision” (p. 3). Thus sociocultural contexts are viewed independently of essay rating in Weigle’s and Barkaoui’s definitions, outlined as the sociocultural background against which the essay rating takes place and dictates the rules of carrying out the activity.

A number of studies have started to analyze sociocultural influences on essay rating (e.g., Baker, 2010; Furneaux & Rignall, 2007; Lumley, 2005; Smith, 2000; Weigle, 1994, 1998). These studies attempt, explicitly or implicitly, to elicit some factors (e.g., test stakes, rater training, institutional constraints) that can be isolated from the sociocultural context occurring “across a wide range of language assessment environments” (Baker, 2010, p. 135). Studying the effects of the stakes of a test, Baker found that not all raters were influenced by the high stakes of a provincial teacher certification test compared to low-stakes research conditions.

Rater training is a process in which raters are introduced to the evaluation criteria and then asked to rate some essay samples according to these criteria in an attempt to arrive at a shared interpretation of their meaning (Elder, Barkhuizen, Knoch, & von Randow, 2007). Rather than exploring the socializing processes of rater training, existing studies have been more concerned about the effectiveness of training, or norming, in standardizing rater behaviours (e.g., Elder, Knoch, Barkhuizen, & von Randow, 2005; Furneaux & Rignall, 2007; Lu, 2010; Weigle,
For example, Weigle (1994) found training helped raters to clarify the intended scoring criteria, modify their expectations of writing quality, and provided a reference group to allow raters to compare scores. Furneaux and Rignall (2007), however, found that while raters demonstrated increased attention to intended evaluation criteria after training, some criteria were given markedly less attention than others, indicating that training may not be as effective as expected.

In comparison to studies that examine one specific sociocultural factor in each study, Lumley’s (2005) study viewed sociocultural contexts more broadly, observing that socially motivated components were involved in the rating process, whereby a range of institutional constraints (e.g., rating scale, training, rater experience, professionalism) were in play to achieve the institutional goal. Lumley further explained that rating scales provided the most explicit statement of the institution’s intended construct, which was supplemented by rater training. Training allowed the test institution to elaborate on how the rating scale ought to be interpreted and applied. In addition, rater experience and professionalism were both essential qualities for the institution to select reliable raters who worked conscientiously to meet its requirements. Lumley’s work has drawn much attention to the social aspect of essay rating.

In existing studies, essay rating has been depicted as a process of raters processing information (e.g., essays, rating scales, writing task; Bejar, 2012) or trying to solve problems with a shared goal of making decisions about the quality of essays based on a predetermined set of evaluation criteria (DeRemer, 1998). These studies have provided rich information about the cognitive aspect of essay rating; however, they often yield inconsistent findings, which implies a complex relationship between rater decision-making and subsequent essay scores. One possible reason, among others, is that most of these studies decontextualize their research findings and
ignore the fact that essay rating is a socially situated human activity. Although some studies have started to take sociocultural contexts into account, they have provided limited insight into the human-context interaction in essay rating because of the confined cognitive approach, and hence were unable to fully capture and explicate the socially mediated nature of essay rating. Three major limitations arise. First, the research focus remains on individual raters. Individuals’ essay rating practices need to be situated in broader sociocultural contexts in order to produce more meaningful interpretations. Second, sociocultural contexts are narrowly defined and reduced to a few factors, which is insufficient to encompass the complex, dynamic nature of sociocultural contexts. Third, these studies postulate that rater cognition is firmly located inside individual raters, ignoring the interactional, dynamic relationship between individuals and the outside world in human activities. It is evident that the cognitive approach to essay rating research is limited and inadequate to embrace and explain the richness and dynamics of human ERA. Hence I propose to use the cultural-historical activity theory (CHAT; Engeström, 1987, 2001) to reconceptualize and understand essay rating within its sociocultural context.

**CHAT Theoretical Framework**

CHAT was initiated by the Soviet psychologist Lev Vygotsky (1978) in the late 1920s. Vygotsky lived in the era of the 1917 Soviet Revolution and his works were greatly influenced by Marxist philosophical principles (Wertsch, 1985). CHAT stems from Vygotsky’s cultural-historical psychology, often referred to as the sociocultural theory of mind, contending that cognition should be investigated without isolating it from its social context. Sociocultural theory aims to explain how human mental functioning is related to cultural, institutional, and historical contexts (Wertsch, 1998). It acknowledges the central role of social relationships and culturally constructed artifacts in organizing human cognition (Lantolf & Thorne, 2006). CHAT has so far
evolved through three generations of research (Engeström, 2001). I will introduce the fundamental concepts that characterize milestones in the development of this theory, with a focus on the first two generations applicable to this study.

**Mediation**

The concept of mediation is the key to understanding CHAT. According to Vygotsky, humans are not restricted to simple stimulus-response reflexes; they are able to make indirect connections between incoming stimulation and their responses through various links (Cole, 1976). Such indirect connections are mediation. Vygotsky’s triangular model of mediated action is acknowledged as the cornerstone of CHAT (Engeström, 2001). It is commonly expressed as the triad of subject, object, and mediating tools (see Figure 1). The subject (a person or a group of persons) and the object (an objective or goal) are interlinked through a tool (e.g., material or symbolic artifacts, social others) that mediates the action.

![Figure 1. Vygotsky’s model of mediated action (adapted from Engeström, 2001)](image)

This concept of mediation postulates a dialectical relationship between individuals and the society in which the latter empowers the former to regulate their minds from the outside social world and the inside mental world through the use of mediating sources (Cole & Engeström, 1993). Vygotsky’s idea of mediated action challenged his contemporary stimulus-response theorists by arguing that cognition and act, individuals and society are integrated, rather
than separated, which characterizes the first generation of CHAT (Engeström, 2001; Swain, Kinnear, & Steinman, 2010). The concept of mediated action revolutionarily breaks down the dualistic Cartesian walls which isolate individuals’ minds from the society; however, its unit of analysis remains focused on the individual and hence does not fully explicate the societal and collaborative nature of actions (Engeström, 1999).

**Operation, Action, and Activity**

To overcome the limitation of the first generation model, Leont’ev (1981) expanded the scope of Vygotsky’s mediated action by introducing a collective aspect through the concept of “activity”, which characterizes the second generation of CHAT (Engeström, 1987). He postulated activity as a three-level hierarchical process relating to corresponding levels of consciousness (Engeström, 1999; see Figure 2).

<table>
<thead>
<tr>
<th>Activity</th>
<th>→</th>
<th>Motive</th>
</tr>
</thead>
<tbody>
<tr>
<td>Action</td>
<td>→</td>
<td>Goal</td>
</tr>
<tr>
<td>Operation</td>
<td>→</td>
<td>Condition</td>
</tr>
</tbody>
</table>

*Figure 2. Leont’ev’s three levels of activity (adapted from Engeström, 1999; Lantolf, 2006)*

The bottom level is *operations*, the automated or unconscious routines of an individual when responding to conditions (i.e., physical or mental setting in which actions take place). The middle level is *actions*, the physical or cognitive acts an individual does to achieve desired results. Actions are carried out by means of operations and driven by goals. Under certain conditions the acts are carried out unconsciously—they are operations—while the same unconscious operations may become conscious actions if the conditions change. Different actions may be directed toward the same goal or the same action may be directed toward
different goals. The top level is *activity*, which include sets of various actions motivated by socially or culturally constructed objectives or purposes.

In his famous example of “primeval collective hunt”, Leont’ev (1981) illustrated the integration of the three levels and how division of labour brings about the crucial difference between an individual action and a collective activity. Collective hunting is an activity with a shared motive, to catch deer for consumption. For example, the responsibility of a beater is to frighten the deer in a specific direction, away from himself, by making a noise such as clapping hands. The clapping of the beater’s hands is an operation. The beating as a whole is an action, directed by the goal to frighten the deer. However, the goal of beating action contradicts the collective motive to hunt the deer. This paradoxical action can only be understood when considered in relation to others participating in the same hunt, whose action is to kill the deer through the use of weaponry: Their mutual motive is to catch the deer and distribute and consume food. The beating action is part of the collective hunting activity on the basis of the beater’s conscious knowledge that he frightens the deer and cooperates with others so that the socially constructed objective of catching the deer can be achieved (Axel, 1997).

This hierarchical or vertical structure extends the unit of analysis to object-oriented activity, bringing CHAT a huge step forward (Engeström, 2001). The three-level scheme directs our attention to explaining and understanding from a sociocultural perspective how mental and observable activity can be regarded as an integrated unit of analysis, and how interactions between the two affect both the individual and the environment (Yamagata-Lynch, 2010).

**Activity Systems**

Although Leont’ev developed Vygotsky’s mediated action into a vertically structured activity scheme, it was Engeström (1987, 1999) who accepted Leont’ev’s three-level scheme and
further graphically modeled Leont’ev’s activity system (see Figure 3). This visual presentation of the activity system was based on Vygotsky’s original model of mediation triangle, to make the context visible and its constituent elements explicit. The uppermost subtriangle of Figure 3 represents individual or group actions embedded in a collective activity system, and is Vygotsky’s mediation triangle, which is the “tip of the iceberg” (Engeström, 2001, p. 134). In addition to tools, Engeström (1987) added three mediating sources—rules, community, and division of labour—to the human activity system. As visualized in Figure 3, this activity system encompasses (a) individuals as subjects or actors who are performing goal- or object-oriented activities; (b) tools which include physical objects, signs, and social others that facilitate the activities; (c) rules that dictate how the tools are to be used and regulate how the subjects act to obtain the objects; (d) the community of which the subjects are members when engaging in the activities; (e) the division of labour that describes the continuously negotiated distribution of powers and responsibilities among members of the community, or roles taken by the subjects when participating in the activities; and (f) the outcomes are the results or consequences of an activity. Following Lei (2008), I define the four mediating sources—tools, rules, community, and division of labour—in an activity system as mediators.

![Figure 3. Activity system (adapted from Engeström, 1999, p. 31)](image)

It should be noted that although Figure 3 is often called the activity system, goal-directed
actions (the middle level) are often visualized using these extended triangles as well. Based on this conceptualization, Engeström (2001) further expanded the theory and included interacting activity systems to understand multiple perspectives and networks among interrelated activity systems, which is acknowledged as the third generation of CHAT. In an activity system, the four mediators are in constant interaction with each other, during which systemic contradictions emerge. These contradictions and nature of each mediator can create tensions within the system. The tensions can affect the interactions between mediators, which may either promote or hinder a subject’s ability to attain an object (Yamagata-Lynch, 2010). As Engeström (2001) pointed out, systemic contradictions and tensions are inherent in human activities and do not occur accidentally or arbitrarily.

Applications of CHAT in Education Research

Over three generations of development, CHAT has evolved into a well-established approach to examining human activities from a sociocultural perspective. It has been widely applied in research areas such as psychology, linguistics, human-computer interaction, cognitive science, anthropology, communication, workplace studies, and education (Lantolf & Thorne, 2006). In education, CHAT has been applied in various research fields such as formative assessment (e.g., Black & Wiliam, 2009; J. Li, 2016; Pryor & Crossouard, 2008), second language learning and assessment (e.g., Coughlan & Duff, 1994; Haneda, 2007; Lei, 2008; Swain et al., 2010), e-learning (e.g., Benson, Lawler, & Whitworth, 2008; Mwanza, 2002), and teacher education (e.g., Twiselton, 2004; Yamagata-Lynch, 2003).

Taking the application of CHAT in formative assessment research as an example, teachers’ formative assessment activity has been theorized as a cultural-historical activity system oriented by supporting and improving student learning, which involves teacher’s (assessor)
activity interacting with student’s (learner) activity in classroom assessment (Black & Wiliam, 2009; Pryor & Crossouard, 2008). This line of research often focuses on teachers and students, and their respective roles and power relations in the regulation of learning, providing and receiving feedback, and teachers’ emotion, cognition and action in the assessment activity (e.g., Crossouard, 2009; J. Li, 2016; Webb & Jones, 2009). For instance, J. Li (2016) examined how tutors interacted with their contexts in a New Zealand university when assessing undergraduates’ written assignments. She observed complex interactions among tutors’ emotion, cognition, and grading actions when they intended to provide formative assessment feedback to students’ work, including weighing and comparing students’ grades against the whole cohort, or consulting their supervisor when having to fail an assignment.

Those CHAT-based formative assessment studies are a related body of scholarship investigating interactions between the assessor/rater and the sociocultural context when making assessment decisions. From a CHAT perspective, however, a formative assessment activity and an ERA are different in nature. Formative assessment activity often takes place in the classroom where the assessor (teacher) directly interacts with the assessee (student) for the purpose of supporting student learning. It involves different dynamics within the assessment activity system than an ERA, where raters (oftentimes teachers) usually do not know the assessee (writer) and do not have any direct relationship with assessees, and the assessment purpose often involves summative use of essay scores (e.g., selection, certification, immigration). Therefore, a formative assessment activity and an ERA are related, yet different types of assessment activities. Few, if any, instances of CHAT-based essay rating research have been identified; however, findings from formative assessment research indicate that CHAT may bring promising insights into this cognitive-approach-dominated research field.
Operationalizing CHAT to Reconceptualize Essay Rating

As previously discussed, CHAT takes the view of human activity as a complex, socially mediated process rather than a linear, individually focused cognitive process. In this section, I compare how essay rating is reconceptualized through CHAT versus a cognitive approach in terms of their different conceptual views on essay rating, sociocultural contexts, the relationship between cognition and sociocultural contexts, and the unit of analysis in researching essay rating (see Table 1 for a summary).

Table 1

*Essay Rating Approaches: Cognitive and CHAT*

<table>
<thead>
<tr>
<th></th>
<th>Cognitive</th>
<th>CHAT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Views about essay rating</td>
<td>Linear, individually focused process of information processing or problem-solving</td>
<td>Complex, socially mediated activity, with individual raters’ actions (cognitive layer) interpreted against their collective ERA system (social layer)</td>
</tr>
<tr>
<td>Views about sociocultural contexts</td>
<td>Social milieu of essay rating that dictates the rules of carrying out essay rating, often in the form of independent, separable factors</td>
<td>Social world outside individual raters, including all the mediators and their interactions within the entire activity system</td>
</tr>
<tr>
<td>Relationship between cognition and sociocultural contexts</td>
<td>Dualistic, unidirectional, stable</td>
<td>Dialectic, interactional, dynamic</td>
</tr>
<tr>
<td>Unit of analysis</td>
<td>Individually focused actions or operations</td>
<td>Collective object-oriented activity</td>
</tr>
</tbody>
</table>

The cognitive approach of essay rating is a linear process of information processing where raters respond to conditions or stimuli (e.g., Baker, 2012; Bejar, 2012; Freedman & Calfee, 1983), or as a process of problem-solving of which raters’ goals are to make decisions about essay quality based on a given set of rating criteria (e.g., Crisp, 2010; DeRemer, 1998; Lumley, 2005). In contrast, the CHAT approach of essay rating is more complex, a socially
mediated activity with three hierarchical levels. Among the levels, the distinction between short-lived goal-directed action (middle level) and durable object-oriented activity (top level) is of central importance (Engeström, 2000); therefore, the key to understanding ERA lies in individual raters’ goal-directed actions and their collective object-oriented activity. When performing the ERA, individual raters’ goal-directed decision-making actions (cognitive layer), are more understandable when interpreted against the background of their collective object-oriented ERA system (social layer). This activity system consists of the central ERA system in relation to other interrelated practices, which is always coupled with systemic tensions brought about by interactions between different mediators within the activity system. From the CHAT perspective, the cognitive approach deals solely with the cognitive layer of ERA, either on individual raters’ automatic rating operations or on raters’ goal-directed decision-making actions, and assumes that goals and underlying purposes of essay rating are shared by all participating raters. Without situating those cognitive details within their sociocultural contexts, the cognitive approach may lose sight of the forest for the trees.

The cognitive approach defines sociocultural contexts as a social milieu against which essay rating takes place and that dictates the rules of conducting the activity, often researched in the form of a few independent, separable factors (e.g., test stakes, training) elicited from sociocultural contexts. In contrast, the CHAT approach views sociocultural contexts as the entire social world outside the individual rater, including all four mediators and their interactions within the activity system. From a CHAT perspective, the cognitive approach defines sociocultural contexts in a narrow sense, mostly counting only one mediator (i.e., rules) in the activity system as sociocultural contexts while ignoring the sociocultural roles of other mediators. Also, the sociocultural context being represented as independent, separable factors
oversimplifies the dynamic, complex social world in real-life settings.

The cognitive approach takes a dualistic view, isolating sociocultural contexts from cognition, the latter firmly located inside the individual. The influences of sociocultural contexts on essay rating are often deemed as unidirectional and stable (i.e., rater judgement is shaped by their responses to sociocultural factors). However, the CHAT approach takes a dialectic view, regarding sociocultural contexts as integral to cognition, the latter distributed in the individual and in the sociocultural context (Cole & Engeström, 1993). The influences of sociocultural contexts on essay rating are deemed interactional and dynamic (i.e., ERA is shaping and being shaped by the activity system, with mediators constantly interacting with each other). The CHAT approach challenges the linear causal model of decision-making centered in the mind of an individual rater by offering a model that distributes decision-making across the individual raters, the sociocultural contexts, and the varied interactions between them (Cole & Engeström, 1993). According to CHAT, the cognitive approach separating sociocultural contexts from cognition is inadequate for grasping the essential nature of human experience and behaviour, as it fails to fully explain why effects of the same characteristics of the four key elements often yield different or conflicting results when examined in different sociocultural contexts.

The cognitive approach takes the individually focused actions or operations as its unit of analysis and the key to understanding human cognitive functioning. The CHAT approach adds a collective, social layer to understanding essay rating. The durable, object-oriented activity makes individuals’ actions more understandable by bringing social meanings to the short-lived, goal-directed actions. The CHAT approach hence takes the collective object-oriented activity as its unit of analysis.
Reinterpreting Previous Research on Essay Rating from a CHAT Perspective

This section builds on the earlier conceptual comparison between CHAT and the cognitive approach by interpreting some previous studies afresh from a CHAT perspective. My intention is to showcase how CHAT, as a conceptual tool, may provide new understanding of essay rating. I will revisit five important studies I reviewed earlier through a CHAT lens to highlight how CHAT could offer a better, or different, understanding about essay rating processes and results. These five reinterpretations are organized under three topics: (a) different communities, different activities (e.g., Cumming, 1990); (b) same community, different activities (e.g., Vaughan, 1991); and (c) systemic contradictions and tensions. The third one includes three subtopics: inherent tensions within a system (e.g., Lumley, 2005), tensions between interrelated systems (e.g., Weigle, 1994, 1998), and contextual tensions (e.g., Baker, 2010).

Different Communities, Different Activities

One line of essay rating research is interested in finding out how raters are affected by rater characteristics, such as their teaching and rating experience, and linguistic and professional backgrounds (e.g., Cumming, 1990; Lim, 2011; Sakyi, 2003; Wiseman, 2012). Some of those studies overlook the fact that within the same study raters with different experience and backgrounds may have engaged in different communities with different objectives to achieve when approaching the seemingly same rating task, which could have contributed to rater variability. As an example, Cumming (1990) compared ratings and rating behaviours of experienced and novice teachers scoring essays for an ESL placement test in a Canadian university. These essays were written by adult students with different levels of ESL proficiency and writing expertise. Multivariate analyses showed that novice teachers tended to rate language use differently than content and rhetoric, while experienced raters showed no significant
differences between ratings for those three aspects. Cumming also noticed that many decision-making behaviours used by the two groups of raters varied significantly. Experienced teachers appeared to have a fuller mental representation of the assessment task and incorporated a larger number of criteria, self-control strategies, and knowledge sources to read and evaluate the essays. In contrast, novice teachers tended to rate essays with few skills and criteria, using skills that may have been derived from their general reading abilities or other knowledge (e.g., editing essays). In addition to Cumming’s conclusion that such differences were accounted for by the two groups’ professional experience, a CHAT analysis suggests that the two groups of teachers seemed to engage in different communities with different rating objectives, which may also account for their differing behaviours.

Figure 4. Activity system of the experienced teachers in Cumming (1990)

Figure 4 simulates the experienced teachers’ activity based on Cumming’s work. In this activity system, the subjects are the expert teacher-raters who had many years of ESL
composition teaching experience in the university’s ESL program. These teachers’ rating objective seemed to be placing ESL students, which is the intended purpose of the writing assessment. As Cumming observed, these teachers tended more frequently to envision the personal situation of the writer (usually as an ESL student with specified learning needs), to count the number of main ideas in each essay in order to assess the students’ total written output, to establish a general impression of the ESL students’ command of syntax in English, and to engage in more “classifying errors” behaviours, among other strategies and behaviours. These behaviours seemed to be associated with placement purposes. Tools that mediated these experienced teachers’ activity may include the writing assessment task, the rating criteria, students’ essays, their ESL teaching experience, and students’ learning needs, the knowledge of which the teachers could have gained from years of teaching experience and inferred from the essays. The rules guiding this activity could include marking holistically based on the rating criteria. In addition, since essays were used for ESL placement decisions, it is safe to include a rule such as the program’s requirement of placing ESL students based on their learning needs. This activity was situated in the community of ESL program colleagues and the program coordinator, who shared responsibilities within this ESL program as their division of labour. The outcome of this activity was, according to Cumming’s findings, that these experienced raters showed no significant differences between ratings for the three writing aspects (language use, content, and rhetoric), and that they tended to have a fuller mental representation of “the problem” when evaluating student essays. This CHAT analysis suggests that those experienced teachers seemed to have performed an ESL-student-placement activity, trying to elicit from the essays useful information about the ESL students’ potential learning needs, which may account for such an outcome.
Figure 5. Activity system of the novice teachers’ in Cumming (1990)

Figure 5 is a simulation of the novice teachers’ activity. In this activity system, the subjects are the preservice teachers recruited from TESL (Teaching English as a Second Language) classes in the university. These teachers had no prior teaching experience. Their object for participation seemed to be practicing rating ESL essays. Cumming mentioned that the preservice teachers were promised feedback on their rating performance from the researcher in return for their participation. When offered the option of rating the essays alone or in the company of the researcher, two novice teachers chose the latter, in hope of having feedback immediately after rating. Therefore, practicing for feedback seemed a plausible objective for these novice teachers. Some of the tools available to these novice teachers were different from those of expert teachers’. Without prior ESL teaching experience, they seemed to compensate by drawing on their reading abilities and editing skills, as Cumming posited. The rules guiding their activity could include the same rule of holistic marking, and perhaps the TESL teaching and
assessment conventions they learned from their TESL class. This activity was situated in the community of their TESL class where their division of labour may include improving TESL practices, and the researcher who promised to provide feedback on their rating performance. The outcome of this activity was, according to Cumming’s findings, that these novice teachers’ ratings for language use was significantly different than their ratings for content and rhetoric, and thus engaged in more “editing phrases” behaviours to facilitate their interpretation of the essays. All these findings indicate that these novice teachers might be trying to apply their TESL expertise to reading and rating the essays, which may account for why they looked at language use differently than the other two aspects, content and rhetoric, and involved more editing or correcting error behaviours. This may be associated with the expertise they acquired from their TESL class, which needs to be confirmed with additional information. This CHAT analysis indicates that these novice teachers seemed not to be making assessment decisions but rather performing a practicing-ESL-rating activity.

This CHAT analysis of Cumming’s (1990) work suggests that, apart from the varied teaching experience leading to the differences in raters’ rating behaviours and results, the two groups of raters from two communities seemed to aim for different goals when approaching the “same” rating task. Their different goals could have led them to interact with different mediators and consequently may have also contributed to the differences in their rating behaviours and results. From a CHAT perspective, the “same” rating activity seemed to be two different activities for the two groups, which should also be taken into account when explaining these differences. Similar CHAT analysis can be applied to reinterpreting other studies where participating raters came from different communities (e.g., Cumming et al., 2002; Sakyi, 2003).
Same Community, Different Activities

Sometimes, raters engage in what seems to be the same rating task in the same community but they may hold different objectives when approaching the task, which could also have led to different activities. For instance, Vaughan (1991) examined what went on in experienced raters’ minds when the raters made holistic scoring decisions on borderline essays from an essay test in an American university (the purpose of this essay test was not specified). Based on analysis of think-aloud protocols, she found that the raters demonstrated great variation in their focuses and rating styles, including “two-category”, “grammar-oriented”, “single-focus”, “laughing rater”, and “first-impression dominates” styles. From a cognitive perspective, such variation may indicate that the raters probably involved “idiosyncratic criteria” (Charney, 1984), that is, raters adopted their own rating criteria in addition to the prescribed ones. However, from a CHAT perspective these findings suggest that those raters may have had different objectives when engaging in the seemingly same task in the same community, which consequently may have led to variation in other components in the activity system and ultimately the varied nature of their rating activities.

Figure 6 is a combined simulation² of three potential activity systems based on Vaughan’s (1991) study. The subjects of activity (A) are (A) raters, who were the “two-category” and the “grammar-oriented” raters. These raters’ object(ive) seemed to be making placement decisions. They concentrated on only one or two categories based on the rating scale and seemed to have their personal criteria regarding which categories or writing aspects were most important for a pass decision (e.g., grammar, organization, content). This action could be because these raters considered the purpose of the rating task was to make placement decisions. To achieve this

---

² This combination of activity systems is to facilitate comparison and highlight the differences between raters’ potential activities.
Figure 6. Activity system in Vaughan (1991)

object, (A) raters used borderline student essays and the rating scale as their mediating tools. Their activity was guided by rules including the rating criteria used in the university system and possibly the raters’ personal criteria of good writing, that is, one or two categories in the rating scale were most useful to judge student writing ability. This activity was situated in the community of university system where (A) raters’ responsibility might include marking essays. I call activity (A) placement activity.

The subjects of activity (B) are (B) raters, the “single-focus” raters. (B) raters’ objective seemed to be making pass/fail decisions, as they kept looking for evidence to pass or fail the essays. They used the same tools as the other raters: borderline student essays and the rating scale. They shared the same rule with other raters, the rating criteria used in the university system. Additionally, another guiding rule might be their personally held minimum requirements for essays deserving a pass. This activity was performed in the same community of the university.
program where (B) raters’ responsibility might be the same as (A) raters’. I call activity (B) pass/fail activity.

(C) rater, the “laughing rater”, is the subject of activity (C). This rater’s objective seemed to be making judgement on the imagined writer. Although Vaughan reported (C) rater’s major focus was on content, the aspects of content this rater focused on appeared slightly different than (A) raters who seemed more concerned about rhetorical aspects of the essay content (e.g., “I think the content is really very thin”, p. 118). (C) rater appeared more concerned about ideational aspects of the content, making judgement on the nature of the presented ideas instead of solely on the writer’s writing ability. For example, Vaughan described that (C) rater got annoyed at one essay on taking attendance in college and concluded that “I don’t like the way it argues, anyway…” (p. 120). This rater chuckled through several essays and reacted strongly to the contained ideas on which he or she commented in another essay as “written by a rather immature” writer (p. 120). As Vaughan commented, this rater seemed to have established a psychological link with the imagined writer. Consequently, the rater seemed to be making judgement based on the imagined writers rather than their essays. (C) rater shared the same tools, community, and responsibility with other raters. Besides following the same university rating criteria, there seemed to be a different rule guiding this rater: judging essays based on ideational aspects of content. I call activity (C) judgement activity.

There was another type of rater, the “first-impression-dominates” raters whose object seemed to be making judgement by first impressions. However, there was inadequate information from Vaughan’s study to interpret this type of raters’ objectives when they conducted the essay rating task. They might have the same objective as any of the above three types of raters, or have different objectives and perform an activity differently than any of the other three
types of raters. Regardless of their different approaches or activities, the outcome for all those raters’ activities was the same, to complete rating borderline essays.

What the cognitive perspective explained as raters with different rating styles or idiosyncratic criteria in Vaughan’s (1991) study, when revisited from a CHAT perspective, it might have been a result of raters holding differing objectives when approaching the same rating task. The raters’ potentially different objectives could have led them to interact with different rules, and hence perform the seemingly “same” ERA in different ways. Similar CHAT analysis may be applied to other studies where raters sharing the same community approached the rating tasks differently (e.g., Erdosy, 2004; Smith, 2000).

**Systemic Contradictions and Tensions**

Most previous studies on essay rating have left the impression that it is an activity with few tensions. However, CHAT contends that constant interaction exists in activity systems and often generates systemic contradictions and tensions. Any human activity features these inherent, complex contradictions and tensions, which account for the richness and dynamics of human activities. Although systemic tensions were not salient in most previous studies, such as Cumming (1990) and Vaughan (1991), some studies implied presence of such tensions in rating activities.

**Inherent tensions within a system.** Most studies tended to crystallize essay rating as a linear process (e.g., Crisp, 2010; Freedman & Calfee, 1983; Homburg, 1984), while some recent studies show that it is indeed a complex process with tensions (e.g., Elliott, 2013; Lumley, 2005; Sakyi, 2003). For example, Lumley explored the nature of the rating process by examining four trained, reliable raters marking essays from an essay test used for immigration purposes in Australia. He observed that raters followed a fundamentally similar rating process, but the
relationship between scale descriptors and essay quality remained obscure. He explained that the rules and the rating scale did not cover all eventualities, forcing the raters to develop various strategies to help them deal with problematic aspects of the rating process. In so doing, raters tried to remain close to the scale, but at the same time were heavily influenced by their first impressions of the essays. This established a tension between the rules and the intuitive impression. As Hamp-Lyons (2007) commented, Lumley’s work has greatly contributed to our understanding of essay rating, but at the same time has complicated the situation rather than solved the problem. Delimited by the cognitive view, the findings could not explain how the raters were able to resolve this tension, which ended up being an indeterminate process.

![Activity system in Lumley (2005)](image)

*Figure 7. Activity system in Lumley (2005)*

I reexamine Lumley’s (2005) study from a CHAT perspective (see Figure 7 for a simulated representation of the activity system). The subjects are the four trained experienced raters, whose goal was reading and rating to meet institutional requirements. The tools used by
these raters could include resources provided by the testing institution (e.g., essays, writing task, rating scale, rater training and reorientation), raters’ experience and professionalism, previously scored essays, and other undefined tools that raters used to assist with decision-making. The rules that guided these raters may include, according to Lumley, making judgement based on institutional rules embodied in the rating scale, rater training and reorientation, and maintaining high reliability. This activity took place in the community of the testing institution, where these raters’ responsibilities could include marking by following the institutional rules, and coping with difficulties that the institutional rules failed to cover.

As Figure 7 shows, the conditions of this activity could have brought forth three systemic tensions (in the shape of lightning-shaped arrows). A tension seems to exist between the tools and the object. The tools provided by the testing institution (rating scale descriptors, training, and rating guidelines) were not able to cover all the eventualities that raters encountered in the essays, and seemed inadequate to help the raters to achieve their object of meeting the institutional rules (e.g. following the rating scale). This tension could have triggered another tension between the rules and the object, making the raters struggle, as the rules required that raters must make a judgement based on the institutional rules and maintain high reliability. These two tensions could have brought about a subsequent third tension, between division of labour and the object. Raters may have been drawn by an attempt to meet the institutional requirements of reliable marking while accepting the fact that the tools provided were inadequate and had to individually cope with difficulties that institutional rules failed to cover. To solve these tensions raters may have developed strategies by employing various tools, such as previously marked essays (for comparison), the rating scale (to arbitrate between conflicting components of descriptors), their teaching experience and professional judgement. If still unsuccessful, they
might have even discarded the rating scale and adopted other undefined tools. Whatever strategies they used could have been driven by their ultimate goal of meeting the institutional requirements for rater reliability. The outcome of this activity was that raters managed to assign reliable scores.

This CHAT analysis of Lumley (2005) reveals that what he called the tension between the rules and the intuitive impression, plus the indeterminacy of how raters solved the tension, could have been a result of three interrelated tensions among four components of the activity system when the raters were striving for their object. My revisit of Lumley’s work provides a more insightful understanding of the nature of essay rating. Similar CHAT analyses can be applied to other studies describing contradictions and tensions within an activity system (e.g., Elliot, 2013; Furneaux & Rignall, 2007; Sakyi, 2003).

**Tensions between interrelated systems.** As previously discussed, the cognitive approach often ignores contexts and connections between ERA with other neighbouring activities. According to CHAT, essay rating may not be a standalone activity and can influence and be influenced by interrelated activity systems. For example, Weigle (1994, 1998) examined the effect of training on novice and experienced raters’ ratings and behaviours when marking an ESL placement examination (ESLPE) in an American university by comparing pre- and posttraining scores and decision-making behaviours. Her multi-faceted Rasch measurement analysis showed that before training the novice raters tended to be more severe and less consistent in their ratings than the experienced raters, but the difference was less pronounced after training. Furthermore, after training both groups assigned more consistent scores and novice raters also assigned less severe scores. However, significant differences in severity were still found among raters in both groups. Weigle concluded that rater training was more successful in improving raters’ self-
consistency than interrater agreement, which was particularly more obvious on the novice raters. Her analysis of think-aloud protocols and interviews with four novice raters who demonstrated the greatest change after training revealed that training influenced raters’ decision-making in three ways: helping to clarify the intended rating criteria, modifying the raters’ expectations of student writing, and increasing their concern over interrater agreement. From a CHAT perspective, the effect of rater training on essay rating may involve interactions between interrelated activity systems, which could be a far more complex process than indicated from a cognitive approach. I conduct a CHAT analysis of Weigle’s study to illustrate the training effect on the novice raters when seen through a CHAT lens. There are two reasons for the focus on novice raters. First, novice raters demonstrated more considerable change, hence a better example to illustrate my point; second, there is insufficient information about experienced raters’ decision-making in Weigle’s two articles. Experienced raters are only cited as a reference (to compare with novice raters).

Figure 8. Activity system before training in Weigle (1994, 1998)

Figure 8 is a simulated representation of the activity system before training (shortened as
pretraining activity) in Weigle (1994, 1998). The subjects in the activity are novice raters, teaching assistants (TAs) without prior ESLPE rating experience who were at the same time graduate students in TESL and applied linguistics in the university. Their object in pretraining activity could be marking essays following the rating scale. According to Weigle, the tools they used may have included student essays, writing tasks, the ESLPE rating scale, and their personal experience. Their personal experience may include expertise in TESL or applied linguistics, which could have affected their idiosyncratic interpretations of the rating scale and their expectations about student writing. For example, one rater reportedly commented that some essays did not appear “academic enough” (p. 212) to her. The rule guiding these novice raters could include marking essays based on the rating scale, given that these raters were only provided with the rating scale when asked to mark essays before training. This activity was situated in the community of the university’s ESL section where these novice raters were hired as TAs. Their responsibilities in this activity may include marking essays based on the rating scale. However, lack of clarification and elaboration on the intended interpretation of the rating scale caused some difficulties for these untrained raters in interpreting vague score band descriptors and hence the difficulties in applying the rating scale as intended, as Weigle stated. This could have caused a tension between the tools provided to these novice raters and their object of following the rating scale, as raters may have struggled with insufficient supporting resources to follow the rating scale, similar to the situation encountered by the raters in Lumley (2005). This tension between tools and object could have brought about further tension between rules and object, because raters were required to mark essays based on the rating scale while struggling with following the underdefined rating scale. To solve these two tensions, raters may have drawn on their personal experience to generate personalized interpretations of the rating scale, which
led to literal interpretations of the rating scale and personalized expectations about student writing. The outcome of the activity included that novice raters assigned more severe and inconsistent scores and applied stricter standards overall than experienced raters. In comparison to experienced raters, the score variation in untrained raters might be attributed to their personalized interpretations and elementary understanding of the rating scale.

Figure 9. Activity system during training in Weigle (1994, 1998)

Figure 9 is a simulation of the activity system during training (shortened as training activity). The subjects in the activity are the same novice raters. Their goal for attending rater training could be developing a shared interpretation of the scale and learning how to apply it reliably, in order to solve the tensions associated with their rating difficulties in the pretraining activity. The tools in this training activity may include the rating scale, rater training, sample essays and their scores, and the supervisor who trained the raters and monitored their ratings. According to Weigle, rater training involved clarifying the intended rating criteria, which included making wording in the rating scale less ambiguous, weighting descriptors in the various
score bands, and explaining rating criteria that were not explicitly stated in the rating scale. The rules for this activity may include norming raters’ interpretation of the rating scale and making explicit the desire for interrater agreement. Weigle reported that training procedures included asking raters to compare their scores with each other, such as comparing with the consensus for sample essays and comparing scores with a second rater who rated the same essay as part of the norming process. Weigle noticed that after rater training raters expressed more concern for interrater agreement, suggesting that desirability for interrater agreement may have become explicit during the training session. This training activity was also situated in the same community of ESL section, and these raters’ responsibility may include engagement in the training session. Within this training activity system there was no obvious tension identified based on Weigle’s description. The outcome of this activity included clarified rating criteria, revised personal expectations, and possibly raters’ increased awareness of interrater agreement,. This training activity served as a brokerage activity that aligned the two disconnected rating activities before and after training. The outcomes of this brokerage activity alleviated the tensions in pretraining activity while providing tools for raters in future activities.

Figure 10 is a simulation of the activity system after training (shortened as posttraining activity). The subjects are the same raters who were probably driven by the goal of marking essays based on the intended rating criteria and conforming to other raters’ scores. The tools in this activity may include the same tools available to these raters in pretraining and training activities, and some new tools they acquired from the outcome of the training activity. The rules were probably similar to those in the training activity. This posttraining activity was situated in the same community and these raters’ responsibilities may include marking based on the clarified rating criteria and coping with difficulties not covered by training and the rating scale.
Figure 10. Activity system after training in Weigle (1994, 1998)

This posttraining activity resembles the ERA in Lumley (2005) as illustrated in Figure 7: Raters may hold similar goals and were probably facing a similar difficulty with inadequately defined rating criteria, because rating scale and training can never cover all the eventualities (Lumley, 2005). Therefore, the same kind of tensions among tools, object, rules, and division of labour in Lumley’s (2005) study may apply in this activity too. However, the training activity yielded some outcome that could feed into the tool component in the current activity, as shown in Figure 10, which could have helped to alleviate to some extent tensions in the pretraining activity. Yet those tensions remained; raters may have to turn to the same tool they used in the pretraining activity, that is, their own experience, to compensate for the inadequate rating criteria. The outcome of this posttraining activity included that these novice raters rated more consistently and less severely, which was associated with the tensions alleviated by the outcome of the training activity. However, severity remained an issue among both novice and experienced raters. This
second outcome may be due to the fact that rater training and the rating scale remained inadequate. Another outcome of the activity, raters’ concerns about interrater agreement, may be a result of the training effect which may have increased the raters’ awareness of such desirability.

Rater training examined from a cognitive view attempted to determine linear causality between training effect and essay scores (e.g., Knoch, Read, & von Randow, 2007; Shohamy, Gordon, & Kraemer, 1992; Weigle, 1998), or to explore changes in rater behaviours caused by training (e.g., Furneaux & Rignall, 2007; Weigle, 1994). However, a reanalysis of Weigle’s study from a CHAT perspective suggests that essay rating before and after training may have been two different activities, with different goals to achieve. These two activities involved different kinds of interactions among mediators within their own activity systems, generating different kinds of tensions within each system. However, the two activities were connected by a brokerage activity, the training activity, which could have helped alleviate the tensions in the pretraining activity and provided new tools for the posttraining activity. In brief, this CHAT analysis reveals a more dynamic and holistic picture of how training may have caused change in essay rating, which potentially enhances the understanding of this socializing practice. Similar CHAT analysis can be applied to reinterpreting other studies on rater training (e.g., Elder et al., 2005; Furneaux & Rignall, 2007).

**Contextual tensions.** Some essay rating research examines effects of given factors or variables on rating activity by manipulating these variables while controlling for other variables, with the assumption that any change in the observed outcomes is the result of influences from the manipulated variables. From a CHAT view, change is a complex phenomenon: Any change in one mediator can create tensions in individual activities and further affect the nature of those activities. I use Baker’s (2010) study as an example. In the context of a writing test for a
provincial teacher certification in Canada, Baker compared raters’ ratings for the same essays under the operational rating session (defined as high-stakes condition) and under the research condition (defined as low-stakes condition), and examined whether or not raters marked differently when the consequences to test-takers were different. While essay, rater, rating scale and rater training were unchanged, the t-test results showed that some raters’ ratings remained similar under both conditions, while some assigned more lenient or borderline scores under the research condition. Based on the participants’ responses to a postrating question, Baker noticed that the changed consequences to test-takers were not a salient focus for all participating raters; in contrast, after rerating the essays all the raters expressed their concern about self-consistency with the scores they previously assigned in the operational setting. Baker suggested that this might be due to raters’ concern about rating consequences to their future employment opportunities, and explained that the unexpected findings indicate the different notions of “high-stakes” defined by the scholarship as serious consequences to test-takers, as opposed to the raters’ understanding of “high-stakes” as serious consequences to themselves. Reinterpreting this study with CHAT analysis shows that the changed “test stakes” may have triggered contextual tensions leading to raters’ different activities under the two conditions, which could have explained the unexpected raters’ concern about their own positions over the consequences to test-takers.

Figure 11 is a simulated representation of the raters’ activity in the operational setting of Baker’s (2010) study. The subjects in the activity are the raters for the provincial test. Their goal might include assigning reliable scores for teacher certification purposes, which was aligned with the purpose of this rating employment. The tools they used may include exemplar essays, rating scale, training, essays, and, for some raters, high-stakes consequences to test-takers. The rules
Figure 11. Activity system in operational setting in Baker (2010)

guiding their activity could have included, like most high-stakes tests, marking and conforming to the rating criteria set by the testing institution, maintaining high rater reliability, and certification decisions based on essay scores. Their activity was situated in the testing institution where raters and their trainer (the researcher) both engaged in the activity, with shared responsibilities of facilitating essay marking. The outcome of the activity was that these raters assigned reliable scores that met the requirements of the testing institution, based on which teacher certification decisions were made.

Figure 12 is a simulation of the same raters’ activity in the research setting based on Baker (2010). The researcher made great efforts to establish the same essay rating conditions in the research and operational settings except for consequences of ratings to test-takers. However, from a CHAT view, this rating activity might be completely different than the activity in the operational setting, with only some common elements. Raters’ object in this activity could have
changed; they were no longer marking for teacher certification purposes. Instead, their object might include marking essays by following their operational rating practices to support research (they volunteered to participate in this research project), while maintaining their future employment (based on Baker’s speculation). Most tools in this activity were intended to be the same as in the operational setting, for research purposes. Those same tools may include the same exemplar essays, test-taker essays, rating scale, and rater training. According to Baker’s findings, in spite of the researcher’s efforts, some other tools were unintentionally used in this activity too: consequences to the rater, and scores previously assigned to the same essays. In this research setting, the guiding rules may include (a) marking essays based on the institutional rating criteria, as required by the researcher; (b) not reporting scores to test-takers or the testing institution, as raters were told prior to participation in the research; and (c) raters’ belief that consistent raters are highly valued and desired by the testing institution, which was noticed by the researcher after
the rerating session. The first rule remained the same; while the second and third rules were new, different than the rule of certification decisions based on essay scores in the operational setting. The community of the activity remained the same. The division of labour in this activity could have included (a) facilitating essay marking, which remained the same as the operational setting; (b) the researcher conducting research; and (c) the raters showing self-consistency across settings. The second and third responsibilities were new in this research setting.

Figure 12 illustrates how change in one variable, “test stakes”, could have generated systemic tensions that ultimately changed the nature of these raters’ activities. Three tensions in this activity could have been triggered by a change in activity settings as a result of the trainer/researcher inviting raters to rerate the same essays for her study. The researcher promised that raters’ reratings would be used for research purposes and not be reported to the testing institution. In spite of this debriefing, these raters might still see the researcher—who was their trainer in both settings—as a representative of the testing institution, hence the risk of the testing institution knowing their reratings. This contradiction associated with the dual roles played by the researcher in this study could have generated a circular tension (in the shape of dotted circle) within these raters’ object: Raters were participating in the research while worrying that the testing institution would compare their reratings against their previous ratings and possibly jeopardize their future employment. This systemic tension could have caused the corresponding change in tools, rules, and distribution of labour on raters’ part in the research setting. The new, unintentionally used tools, previous scores and consequences to self, were likely used to ponder how to assign consistent scores and maintain future employment. Although self-consistency was probably also desired in the operational setting, in this research setting desirability of self-consistency across settings seemed to have been prioritized. The first circular tension within
raters’ object may have caused a second tension between rules and object, as it seemed difficult for raters to trust that their reratings would not be revealed to the testing institution. Furthermore, these two tensions may have brought a third circular tension, or contradiction, within the component of division of labour. Raters may attempt to show their self-consistency across settings to the trainer/researcher due to concern about consequences to themselves, while the researcher wanted to find out whether raters would mark differently due to change in consequences to test-takers. The outcomes of this activity in the research setting consequently included (a) raters’ highlighted concern about their self-consistency; and (b) raters assigning similar scores, or more lenient and borderline scores compared with their ratings in the operational setting. One possible explanation to such mixed findings on essay scores is that the raters who showed consistency across the two settings may have been influenced by the self-consistency concern alone; while those who showed more variability could have been influenced by both the self-consistency concern and the absence of high-stakes consequences to test-takers. More information about these raters’ thoughts is needed to better understand what had happened in their minds that have led to such outcomes.

The above reinterpretation of Baker’s (2010) work illustrates that the “only” change in consequences of ratings to test-takers between the operational and research settings might have brought about a series of changes, systemic contradictions and tensions in the ERA system and ultimately changed the nature of the activity. Rather than reducing the effect of change to linear causality, the CHAT analysis could potentially enable a more holistic, in-depth understanding of change in any element of the activity system and its resulting outcomes. Similar CHAT analysis can be applied to other studies comparing the effects of changed conditions (e.g., Elder et al., 2005; Freedman & Calfee, 1983).
Summary

In this section I revisited five studies that follow a cognitive approach to essay rating research to further illustrate differences between the cognitive and CHAT approaches. This CHAT reinterpretation highlights a different, or richer, understanding of their respective findings about rating processes and results. However, all my interpretations are tentative, as there is limited information reported in those studies about the specific research contexts and raters’ thinking about the contexts, which limits the extent of my conclusions. This limitation is because those cited authors did not provide all the details in their publications, and/or their cognitive-approach research designs were not intended to include collecting such information. In the following section, I discuss how essay rating should be investigated when following a CHAT approach, in addition to considerations and qualifications for a CHAT-based essay rating study.

Investigating Essay Rating through the CHAT Approach

Although CHAT conceptualizes essay rating differently than the cognitive approach, it acknowledges strengths of the cognitive approach in addressing cognitive details about raters’ goal-directed actions. Adopting the CHAT approach does not call for completely different research methods, but rather some methodological considerations adapted to assumptions and qualifications of the CHAT approach. Informed by Yamagata-Lynch’s (2010) recommendations for qualitative research using a CHAT approach, this section illustrates how to incorporate this sociocultural view into an investigation of essay rating as a socially mediated activity. As this is an original study incorporating both cognitive and social layers, I use my study as an example to establish a rationale for future studies that adopt this approach. This section explains considerations of CHAT-based essay rating studies in terms of research design. The section then discusses qualifications and accommodations for the corresponding data collection, analysis, and
integration methods that can be used to investigate essay rating in the CHAT approach.

**Research Design Considerations**

From a CHAT perspective, essay rating is no longer an isolated event but rather a situated activity. Therefore, the unit of analysis is no longer individually focused rater decision-making actions or operations, but expands to raters’ collective ERA. CHAT-based essay rating studies uncover not only cognitive details of rater decision-making, the “what”, but also the sociocultural contexts with which raters are interacting when making scoring decisions, to support an understanding of the “how” and “why”. This expanded view drives the researcher to investigate both individual raters’ decision-making processes (goal-directed actions), and collective activities that are directly related to their essay rating practices (object-oriented activity) and constitute the sociocultural contexts of their decision-making.

The research design should thus include identification and selection of activity-related settings, key activities and actions, and participants to decide the “where”, “what”, and “who” engaging in data collection and ensure research questions will be properly addressed (Merriam, 2009). The activity-related setting of a CHAT approach is the environment where essay raters’ object-oriented activities take place and where raters have communal motives and are bound together by the activities in which they engage (Yamagata-Lynch, 2010). This is often ignored by the cognitive approach, which decontextualizes or isolates essay rating. Differing from settings for other qualitative inquiries, the activity-related setting for CHAT inquiries is not necessarily bounded to one physical location. It can cut across multiple physical location boundaries (of both central ERA and interrelated practices) that relate to the object-oriented activities under investigation (Yamagata-Lynch, 2010). In my study, activity-related settings included both the provincial rating centre where the central ERA took place, and the high schools where
In the selected activity-related settings, key activities and actions need to be identified to narrow the scope of the study and ensure that the data collected is relevant, sufficient, and essential to addressing research questions (Yamagata-Lynch, 2010). Since essay rating is not a standalone activity, the key activity should include the central ERA and its related practices that directly interact with the central activity. The central activity, consisting of raters’ decision-making actions, is where the essay rating phenomenon under study occurs. The related practices either contribute to mediating components of the central activity (e.g., forming preparticipation contexts), or are directly affected by the outcome of the central activity (e.g., forming postparticipation consequences). These related practices, together with the central activity, form an activity system as the sociocultural context of essay rating. The cognitive approach is only interested in individual raters’ decision-making actions, regardless of raters’ interactions with rules, their responsibilities in the rating community, and with other practices in related communities (i.e., the sociocultural context). In my study, the focus is on the central NMET ERA at the provincial rating centre, which consisted of NMET raters’ decision-making actions of interest. This central activity was interconnected with raters’ practices at their high schools. Their practices in the school context were considered as relevant, due to two considerations. One is the theoretical consideration about the testing-driven culture in China and the extreme significance of NMET in the secondary education system (Cheng & L. Qi, 2006), while another is the empirical consideration about potential influences from NMET raters’ teaching community (Mei & Cheng, 2014). There can be many everyday practices in the school context that have informed high school English teachers’ perceptions and essay rating activities when serving as NMET raters. Such practices may include the NMET writing test preparation, marking mock NMET
writing tests, incorporating the use of documents and policies concerning NMET writing tests and rating practices, recommending or supporting teachers to participate in NMET rating, and communications about rating practices by novice NMET raters with their teacher colleagues with prior NMET rating experience. All of these practices may influence or be influenced by the central ERA, and hence should be taken into account. The central ERA at the provincial centre and its interrelated practices at high schools constituted the sociocultural context of the NMET essay rating in this study.

The identified activity-related settings and their related key activities and practices can then be used to identify participants in these activities and whose inclusion in the study would be valuable. Whereas participant groups are confined to individual raters in the cognitive approach, the CHAT approach requires involving not only individual raters as participants for investigation of goal-directed decision-making actions, but also other participants who share the same communities with raters in the central activity and its interconnected practices, in order to obtain multiple perspectives about the object-oriented activity. In my study, participants from the rating community included directors of the provincial centre, essay rating team leaders with trainer roles, and NMET essay raters. Participants from raters’ teaching communities included the aforementioned NMET raters, their principals, and teaching colleagues who were prior NMET raters. The same NMET raters were my participants for investigation of NMET raters’ goal-directed actions, in addition to their activities across the two settings.

**Data Collection Methods**

Different data collection methods are involved in addressing research questions concerning ERA, by examining individual raters’ goal-directed decision-making actions (cognitive layer) and raters’ collective object-oriented ERA (social layer). The cognitive
approach has important implications for investigating the cognitive layer of ERA, because its focus is on individual raters’ decision-making actions, investigating individual raters’ thought processes when they assess essay quality. Various techniques have been adopted to explore raters’ cognitive processes, such as think-aloud protocols (e.g., Crisp, 2008; Cumming, 1990), write-aloud protocols (e.g., Baker, 2012), stimulated recalls (e.g., Erdosy, 2005; Huot, 1993), follow-up interviews (e.g., Huot, 1993; Milanovic et al., 1996), and immediate retrospective written reports (e.g., Milanovic et al., 1996). Despite its strengths in tapping into raters’ thought processes and addressing the cognitive layer of ERA, the cognitive approach usually makes assumptions about raters’ goals or purposes of rating (i.e., all raters hold the same goals as expected by the researcher or the testing institution). My reinterpretation of previous studies in this chapter suggested that this may not always be true. Therefore, it is important to ask raters questions about their rating goals in addition to collecting data on their thought processes for a complete investigation of the cognitive layer. Investigation of the social layer, however, goes beyond the scope of the cognitive approach. According to Yamagata-Lynch (2010), frequently-used data collection methods for the social layer include interviews (to learn about participants’ experiences, opinions, feelings), document analysis (to extract excerpts, or entire passages of documents relevant to participants’ activities), and observation (to obtain detailed descriptions of participants’ activities). All data collection methods have their strengths and limitations; hence the choice of methods will be based on practical constraints and feasibility considerations. In the rest of this subsection, I focus on four of my selected methods (think-aloud protocols, stimulated recalls, interviews, document analysis) for this study, discussing their strengths and limitations when used in CHAT-approach studies.

**Think-aloud protocols (TAPs).** TAPs (also referred to as concurrent verbal reports) have
been most frequently used in exploring raters’ thought processes (e.g., Crisp, 2008; Cumming, 1990; H. Li & He, 2015; Lumley, 2005; Sanderson, 2001; White, 1993). To produce TAPs, raters are usually trained and instructed to verbalize their thoughts while assessing a set of essays. Raters’ verbal protocols are recorded, transcribed, and analyzed to identify raters’ decision-making processes in terms of their behaviours and aspects of writing they attend to when assessing essays. In my study, TAPs were used to examine NMET raters’ decision-making actions while assessing essays.

TAPs have several advantages in tapping cognitive processes of raters’ goal-directed actions and automatic operations. First of all, TAPs can help make thought processes—not directly observable in the external world—observable at some level (Gass & Mackey, 2007). Based on TAPs, inferences can be drawn about what takes place in rater’s minds, so as to explain how and why a rater comes to a scoring decision (Barkaoui, 2011b). Compared to other introspective methods, such as stimulated recalls, TAPs provide an online or concurrent report of what is going through a rater’s mind when assessing an essay, rather than a delayed or retrospective report of a rater’s thought processes, thus avoiding problems of information retrieval and filtering (Green, 1998; Mackey & Gass, 2005). Furthermore, compared to other self-report tools, such as interviews, TAPs are more likely to reflect what raters actually do and are concerned about as they read and rate essays, rather than what they believe they do and are concerned about (Huot, 1993). Lastly, TAPs provide specific instances of raters’ actual goal-directed actions, rather than general statements about actions usually obtained from self-report instruments such as interviews and questionnaires (Barkaoui, 2011b).

However, TAPs have several limitations. TAPs are not a neutral method; from a CHAT perspective, they are socially situated practices. One major criticism of TAPs is reactivity, that is,
whether asking participants to report thinking processes alters the processes being observed (Barkaoui, 2011b; Bowles, 2010; Ericsson & Simon, 1993; Green, 1998). Verbalization is not just a report of thought; it potentially transforms thinking, as a process of comprehending and reshaping experience (Barkaoui, 2011b; Smagorinsky, 2001; Swain, 2006). During this process, the participants may interact with the researcher’s object-oriented activity (research activity), which could potentially affect the transformation of their thinking. TAPs were found to be dialogic (addressing an audience), hence demonstrating their interactive and social nature (Barkaoui, 2011b; Sasaki, 2003; Smagorinsky, 2001). Another major criticism concerns veridicality of TAPs, that is, whether TAPs represent an accurate and complete report of participants’ thinking processes (Barkaoui, 2011b; Ericsson & Simon, 1993; Lumley, 2005). As indications of what participants are attending to when they try to complete a short task, TAPs make participants focus only on the goals associated with that task, which makes participants seem static and disembodied from their long-term individual development and their social relations (Swain, 2006). In addition, there are individual differences in the quality and quantity of produced TAPs, as participants may not be used to verbalizing their thoughts while focusing on completion of a task (Green, 1998; Smagorinsky, 1994). Also, the entire process of collecting, transcribing, coding, and analyzing data from TAPs is time-consuming and labour-intensive (Barkaoui, 2011b).

Despite those limitations, as Barkaoui (2011b) pointed out, TAPs are still a useful method for identifying raters’ goal-directed decision-making actions. Though incomplete and imprecise, TAPs are the only tool available thus far to provide insight into the kind of processes that raters actually employ to complete rating tasks (Barkaoui, 2011b). Also, reactivity and incompleteness are not unique to TAPs. It is impossible to neutralize the social context in any human research
(Smagorinsky, 2001). Other methods, such as stimulated recalls, interviews, and observation, have the same problem.

Rather than ignoring the presence of those effects, researchers should be aware of and try to understand them. Researchers need to identify and account for the trajectory of participants’ verbalizing actions (Smagorinsky, 2001). When participants perform verbalizing actions, researchers and prompts become mediators. Hence researchers need to be aware that their relationship with participants can influence the produced protocols. This implies that interpreting TAPs requires establishing rapport with participants before producing TAPs and acquiring knowledge about participants’ prior experience that is related to or mediating their verbalizing actions. These efforts allow for words to be understood in the same way by both researchers and participants. Additionally, researchers can adopt the following four strategies to address those limitations (Barkaoui, 2008, 2011b; Smagorinsky, 2001): (a) providing raters with careful and detailed training and instructions on how to verbalize their thoughts while rating essays (see my description in Chapter 3 as an example); (b) interviewing raters about their perceptions of the TAP task and how their prior experiences mediated that perception, and the corresponding effects on thought processes and completeness of their verbal protocols; (c) inspecting the produced protocols for evidence and explanations of incompleteness, reactivity, and variation in verbalization quality and quantity; and (d) combining the use of other methods (e.g., stimulated recalls), asking raters to describe their rating processes retrospectively, and then comparing these descriptions with concurrent TAPs to assess completeness and reactivity of TAPs.

**Stimulated recalls (SRs).** SRs are another introspective method to investigate raters’ thought processes. Unlike TAPs, SRs are produced retrospectively. Raters are usually prompted with visual or audio recall support (e.g., essays they rated, video or audio recordings of how they
carried out rating tasks) and instructed to report thoughts they had while reading and assessing essays after they performed the rating task. SRs are also used in some CHAT-based studies to probe into participants’ goal-directed actions (e.g., Cumming, 2006; Engeström, 2000; Lei, 2008). In my study, SRs were used immediately after TAPs to further explore raters’ thought processes.

As an introspective technique, SRs offer advantages in exploring raters’ goal-directed actions. Unlike TAPs, SRs do not require training raters to verbalize their thoughts (Gass & Mackey, 2000). SRs can also help to gain information about specific instances of raters’ actions without disrupting or altering their thinking processes to be observed (Erdoesy, 2005; Ericsson & Simon, 1993). Finally, compared with simple follow-up interviews, SRs generate retrospective reports of raters’ concrete actions based on recall support, hence providing protocols of raters’ thought processes with higher accuracy and specificity (Gass & Mackey, 2000).

SRs, however, also have their limitations. Apart from the drawbacks of being time-consuming and labour-intensive, the major criticism against SRs is nonveridicality. Veridicality of data obtained from SRs can be threatened due to three concerns. First, accuracy and completeness of SRs often decreases with time: The more delayed SRs are, the less accurate and complete raters’ reports of their thoughts tend to be (Ericsson & Simon, 1993; Gass & Mackey, 2000; Lyle, 2003; Russo, Johnson, & Stephens, 1989). Like TAPs, SRs are also socially mediated; influences of raters’ prior experience and their relationship with researchers need to be identified and accounted for (Smagorinsky, 2001; Swain, 2006). Therefore, the second concern is that the procedure of producing SRs is a process of raters interacting with researchers’ activity, reconstructing thoughts, presenting raters’ selective evaluations and inferences of what has occurred previously, which makes produced SRs susceptible to fabrication (Ericsson & Simon,
Finally, as in interviews and TAPs, raters may filter or sanitize their thoughts, reporting what they think researchers want them to say, rather than what they actually did (Gass & Mackey, 2000; Green, 1998; Lyle, 2003).

In spite of those limitations, SRs still provide a useful means of accessing raters’ thought processes. It can help identify the type of actions raters employ when assessing essays, whether or not these actions are organized in specific ways, and when and if certain actions are being employed (Gass & Mackey, 2000). In addition, those threats to veridicality are not unique to SRs. Other research methods, such as TAPs and interviews, also face the same problems of incompleteness, inaccuracy, and/or thought sanitization.

To address those limitations, researchers can employ the following three strategies as suggested by Gass and Mackey (2000) and Lyle (2003). The first strategy is to conduct consecutive SRs, that is, carrying out SRs immediately after raters assessed essays and thought aloud, to reduce the effects of the time lag. At the same time, as SRs follow TAPs, researchers have gained knowledge of raters’ prior experience. The second strategy is to provide raters with careful instructions and develop detailed plans for carrying out SRs (see my description in Chapter 3 as an example). The third strategy is, as discussed, to compare SR protocols with TAPs to assess the completeness of SRs.

**Interviews.** Interviews are a widely used technique in qualitative research as they allow researchers to enter into participants’ minds and explore their experiences, feelings, perceptions, and intentions that are not directly observable (Mackey & Gass, 2005; Merriam, 2009; Patton, 2002). In my study, interviews were used primarily to investigate raters’ collective object-oriented activity based on multiple perspectives.

The use of interviews has several advantages in CHAT-based studies. First, interview
participants may introduce researchers to other potential research participants (Yamagata-Lynch, 2010). Second, through interviews, participants may share information about documents and artifacts related to mediators (e.g., existing rules, division of labour) in their object-oriented activities (Yamagata-Lynch, 2010). Third, interviews are targeted, as they focus directly on the research topic, helping the researcher identify information about mediators and goals/objects of the activities of interest (Yamagata-Lynch, 2010; Yin, 2009). Lastly, interviews are insightful, and can provide participants’ perceived causal inferences and explanations to their object-oriented activities (Yin, 2009).

However, interviews cannot be used without caution. As another socially mediated practice, an interview can be inaccurate and incomplete, as a result of interviewees’ selective recall, self-delusion, or memory loss (Hall & Rist, 1999; Yin, 2009). Interviews can also be biased, due to interviewees’ perceptual distortions and researchers’ leading questions (Hall & Rist, 1999; Yin, 2009). Another drawback of interviews is veridicality, not unlike the issue with TAPs and SRs. Interviewees may give what they think the interviewer wants to hear, instead of what they actually did and thought (Yin, 2009).

Interviews, if well designed and used carefully, can help collect rich data that illuminate the research phenomenon. To address the weaknesses of using interviews, researchers can employ two strategies. The first strategy is to develop and pilot interview guides, asking open-ended questions such as hypothetical and interpretive questions (Strauss, Schatzman, Bucher, Ehrlich, & Sabshin, 1981), while avoiding leading questions. The second strategy is to conduct multiple interviews with the same participants and interview different participants to assess completeness of the interview data.

**Document analysis.** Document analysis is a research method for systemically
reviewing and analyzing both printed and electronic documents and artifacts, in order to elicit
meaning, gain understanding, and develop empirical knowledge (Bowen, 2009). This method is
often used in CHAT-based studies to provide contextual information about the activities of
interest (e.g., Barrett-Mynes, 2013; Lei, 2008; Yamagata-Lynch & Haudenschild, 2009). In my
study, document analysis provided archival information about the broader and immediate
sociocultural contexts that potentially interacted with the NMET ERA.

As a research method, document analysis has many advantages. First, unlike TAPs and
interviews, documents are non-obtrusive and non-reactive, in that they are not created for the
purpose of the research, and are unaffected by the research process (Bowen, 2009; Merriam,
2009; Yin, 2009). Second, documents are stable, and can be retrieved repeatedly (Yin, 2009).
Third, documents provide broad coverage: extended time span, many incidents, and many
settings (Yin, 2009). Fourth, documents provide exact information about names, references, and
details of incidents (Yin, 2009). Finally, it is less time-consuming and less costly compared to
other qualitative methods such as observation (Bowen, 2009).

However, document analysis has potential flaws. Documents may be incomplete or
provide insufficient information, since they are not produced for the purpose of the study
(Bowen, 2009; Merriam, 2009). Also, documents are sometimes not retrievable or the access is
deliberately blocked (Bowen, 2009; Yin, 2009).

In spite of its flaws, document analysis provides a valid method for providing contextual
information. An incomplete collection of documents suggests biased selectivity, possibly because
only records in alignment with the organizational policies are kept and available, but can still
provide researchers with some information about the context (Bowen, 2009; Merriam, 2009; Yin,
2009). To address this limitation, researchers can use document analysis in combination with
other methods—such as interviews—to provide multiple perspectives (Patton, 2002) and hence reduce the effect of selectivity biases. Findings from document analysis can be used to inform interview questions, prompting participants to share their perceptions of the contexts to help enhance researchers’ understanding of the sociocultural contexts of interest.

**Data Analysis and Integration Strategies**

CHAT-based studies usually involve collecting qualitative data using research methods such as the four I discussed. The goal of qualitative data analysis is to make sense of collected data, which is primarily an inductive and comparative process (Merriam, 2009). As Saldaña (2009) recommended, data analysis can be carried out in two cycles. In the first cycle, all data sets can be analyzed using the open coding technique from constant comparison method (CCM; Strauss & Corbin, 1990). Due to the different focuses and nature of data on cognitive and social layers, different data sets can be analyzed using slightly different strategies. In the second cycle, data can be examined across data sets using the axial coding technique from CCM, before being integrated using the CHAT framework.

CCM is a systematic qualitative analytic method that engages researchers in an intense, systematic process of examining and reexamining the data while comparing one source with another to find similarities and differences (Strauss & Corbin, 1990). *Open coding* involves “a process of breaking down, examining, comparing, conceptualizing, and categorizing data” (p. 61). In the first cycle of data analysis, data for the social layer (e.g., interview and document data) can be analyzed using the open coding technique to describe the collective ERA. Data for the cognitive layer of NMET ERA (e.g., TAP and SR data) can be analyzed using the open coding technique while conceptually informed by coding frameworks developed by previous studies following the cognitive approach. Among the various frameworks about rating processes,
I find Cumming et al.’s (2002) framework useful to describe instances of raters’ thought processes. This framework includes 27 decision-making behaviours, categorized dimensionally by focus and strategy. The focuses include: (a) self-monitoring of one’s own rating process; (b) ideational and rhetorical elements of the essay; and (c) language accuracy and fluency of the essay. The strategies include (a) interpretation strategies used to comprehend the essay, and (b) judgment strategies to evaluate the quality of the essay. This framework accounts for both raters’ rating behaviours and the aspects to which raters attend. The matrix format is also efficient at identifying raters’ decision-making behaviours, categorizing data, and making adaptations by revising, removing, and adding new codes that best describe specific data. In sum, the analysis of data for the cognitive and social layers can adopt slightly different strategies to accommodate their different focuses. Axial coding includes a set of procedures making connections between categories and putting the data back together in new ways, by “utilizing a coding paradigm involving conditions, context, action/interactional strategies and consequences” (Strauss & Corbin, 1990, p. 96). In the second cycle of data analysis, the axial coding technique can be used by linking the codes generated across data sets from the first cycle and grouping them into categories and subcategories, based on Strauss and Corbin’s codes relationship paradigm. In addition to this coding process, findings can be further integrated based on the CHAT conceptualization to validate relationships between categories and subcategories, making sense of the data by forming a comprehensive picture of the two-layered ERA.
Chapter 3 Method

In this chapter I describe the study settings before justifying the overall research approach I took and explaining my roles in this study. I then provide an overview of the research design and details of corresponding participants, materials and instruments, and procedures of data collection, analysis, and integration employed in addressing the three research questions. I describe my efforts in enhancing study validity and reliability to close this chapter.

Study Settings

The settings of interest for this study included two types of settings related to the NMET raters’ ERA: one provincial rating centre where the participating raters rated NMET essays, and four high schools where the participating raters taught. To provide background information of the NMET ERA, I describe these two activity-related settings and the activities/practices related to the NMET ERA at each setting.

Rating Centre

The rating centre was located on a university campus in a provincial capital in China. The centre was isolated from other campus buildings and its entrance was monitored by security guards. Some banners, written in Chinese, hung inside and outside the building, displaying slogans such as: “Iron shoulders take on morality and justice; every point assigned shall not fail in test-takers’ ten years of toiling and moiling”; “Select the talented individuals for our country; dedicate selflessly”; and “Be careful and serious; be meticulous; rate reasonably, neither too strict nor too lenient.” Inside, each room had computer terminals connected to the centre’s on-screen marking (OSM) system, and raters worked in teams. At the front of each room, a computer reserved for team leaders was connected to a projector.

This rating centre, like rating centres for other Gaokao subjects at the same location, was
organized and administered by a local university. The centre had a director and an associate
director. They were under the supervision of the Provincial Administrative Centre for Enrollment
and Examinations and supported by a working group mostly made up of team leaders responsible
for major tasks. There were three types of tasks at this centre: rating (to mark essays and other
items), arbitration (to resolve discrepancy ratings), and quality inspection (to monitor rating
quality). Each essay rating team was led by an expert rater, who reported directly to the directors.
There were also one arbitration team and one quality inspection team for essay rating. In this
centre, essay raters were all high school English teachers from around the province. These raters
were made up of experienced raters who had worked as NMET raters before, and new raters who
were recommended by their high schools as a response to a mandate from the provincial
Education Department.

The rating procedures at this centre took place over a 9-day intensive work period
typified as follows:

- Day 1: The working group discusses operational rating specifications to complement the
  NMET rating scale, and selects anchor essays and testing essays for training purposes;
- Day 2: Rater training (including plenary mobilization meeting, OSM system operation
  practices, application of rating scale and specifications, and pilot rating);
- Days 3-9: Operational rating (including rating essays, ongoing informal rater training
  and reorientation, arbitrating discrepancy ratings, and quality inspection).

The rating procedures were facilitated by the OSM system. Test-takers’ answer sheets
were digitally scanned, organized by test item section, and uploaded to the system. Essay raters
retrieved the scanned essay section from the system and directly assigned scores to essays on
screen. Each essay was randomly assigned to two raters and their ratings were averaged to obtain
a composite score. In cases of rating discrepancies beyond the acceptable level (in this case, greater than four points out of 25), the OSM system automatically sent disputed essays to a third rater. If the discrepancy among the three ratings remained unacceptable, a fourth rater was used to finalize the scoring. Individual raters’ statistics (e.g., number of completed ratings, [in]valid rating rates, serious rating error rates) were computed and accessible to essay team leaders and the quality inspection team for monitoring purposes. The OSM system was also used for purposes such as rater training and reorientation, monitoring of ratings, and providing feedback for ratings.

**Schools**

The four high schools were located in different places within the province (two in urban and two in suburban areas), but they had much in common in terms of their physical environment, school structure, and Gaokao-related practices.

Each school had more than 1,000 students at each grade with a class size of 60 to 80 students (about 15 to 18 classrooms per school). On the walls of their Grade 12 teaching buildings hung posters with profiles of graduates who entered top universities, and rankings of students based on results of monthly Gaokao mock tests. Outside each classroom were profiles of high-achieving students or students who made the biggest progress in the previous month’s test results. There were also hanging banners with Gaokao-motivating slogans inside and outside the classrooms, such as: “One year’s toiling and moiling, life-long success with no regrets”; “The wise take fame and fortune as a stepping stone and always remind themselves to keep working hard; the smart use success as an opportunity to reflect on themselves and keep cool-headed”; “Wish you to be a fish leaping over the dragon gate (“to be successful”) and have a prosperous future!” A countdown of Gaokao test-taking days was noted in some Grade 12 classrooms.
reminding students of the time left for test preparation.

In all four high schools, everyday teaching was organized and supervised by the principal and vice principals, supported by the office of teaching affairs. Major duties of the principals included defining teaching goals for each grade, coordinating and supporting teaching and learning activities, monitoring teaching progress, ensuring and improving the quality of teaching and learning, and supporting student test preparation for Gaokao. Gaokao preparation played a dominant role in everyday teaching and learning practices, especially in Grade 12. School administrators, with the help of head teachers, developed Gaokao preparation plans to guide Grade 12 teaching practices. The head teachers then monitored how well teachers carried out these plans.

In all four schools, formal Gaokao preparation began immediately after students entered Grade 12 in September and continued until the Gaokao administration in early June. All these schools compressed three years of high school teaching tasks into two years, so the third year could be devoted to Gaokao preparation. It was a common practice to teach all 14 subjects required in high school curricula in Grade 10. The nine mandatory Gaokao–related subjects and remaining curricular requirements occurred in Grade 11. In the first half of Grade 12, teachers focused on reviewing the Gaokao mandatory subjects, and preparing students for Gaokao based on the official test syllabi. Monthly school-based tests on Gaokao mandatory subjects were organized and administered throughout the three years in high school, while in Grade 12 the tests were more targeted and focused. Test papers were usually modelled after previous Gaokao test papers, and students’ answers were marked following Gaokao rating criteria. Students’ performance on these monthly tests were ranked and posted on campus as a means of keeping

---

3 In China, high school education usually lasts three years, from Grade 10 to Grade 12.
both students and teachers motivated for better performance in future tests. The second half of Grade 12, known as Chongci (dashing to the finish), focused on familiarizing students with the test formats and improving their test-taking skills. At this stage, students wrote several rounds of mock Gaokao tests, usually organized by the city- or county-level education departments. These tests simulated Gaokao tests in terms of test contents, formats, and test-taking procedures. The test results were used to assess students’ readiness for Gaokao and predict their Gaokao performance.

**Qualitative Case Study Approach**

I employed a qualitative case study approach for this study, with the purpose of developing an in-depth understanding of a phenomenon (i.e., NMET ERA) within a bounded system (Merriam, 2009). Yin (2014) defined the scope and features of a case study that makes it distinct from other forms of empirical inquiries used in social science studies, especially those using qualitative methods.

According to Yin (2014), the scope of a case study involves “investigating a contemporary phenomenon (the ‘case’) in depth and within its real-world context, especially when the boundaries between phenomenon and context may not be clearly evident” (p. 16). This study examined how participants performed their roles when engaging in the NMET ERA and why they performed in such ways, in order to develop an in-depth, situated understanding of the NMET ERA. Participants’ performance was studied with limited intrusion by research procedures, and they had the freedom to express themselves through interviews and their own notes or writing without being constrained by preestablished research tools. The boundaries between the NMET ERA of study and the context were blurred, as the context is considered an essential part to the ERA through a CHAT lens.
Yin (2014, p. 17) provided three defining features of a case study: (a) involving multiple sources of evidence, (b) coping with the technically distinctive situation in which there will be many more variables of interest than data points, and (c) data collection and analysis guided by prior theoretical propositions. This study collected and integrated data from multiple sources to capture the complexity and dynamics of the NMET ERA of interest. By including perspectives of multiple groups of participants, this study revealed how NMET raters interacted with the sociocultural context, and how their decision-making affected and was affected by such interactions. Such complexity cannot be fully captured if studied using a singular vantage data point. The operationalized framework of cultural-historical activity theory in essay rating research, which I outlined in Chapter 2, guided the research design and corresponding data collection and analysis procedures to address my research purpose and questions.

In addition to the above reasons, I chose a case study approach because it is compatible both theoretically and conceptually with CHAT inquiries. The case study approach is frequently used in CHAT-based research (e.g., Barrett-Mynes, 2013; Greig, 2008; Lei, 2009; Marken, 2006; Parks, 2000). In theory, case study focuses on describing, explaining, or exploring the “how” and “why” of a contemporary phenomenon (Yin, 2014). The CHAT approach focuses on understanding ERA as socially mediated, by situating raters’ goal-directed decision-making actions within their object-oriented ERAs in natural settings, to develop an in-depth description and explanation in terms of how and why raters perform the ERA, which determines how they assess essays. Conceptually, case study is interested in examining a phenomenon within bounded systems, or “cases” (Merriam, 2009). A case is the unit of analysis; it can be an individual person, program, institution, community, relationship, or activity (Yin, 2009). In CHAT inquiry, an object-oriented activity constitutes the case of study, often represented as a bounded activity.
system including subject, object, and the four mediators. In my study, the object-oriented NMET ERA (i.e., case) was my unit of analysis and embodied the global nature of the activity. Considering the unique testing-driven culture and the high impact of NMET in Chinese society, selecting the NMET ERA as a critical “case” enabled me to study how raters interacted with the sociocultural context when rating NMET essays within the Chinese context.

The boundaries of this case were defined in terms of place (where), time (when), relevant participant group (who), and the type of collected evidence (what; Yin, 2014). The study examined raters’ NMET ERA at the two activity-related settings (rating centre and schools) around two (2014 and 2015) NMET operational rating sessions. The overall design was a single case of raters’ object-oriented NMET ERA. Multiple groups of participants from the two settings who interacted closely with raters were chosen to gain a multi-faceted perspective of this selected case. Multiple sources of evidence were collected to explore both cognitive and social layers of this NMET ERA and gather a rich and contextualized understanding about essay rating in this NMET context.

**Role of the Researcher**

For a qualitative case study, it is important to provide information about the researcher’s role. The researcher is a human instrument who interacts with participants in the field; the role the researcher assumes thus has an impact on the collected data. Also, data analysis and interpretation of qualitative research findings are heavily influenced by the researcher’s prior knowledge of the research site and perspectives about the phenomenon of interest. Reporting how my perspective affected the interpretations of participant perspectives and experiences in the field both maintains trustworthiness throughout the investigation (Yamagata-Lynch, 2010) and allows the reader to assess my potential prejudices (Yin, 2011).
In this study I assumed different roles in two rounds of data collection. In the first round, I was an observer; I collected data six months after the 2014 NMET operational rating session and therefore there was no overlap between participants’ activities and mine (Yamagata-Lynch, 2010). As I had prior work and research experience (working as a rater for essay writing and other items, and as an investigator of two research projects involving essay raters) at the provincial rating centre of study, I was familiar with the research site and to some extent had an insider’s perspective on the central ERA. In the second round, I was a participating member in the major tasks for 2015 NMET rating at the rating centre, assuming the role of participant as observer who engaged in participants’ everyday activities but maintained the role as a researcher (Yamagata-Lynch, 2010). Before the operational session, as a working group member I attended two preparation meetings, developing rating specifications with other members in the group and selecting anchor essays. During the session, I worked as an essay rating quality inspector, monitoring ratings of all essay rating teams and reporting to the quality inspection team leader (a study participant), but I did not have direct contact with other raters. At work, I occasionally consulted with an arbitration team leader (another participant) to discuss essay scores. In sum, I contributed to the realization of the NMET essay rating. I approached participants outside of work and collected data from these participants after the operational rating session. I assumed an observer role at the high school setting in both rounds.

The two roles (participant as observer and observer) I assumed in this study potentially influenced, both positively and negatively, the trustworthiness of the study and introduced bias. I maintained a good rapport with participants from the rating centre, some of whom I have known from prior experiences at this rating centre or in previous research projects. I also established rapport with participants at the four schools during data collection from connections made at the
rating centre and relationships with team leaders. The role of *participant as observer* allowed me to acquire a rich collection of first-hand knowledge about participant experiences at the rating centre (Yamagata-Lynch, 2010), contributing to an in-depth understanding of the central NMET ERA. Meanwhile, as a researcher and member in the rating community, my interactions with participants may cause change in their object-oriented activities under investigation. I might also lose sight of an outsider perspective in this role, as I acclimated to participant practices and it became less difficult to understand participant experiences (Yamagata-Lynch, 2010). This would introduce bias during my data collection and analysis. The role of *observer* allowed me not to interfere in participant activities. Also, there may be information that participants were more comfortable sharing with “outsiders” than “insiders” (Yamagata-Lynch, 2010). However, at the same time, my inferences about participant activities and practices lacked first-hand knowledge. On the whole, there were both potential costs and benefits in either role I took. Besides understanding and evaluating those aspects, I employed some strategies to help reduce the negative influences, which I summarize in the last section of this chapter.

**Overall Research Design**

The purpose of this study was to examine how the NMET ERA was mediated by its sociocultural context of the high-stakes Chinese university entrance English examination. Drawing on the CHAT framework, this study explored two layers of the NMET ERA, individual raters’ goal-directed actions (cognitive), and collective object-oriented activity (social). I explain the overall research design in relation to the study’s research questions.

**Research Questions**

The following three research questions were addressed in this study:

**RQ1: How do raters assess NMET essays to achieve their goals?**
This research question explores the cognitive layer of NMET ERA by examining two aspects:

- What rating behaviours do NMET raters engage in while marking NMET essays?
- What are their goals when marking specific NMET essays?

Rating behaviours refer to the actions raters take while reading and rating essays, including aspects of writing they attend to, sequential rating procedures they follow, and what they consider when deciding on final scores. Goals refer to raters’ underlying purposes of assessing individual essays. These aspects describe NMET raters’ goal-directed actions while they assess NMET essays. Individual NMET raters were the participants in the investigation of this cognitive layer. To answer the above two subquestions, I chose to listen (concurrent TAPs, retrospective SRs) and ask (interviews).

RQ2: What are the broader (Chinese society) and immediate (rating centre and school) sociocultural contexts in which the NMET ERA is situated?

This research question explores the social layer of the NMET ERA. The sociocultural contexts are operationally defined as *social, cultural, institutional, and political environment* where the NMET ERA takes place, which specifies tools, rules, community, division of labour, and their interactions within the NMET ERA. The broader context focuses on the Chinese society in general, while the immediate context focuses on the high schools and the rating centre where the NMET ERA of question takes place. The high schools were considered a related setting based on theoretical (i.e., the testing-driven culture in China and the enormous significance of NMET in secondary education system; see Cheng & Qi, 2006) and empirical considerations (i.e., potential influences from NMET raters’ teaching community; e.g., Mei & Cheng, 2014). It should be noted that NMET, the English component of Gaokao, is literally
translated as “Gaokao English” in Chinese. As mentioned in Chapter 1, Gaokao includes other components, such as Chinese, mathematics, sciences or social sciences, all of which include some involvement by human raters. All of these Gaokao components share similar sociocultural contexts with regard to rating activity. Therefore, I employ the term “NMET rating” when issues involved in rating essays of Gaokao English component are discussed. In other situations, the term “Gaokao rating” is a broader term that encompasses NMET rating. Participants for investigation of this social layer included the same raters who participated in TAPs and SRs for RQ1 and additionally others who shared the raters’ rating and teaching communities. I used the document analysis and interview research methods to address RQ2. I used document analysis to study the following sociocultural contexts:

- the impact of Gaokao rating on the society, in general;
- the NMET rating centre; and
- Gaokao rating in high schools.

Document analysis findings informed the design of my interview instrument to explore participants’ perceptions regarding the three contexts. Document analysis and interviews provided multiple perspectives to help me understand the broader and immediate sociocultural contexts where the NMET ERA took place.

**RQ3: What is the nature of NMET ERA as an activity system within the above sociocultural contexts?**

This research question explores the NMET ERA as a whole (i.e., both cognitive and social layers). The NMET ERA is operationally defined as an activity system consisting of the central rating activity and its interrelated practices at the two settings. The NMET ERA, therefore, is studied from the following aspects:
• the central activity system at the rating centre with different mediators (i.e., tools, rules, community, division of labour) and systemic tensions;
• the interrelated practices in the school context that influence and are influenced by the central activity; and
• interconnectedness between the central activity and these practices.

The above aspects were studied by integrating the results from all data sources (TAPs, SRs, document analysis, and interviews) and understanding them through a CHAT lens. This CHAT lens allowed the examination of the cognitive layer in relation to the social layer, which provided a contextualized understanding of NMET raters’ purposeful rating behaviours.

Research Design

To address the three research questions, I employed a qualitative multiple-method case study design based on multiple perspectives. The overall design is summarized in Table 2.

Table 2

<table>
<thead>
<tr>
<th>RQs</th>
<th>CHAT layers</th>
<th>Participants</th>
<th>Data</th>
<th>Data analysis and integration</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. How do raters assess NMET essays to achieve their goals?</td>
<td>Cognitive</td>
<td>NMET essay raters (n = 8)</td>
<td>TAPs</td>
<td>Open and axial coding; CHAT</td>
</tr>
<tr>
<td></td>
<td></td>
<td>At rating centre:</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1) Directors</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2) Team leaders</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>3) Same essay raters</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>At schools:</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>4) Principals</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>5) Same essay raters</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>6) Teaching colleagues</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(n = 25)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. What are the broader and immediate sociocultural contexts?</td>
<td>Social</td>
<td>All participants (n = 25)</td>
<td>Documents, Interviews</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. What is the nature of NMET ERA?</td>
<td>Cognitive + Social</td>
<td>All participants (n = 25)</td>
<td>All data sources</td>
<td></td>
</tr>
</tbody>
</table>
To address RQ1, which examines the cognitive layer of ERA, eight NMET essay raters were asked to assess NMET sample essays as if in operational NMET rating sessions, while producing concurrent TAPs, followed by consecutive SRs and interviews. These raters first verbalized their thought processes while assessing the essays and then explained their underlying purposes and considerations behind those thought processes.

RQ2 explores the social layer of ERA, which investigates the broader (Chinese society) and immediate (rating centre and school) sociocultural contexts of NMET ERA through documents and interviews. Documents provided archival information about the sociocultural context that potentially informed the ERA of interest. Interviews explored participants’ perceptions of the sociocultural context in which the NMET ERA took place. A total of 25 participants were interviewed, including the same eight essay raters, directors, and team leaders at the rating centre, and the same raters’ principals and teaching colleagues at their schools.

RQ3 examines the nature of NMET ERA by investigating both the cognitive and social layers. All participants from the two activity-related settings and all the four data sources were involved. The findings formed a comprehensive picture of the cultural, historical, and social nature of the NMET ERA in this study.

All four sources of data were analyzed using the open and axial coding techniques, then integrated using the CHAT framework. In the following sections, I elaborate how I carried out the above research design in this study.

**Participants**

This study included 25 participants from the rating centre and four high schools (Schools A, B, C, D). These participants included two directors, seven team leaders (four acted as raters’ teaching colleagues), eight essay raters, three principals, and five teaching colleagues. Due to
practicality and convenience considerations, directors and team leaders were recruited via the researcher’s prior work relationships, while essay raters, principals, and teaching colleagues were recruited through connections with team leaders. Table 3 illustrates the relationship between the participant groups and the two activity-related settings. The four high schools that were selected to study the school practices participated in NMET rating and sent teachers to serve as essay raters in 2014 and 2015. These schools’ essay raters, school principals, and their teaching colleagues (20 participants in all) were included in the study. All participants, except for the three school principals, had recent experience with NMET essay rating at the rating centre and were thus included to study the activity in the rating centre setting. At the four high schools, team leaders, essay raters, and teaching colleagues were involved in examining both settings.

Table 3

*Participants’ Involvement in Relation to Investigation of Activity/Practices in the Two Settings*

<table>
<thead>
<tr>
<th>Participant group</th>
<th>Rating centre</th>
<th>High schools</th>
</tr>
</thead>
<tbody>
<tr>
<td>Directors (n = 2)</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Team leaders (n = 7)</td>
<td>✓ (n = 7)</td>
<td>✓ (n = 4)</td>
</tr>
<tr>
<td>Essay raters (n = 8)</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Principals (n = 3)</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Teaching colleagues (n = 5)</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

I have coded each participant with a prefix: Di- for a director, Ti- for a team leader, Ri- for an essay rater, Pi- for a principal, and Ci- for a teaching colleague. The study had 13 male and 12 female participants. Gender differences were not a focus of this study; an analysis of the findings did not show any major differences between male and female raters. Therefore, I have identified all participants as males to protect their identities; anonymity is crucial due to the high-stakes nature of Gaokao.
Directors

The rating centre’s director (Di-nan) and associate director (Di-guan) participated in this study. Both had been at the centre for seven years. In addition to their responsibilities at the rating centre, they were university colleagues. Their responsibilities as directors included hiring and training raters and support staff, coordinating rating procedures, and ensuring rating quality. They provided perspectives regarding administrative support for the central ERA at the centre.

Team Leaders

The seven team leaders came from different workplaces, one from a university and the remainder from different high schools. Four team leaders were from Schools A, B, and C. They were all English teachers with 17 to 33 years of teaching experience and NMET rating experience of four to 23 years at this rating centre. Their experience with NMET rating was similar: They started as raters and were later appointed as team leaders as they demonstrated good essay rating performance and leadership skills. Four of them (Ti-xian, Ti-zhong, Ti-ren, Ti-jian) led essay rating teams; the other three led teams collaborating with essay rating teams—one (Ti-sheng) leading the arbitration team and the other two (Ti-rong, Ti-zhou) leading the quality inspection teams. These team leaders performed multiple duties, including selecting anchor essays and discussing operational rating guidelines, conducting rater training and reorientation within their own team, arbitrating discrepancy ratings, and monitoring the quality of essay ratings. These team leaders provided their perspectives as “middle persons” who interacted with the directors and raters when engaging in the central activity. Also, their extensive experience assuming different roles at the centre allowed them to offer rich information about the central activity of interest. Besides, the four team leaders from the selected schools were also teaching colleagues of raters from their schools, hence they could provide additional perspectives on
interrelated practices in the school context.

**Raters**

The eight essay raters came from the selected high schools. Table 4 shows their demographic information. All eight raters were experienced teachers who had taught English between seven and 25 years. Six raters participated in NMET rating only once; while two raters, Ri-wei and Ri-ling, had more than one year of NMET experience. In Ri-wei’s case, the two years of his NMET rating coincided with both years of this study. These raters, having been trained and supervised by their leaders at the rating centre, provided information about their rating processes and insights as front-line raters who rated NMET essays. At the same time, they also provided perspectives on interrelated practices at their schools.

Table 4

*Information about the Participating Essay Raters*

<table>
<thead>
<tr>
<th>Coded name</th>
<th>School</th>
<th>Teaching experience</th>
<th>NMET rating experience</th>
<th>Participation in this study</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ri-xiu</td>
<td>A</td>
<td>8</td>
<td>1</td>
<td>2015</td>
</tr>
<tr>
<td>Ri-yue</td>
<td>A</td>
<td>20</td>
<td>1</td>
<td>2014</td>
</tr>
<tr>
<td>Ri-xia</td>
<td>A</td>
<td>10</td>
<td>1</td>
<td>2014</td>
</tr>
<tr>
<td>Ri-juan</td>
<td>B</td>
<td>8</td>
<td>1</td>
<td>2014</td>
</tr>
<tr>
<td>Ri-zhen</td>
<td>B</td>
<td>13</td>
<td>1</td>
<td>2014</td>
</tr>
<tr>
<td>Ri-wei</td>
<td>C</td>
<td>21</td>
<td>2</td>
<td>2014, 2015</td>
</tr>
<tr>
<td>Ri-ling</td>
<td>D</td>
<td>25</td>
<td>7</td>
<td>2015</td>
</tr>
<tr>
<td>Ri-dong</td>
<td>D</td>
<td>7</td>
<td>1</td>
<td>2015</td>
</tr>
</tbody>
</table>

**Principals**

Each of the three principals (Pi-li, Pi-xue, Pi-shi) had approximately 30 years of experience teaching Gaokao mandatory subjects: two taught social sciences, and one taught mathematics. One principal, Pi-shi, had prior Gaokao rating experience in the 1980s. These principals provided perspectives regarding administrative support for interrelated practices in the school context.
Teaching Colleagues

Five teachers with experience rating NMET essays between 2013 and 2015 participated in this study. These teachers had taught English between eight and 31 years at Schools A, B, or D. Two of them had two years of prior NMET rating experience. As the raters’ teaching colleagues, they provided perspectives on interrelated practices at the school context, while their recent NMET rating experience also allowed them to share their experiences as front-line raters at the rating centre context.

Materials and Instruments

This study used three types of materials and three data collection instruments. The materials were NMET essay writing tasks, the NMET rating scale, and NMET sample essays. The instruments were TAPs, SRs combined with follow-up interview guide (shortened as SR interview guide), and an individual interview guide.

NMET Essay Writing Tasks

The essay writing tasks for the 2014 and 2015 NMET were used for two purposes: (a) to elicit NMET essay samples from two high schools by NMET candidates who would sit NMET in the following year; and (b) as one of the materials to provide for the eight raters to probe into their decision-making. Both writing tasks required the test-taker to write a letter of approximately 100 words, as an assumed writer named Li Hua, based on a prompt with hypothetical communicative situations. The 2014 task asked the test-taker to read an advertisement about a summer English language learning program from a university in the United Kingdom and to write to the university for more details. The 2015 task prompted the test-taker to write and invite an American friend named Peter to contribute an article about holiday traditions and student life in the United States to the English student newspaper at the test-taker’s
The NMET Rating Scale

The current NMET rating scale was developed by the Ministry of Education and has been in use since 1999 (L. Qi, 2007). No documents were available to the researcher in terms of its scale development and validation processes. The scale includes two sections: major details to be included in essays and five score bands with band descriptors. The major details that an essay should include are laid out, and change every year based on the specific writing tasks. The scale includes five bands with 5 points within each band (25 points total). The band descriptors detail the standards in six aspects: (a) completion of the writing task; (b) covering major details; (c) range of lexical and grammatical resources; (d) accuracy of lexical and grammatical resources; (e) coherence; and (f) effectiveness of communication to the reader.

NMET Sample Essays

NMET sample essays were used to elicit raters’ TAPs and SRs, and also for rater training purposes. Due to practical constraints, essays used for the two rounds of data collection (2014 and 2015) were collected in slightly different ways. In Round One, I collected 21 2014 NMET sample essays written in class by NMET candidates (Grade 12 students) from one school in November 2014. Candidates were asked to complete an essay in class within 25 minutes (as the estimated time limit for authentic NMET). To ensure that candidates would try their best, they were told that their essays would be marked by expert raters to provide an estimation of their NMET writing proficiency. In Round Two, I collected NMET sample essays a few days after the 2015 NMET administration under testing conditions. I collected 64 essays written by candidates from another school whose common practice was to have Grade 11 students (i.e., NMET 2016 candidates) write that same test as part of Gaokao mock tests shortly after the official
Both rounds of sample essays were later marked blindly by two expert NMET raters. Expert scores for Round One essays ranged between 13 and 24 out of 25 points, with 18 of 21 essays at either Bands 3 or 4 on the NMET rating scale. Expert scores for Round Two essays ranged between 5 and 23 points, with 52 of 64 essays also at Bands 3 and 4. This variation suggested a comparable level of proficiency among students from these two schools. The sample essays used in the two rounds were selected with different criteria. In Round One, eight essays that demonstrated the biggest score differences between the two experts’ ratings were selected as sample essays. Being limited by the small score range in the collected essays, I hoped to explore the extent of variations in raters’ responses to the essays. In contrast, with a wider score range to choose the essays, the eight sample essays collected in Round Two were selected based on two criteria: (a) four with maximum and four with minimum rating discrepancy between the two expert raters; and (b) proficiency levels as varied as possible. By doing so, I hoped that the samples selected were more representative of different situations raters encountered in the operational rating session. Thus 16 essays (coded with the collecting year followed by original series number) were used in this study: Eight essays (Essays 14-01, 14-03, 14-04, 14-06, 14-08, 14-10, 14-12, 14-19) were selected for Round One, and another eight essays (Essays 15-18, 15-21, 15-35, 15-40, 15-53, 15-57, 15-63, 15-64) were selected for Round Two. These selected essays were later used to elicit raters’ TAPs and SRs. At the same time, three essays with minimum discrepancy at different bands were selected for each round, used for training purposes.

**Think-Aloud Training Guide**

My data collection was guided by think-aloud training guidelines and instructions for
producing TAPs while rating essays (see Appendix C). Guidelines and instructions were adapted from Barkaoui (2008).

The training guidelines restated the purpose of the study and explained the think-aloud protocol, its intended use in this study, and its producing procedures. The guidelines provided two math problems for raters to practice and become familiar with TAPs. One was a basic operational math question, based on Ericsson and Simon (1993). Considering essay rating involves assessing language, the second one was a problem-solving math question, based on Mackey and Gass (2005), where the math question was presented in descriptive language, a task closer to the essay rating task.

The instructions for thinking aloud were designed for raters. The instructions repeated the intended use of TAPs in this study, followed by 13 detailed instructions of how to produce TAPs while assessing essays.

**Stimulated Recall Interview Guide**

I designed an SR interview guide with two sections: guidelines and instructions for raters conducting SRs, and follow-up questions (see Appendix D). SR guidelines and instructions were based on procedures recommended by Gass and Mackey (2000), explaining the purpose of SR and its procedures. There were five follow-up questions. Question 1 asked about raters’ perceptions of the essays they just rated, underlying purposes when rating those essays, their rating priorities, possible influences on their ratings, and the reasons behind their perceptions and behaviours. Question 2 asked about raters’ perceived differences between rating the eight sample essays and rating authentic essays in the operational session. The first two questions were designed to explore reasons behind raters’ decision-making in the operational session. Questions 3 and 4 both included subquestions about raters’ perceived effects of two instruments, TAPs and
SRs, on their rating. The questions for TAPs were based on Barkaoui (2008, p. 260), and the questions for SRs based on Gass and Mackey (2000). The guide ended with a concluding question inviting raters to provide additional comments.

**Individual Interview Guide**

I used a semistructured individual interview guide with 14 questions organized into three sections to examine participants’ perspectives regarding the sociocultural context of NMET ERA in this study (see Appendix E). Section One (Question 1) contained five subquestions exploring perceptions about the broader sociocultural context by all participants. The first four subquestions were designed based on preliminary thematic analysis of public documents collected from news articles on Gaokao rating; and the fifth included slogans from hanging banners at the rating venue and three popular ones retrieved from news articles. Sections Two (Questions 2 to 8) and Three (Questions 9 to 14) were used to investigate participants’ perspectives regarding the sociocultural contexts at the rating centre and the high schools, respectively. These questions considered elements that would potentially contribute to components (e.g., mediators, objects/motives) in the relevant activities and practices at the two settings. More specifically, I asked questions about the following aspects: (a) participants’ self-perceived roles and responsibilities when involved in corresponding activities; (b) motives for supporting or engaging in Gaokao rating; (c) perceptions of the sociocultural contexts at the activity-related settings; (d) relationships with other members in the same communities; and (e) use of documents and artifacts. Regarding the central ERA at the rating centre, I asked two extra questions about (f) participants’ efforts for improving rating practices, and (g) the challenges they encountered. The concluding question invited participants to provide additional information.
Data Collection

Due to practical constraints and the time-sensitive nature of the NMET administration and subsequent rating, the data collection fieldwork was conducted in two rounds, the first in December 2014 (six months after the 2014 NMET rating) and the second in June and July 2015 (immediately before and after the 2015 NMET rating). These two rounds gave me prolonged exposure to the research phenomenon and a better understanding of the cultural and historical context of NMET ERA in this study. I was subsequently able to refine and adapt my research design in the second round, obtaining a more in-depth understanding of the research phenomenon. This was consonant with the emergent, flexible, and responsive nature of qualitative research design (Patton, 2015). My data collection continued until I reached theoretical saturation, which Strauss and Corbin (1990) stated occurs when “(a) no new or relevant data seem to emerge regarding a category; (b) the category development is dense; and (c) the relationships between categories are well established and validated” (p. 188).

Before data collection, I translated the TAP instructions, the SR interview guide, and the individual interview guide into Chinese. Given that all of my participants spoke Chinese as their first language, interviewing and communicating with participants in Chinese would make conversations more natural to them and easier to understand. A back translation from Chinese to English was later conducted by another researcher to ensure the validity of translated instruments. Then I piloted the three translated instruments and made revisions as necessary to ensure they fully captured my intentions. The same procedures were used for the second round of data collection. All TAPs, SRs, and interviews were conducted in participants’ offices or empty classrooms, and audio-recorded with participants’ permission. A letter of information and a consent form (see Appendices B.1-B.4) were provided to every participant prior to data
collection. I then met individually with each volunteer to ensure that their concerns and questions were answered, before obtaining their informed consent and arranging a schedule for their participation in this study.

The data collection timeline is summarized in Table 5. As it shows, I collected the four sources of data (i.e., documents, interviews, TAPs, and SRs) in both rounds. The total number of interviews was more than 25, because some participants were interviewed in both rounds. The total number of essay raters who provided TAPs and SRs was 9 because one rater, Ri-wei, participated in both rounds. To avoid repetition, in the following subsections I explain these procedures by data source. The procedures are reported according to the relationship between the data sources and the three research questions (from cognitive to social layers): first TAPs, then SR interviews (SRs with follow-up interviews), individual interviews, and documents.

Table 5

*Data Collection Timeline*

<table>
<thead>
<tr>
<th>Study Round</th>
<th>Timeline</th>
<th>Procedures</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Prior to December 2014</td>
<td>Analysis of documents (2013-2014 news articles)</td>
</tr>
</tbody>
</table>
| Round One   | December 2014 | Individual interviews outside rating centre and at Schools A, B, C (n = 18)  
|             |           | 2014 NMET sample essays  
|             |           | TAPs (n = 5)  
|             |           | SRs and follow-up interviews (n = 5)  
|             |           | Analysis of documents (about the two activity-related settings) |
| Round Two   | June-July 2015 | Individual interviews outside rating centre and at Schools A, C, D (n = 14)  
|             |           | 2015 NMET sample essays  
|             |           | TAPs (n = 4)  
|             |           | SRs and follow-up interviews (n = 4)  
|             |           | Analysis of documents (about the two activity-related settings) |
|             | After August 2015 | Analysis of documents (2015 news articles) |
Think-Aloud Protocols

The TAPs were collected after I interviewed raters about their NMET rating experiences, for the purpose of establishing rapport with raters before producing TAPs and learning about their prior experiences that were related to or mediating their verbalizing actions.

Raters were trained individually prior to collecting TAPs and the training process was slightly different in each round. In Round One (2014), I first showed them PowerPoint slides (based on documents collected from two participants) with photos of their NMET rating venue and slogans hanging in the venue, a description of the NMET rating centre structure, and a summary of NMET rating procedures. I then followed similar rater training procedures as in operational sessions: I presented each participant with the operational rating guidelines and the three selected training essays for them to mark, compare (with scores assigned by expert raters), and discuss (in cases of discrepancy beyond the acceptable level). The purpose of the two-part session was to evoke their memory about the operational rating session and help them stay close to that condition as far as possible. In Round Two (2015), the slide presentation was omitted, since they produced TAPs within two weeks after the operational rating session.

After rater training, I provided each rater with TAP training in Chinese, before asking them to produce TAPs. Guided by the TAP training guidelines and instructions, I gave each rater detailed instructions and careful training on how to think aloud while rating essays. Following Barkaoui (2008), several procedures were adopted to enhance the quality of think-aloud data. During TAP training, I asked raters to “act and talk as if you are talking to yourself,” and emphasized the importance of performing the rating task in the way they did in the operational NMET rating session, and not as if they were providing verbal reports; verbalizing all the thoughts they had while rating essays; and being honest about their rating processes and
reactions in a natural manner. I also made it clear to raters that their rating performance would not be evaluated. I presented each rater with a Chinese version of the think-aloud instructions for them to read before they started to rate the sample essays while verbalizing their thoughts.

To minimize researcher effects on raters’ performance, raters conducted their TAPs individually in a quiet space with sufficient time to verbalize. They produced TAPs in Chinese, except for when reading the essays. Rater training and TAP training lasted about 30 minutes. Most TAPs lasted about 18 minutes, while one participant’s TAP took 45 minutes. Approximately 3.3 hours of TAPs were produced by the eight participants.

**Stimulated Recall Interviews**

Immediately after raters produced TAPs individually, at the same location I asked raters to provide SRs, using the same eight sample essays as stimuli, followed by interviews about their perceptions and underlying purposes behind their thought processes. I conducted one rater’s SR and follow-up interview the day after his TAP due to scheduling availability.

Before asking raters to produce SRs, I explained the technique and procedures and gave them careful instructions about how to produce SRs. I then presented each rater with the eight essays they rated earlier, asking them to go through each one and discuss their scores. I adopted slightly different procedures when eliciting SRs in the two rounds. In Round One, raters were asked to recall their thinking processes, using the essays alone as stimulus. Several raters reported that such “TAPs + SRs” practices reminded them of similar rate-rerate practices in the operational session that were aimed to check raters’ self-consistency. However, more raters felt it was repetitive to verbalize their thought processes for the same essays again within such short time. Based on this feedback, I revised the SR eliciting procedure for Round Two. In addition to presenting essays, I played raters’ TAP recordings while they read the essays and listened to their
recordings. Whenever there was more than three seconds of silence in the recordings, I would ask raters what they were thinking/doing at the moment. I asked raters to comment on anything relevant that came to their minds at any stage while they went through each essay. After producing SRs for all eight essays, I asked raters follow-up questions about their perceptions, reasons, and underlying purposes for rating each essay.

Following Gass and Mackey (2000), I adopted procedures to enhance the quality of SR data. I gave the raters minimal training on how to produce SRs; I also asked the raters general questions like “what were you thinking at that time” during the SRs and if the rater did not remember, I did not pursue further to avoid fishing for answers. Finally, I did not ask them questions about their perceptions until all their SRs were completed. To assess the quality of collected data, I also asked the raters about perceived effects of the instruments (TAPs and SRs) on their thought processes. The duration of SR interviews varied between 30 and 90 minutes. A total number of about 8.2 hours of SR interviews were produced by the eight rater participants.

**Individual Interviews**

I conducted individual interviews before collecting TAPs and SRs. These interviews were conducted at multiple research sites in both rounds. Due to feasibility issues and security requirements, interviews were conducted outside the rating centre. Participants were sometimes interviewed more than once (some as many as four times) for availability reasons, as well as to explore their perceptions in more depth.

In Round One, I visited three research sites (Schools A, B, C) and two other places where participants worked to conduct interviews. Participants in this round included 1 director, 5 team leaders, 5 raters, 3 principals, and 4 teaching colleagues \((n = 18)\). I used the interview guide to ask participants about their perceptions about the sociocultural context of Gaokao rating in
China, and their experiences with respect to Gaokao rating at the rating centre when they participated in NMET rating and at their schools. Based on preliminary analysis of data collected in Round One, I made minor revisions to the interview guide and added some more specific interview questions. The revised instrument was later used in Round Two to follow up with some participants in Round One.

In Round Two, interviews were all conducted before and after the 9-day operational rating sessions. To conduct interviews, I visited one new research site (School D) and two other participants’ workplaces, and revisited two research sites (Schools A and C) where essay raters worked. I interviewed 14 participants in this round: 2 directors, 4 team leaders, 6 raters, and 2 teaching colleagues. I collected a total of approximately 33 hours of interviews in the two rounds, each interview with varied duration between 0.2 and 1.3 hours.

Documents

As shown in Table 5 of the data collection timeline, various documents were collected before, during, and after the two rounds of data collection to examine the sociocultural context of Gaokao rating. I collected public and personal documents in the past three years (2013-2015) for this study (summarized in Table 6).

Table 6

Summary of Collected Documents

<table>
<thead>
<tr>
<th>Type</th>
<th>Document</th>
<th>2013</th>
<th>2014</th>
<th>2015</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Public</td>
<td>Gaokao rating news report</td>
<td>114</td>
<td>81</td>
<td>35</td>
<td>230</td>
</tr>
<tr>
<td></td>
<td>Rating centre slogan</td>
<td>0</td>
<td>6</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>NMET test syllabus</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Government-issued Excellent Gaokao Rater</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Award certificate and document</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Personal</td>
<td>Work note</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>Summary presentation slide</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Reflective writing</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
Gaokao rating news reports were collected before and after the two rounds of data collection. The news reports were selected because, as one type of public documents, they provide rich information about the popular social issues and phenomena, and are a useful data source to provide a broader view of the sociocultural context of Gaokao rating in Chinese society. The reports collected prior to Round One were also analyzed and used to inform the interview questions. During the two rounds, I collected both public and personal documents about Gaokao rating at the two activity-related settings. The public documents included the rating centre’s slogans, NMET test syllabus, and government-issued Excellent Gaokao Rater Award certificates and documents given to awardees and their schools in acknowledgement of their excellent rating performance. I did not collect rating centre slogans and NMET test syllabus in 2015 because they did not change. Personal documents were provided by some team leader participants about their NMET rating work experiences, including work notes, summary presentation slides, and one piece of reflective writing. All of these collected documents potentially informed or reflected participants’ perceptions and practices in the two settings. In brief, these documents provided important archival evidence on the broader and immediate sociocultural contexts of NMET rating. All documents, with the exception of the news reports, were provided by participants.

I collected news reports from three online sources: Xinhua, People’s Daily, and Sina.com.cn. Xinhua and People’s Daily were selected because they are news agencies accountable to the central government and their reports have considerable influences on Chinese society. I searched and retrieved their news from Factiva database through Queen’s University Library portal. Sina.com.cn was selected to complement Xinhua and People’s Daily since it is one of the most popular and comprehensive news websites in China, aggregating news from
reliable local newspapers across the country in addition to the central government’s news agencies. I was able to access a comprehensive list of relevant news articles and multimedia news reports. A second reason for selecting Sina.com.cn is its ability to limit my search based on time constraints, despite having a search engine not as well-developed as Factiva’s. These three sources covered the most important news reports in Chinese daily life and yet are accessible to the public. I narrowed down the search period to the months of June, July and August in 2013, 2014, and 2015. These three months represents the annual period when Gaokao catches the country’s attention, from administration and rating in early June, the release of results in late June, and admissions decisions made by universities by August.

I started the news search with the term “高考 (Gaokao)” in Chinese to form a general idea of this topic. The NMET is literally “Gaokao English” in Chinese; it was entailed by the search term “Gaokao”. I searched headlines and lead paragraphs in Factiva and headlines in the Sina.com.cn search engine⁴. A total of 11,843 news reports about “Gaokao” were retrieved from the three sources in this initial search. The number of search results offered a glimpse of how popular this topic was and how much attention Gaokao attracted across the country during Gaokao seasons in the three years. Considering the wide range of documents related to Gaokao, I narrowed down my search to Gaokao rating, using the search terms “高考阅卷” and “高考评卷”, two most commonly used terms referring to “Gaokao rating” in Chinese. A total of 230 news reports (including 163 articles and 67 videos) about Gaokao rating were retrieved and included (see Table 7), all published in June during the Gaokao rating period of a given year.

⁴ The search engine of Sina.com.cn did not allow search in lead paragraphs; so I only searched in headlines.
Table 7

Search Results of “Gaokao Rating” News Reports

<table>
<thead>
<tr>
<th></th>
<th>2013</th>
<th>2014</th>
<th>2015</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Xinhua</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>People’s Daily</td>
<td>0</td>
<td>3</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>Sina.com</td>
<td>114</td>
<td>78</td>
<td>32</td>
<td>224</td>
</tr>
<tr>
<td>Total</td>
<td>114</td>
<td>81</td>
<td>35</td>
<td>230</td>
</tr>
</tbody>
</table>

Data Analysis and Integration

All audio data were transcribed in full. All transcripts and collected documents were imported into MAXQDA 11, a qualitative data management and analysis software. The coding was conducted by hand in English.

Given the inductive and comparative nature of qualitative research, the open and axial coding techniques in the constant comparison method (CCM; Strauss & Corbin, 1990) guided my data analysis. These coding techniques use two analytic procedures to facilitate data conceptualization. One is to ask questions about data; the other is to make comparisons for similarities and differences between instances. Asking questions like “who” “when” “where” “what” “how” “how much” and “why” can open up the data, thinking of potential categories and hypothesizing relationships between categories and subcategories, to help label phenomena of interest and categorize them. Based on Saldaña’s (2009) recommendations, I used the following questions to guide data coding: (a) How exactly do participants rate essays? What specific strategies do they use? (b) How do participants talk about, characterize, and understand what they have done and what has happened? (c) What assumptions are they making? (d) What do I see going on here? What do I learn from these documents and transcripts? (e) What strikes me? Making comparisons can help to discover emerging themes and patterns and make sense of the data. Following Boeije’s (2002) recommendations of a purposeful approach to the CCM, I made
five types of comparison to obtain a thorough, rich understanding of the collected data: (a) comparison within a single transcript and document (e.g., one rater’s TAP, one interview, one news report); (b) comparison of transcripts and documents within the same participant group (e.g., all the raters’ TAPs, all the principals’ interviews, all the work notes provided by team leaders); (c) comparison of transcripts and documents across participant groups (e.g., interviews with different groups of participants); (d) comparison of transcripts and documents by participants sharing the same activity-related setting (e.g., interviews with participants involved in the activity in the rating centre context, and with participants in the high school context); and (e) comparison across the four data sources. In open and axial coding, these five types of comparison helped me develop categories, identify similarities and differences within and among categories, link categories and subcategories, and relate the core category to other categories.

Given the multiple sources of data, the open and axial coding techniques were used in a two-stage process:

1. Open coding: analysis within each data set (i.e., TAPs, SRs, interviews, documents);
2. Axial coding: analysis across the data sets.

In Stage One, I analyzed the four data sets independently, exploring all possibilities within each data set. In Stage Two, I compared the results across the four data sets, and continued to analyze them in a comparative and inductive way. The analysis results are reported in Chapter 4. Following Stage Two, my data integration was based on axial coding findings while conceptually informed by the CHAT framework, the findings of which are reported in Chapter 5.

**Open Coding**

Open coding uses the two analytic procedures to further facilitate data conceptualization and categorization, first by asking questions and then making comparisons to examine the data
and explore possible meanings in the data. During open coding, data were coded separately and in slightly different ways to accommodate their different uses in this study.

**TAP & SR data.** TAP and SR data were both verbal protocols about raters’ cognitive processes, and were thus analyzed using the same strategy. Open coding was conceptually informed by Cumming, Kantor, and Powers’s (2002) descriptive framework of rater decision-making behaviours. Based on Cumming et al.’s framework, I made some changes and generated a coding scheme including 27 codes (see Figure 13).

<table>
<thead>
<tr>
<th><strong>Self-Monitoring Focus</strong></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Interpretation Strategies</strong></td>
<td>13. Assess relevance to central topic</td>
</tr>
<tr>
<td>1. (Re)Read or translate composition</td>
<td>14. Assess task completion</td>
</tr>
<tr>
<td>2. Envision personal situation of the writer</td>
<td>15. Assess coherence and cohesion</td>
</tr>
<tr>
<td>3. Scan whole composition</td>
<td>16. Assess text organization</td>
</tr>
<tr>
<td>4. Refer to or interpret rating scale</td>
<td></td>
</tr>
<tr>
<td><strong>Judgment Strategies</strong></td>
<td></td>
</tr>
<tr>
<td>5. Decide on macro-strategy for reading and rating</td>
<td></td>
</tr>
<tr>
<td>6. Compare with other essays/test-takers</td>
<td></td>
</tr>
<tr>
<td>7. Define own rating criteria</td>
<td></td>
</tr>
<tr>
<td>8. Articulate general impression</td>
<td></td>
</tr>
<tr>
<td>9. Summarize, distinguish, or tally judgments collectively</td>
<td></td>
</tr>
<tr>
<td>10. Articulate or revise scoring decision</td>
<td></td>
</tr>
<tr>
<td><strong>Rhetorical and Ideational Focus</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Interpretation Strategies</strong></td>
<td>17. Observe layout (e.g., paragraph, text length)</td>
</tr>
<tr>
<td>11. Interpret ambiguous or unclear phrases</td>
<td>18. Consider legibility</td>
</tr>
<tr>
<td><strong>Judgment Strategies</strong></td>
<td>19. Classify errors into types</td>
</tr>
<tr>
<td>12. Assess major details</td>
<td>20. Edit phrases for interpretation</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Language Focus</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Interpretation Strategies</strong></td>
<td>21. Consider syntax or morphology</td>
</tr>
<tr>
<td>17. Observe layout (e.g., paragraph, text length)</td>
<td>22. Consider lexis</td>
</tr>
<tr>
<td>18. Consider legibility</td>
<td>23. Consider frequency and gravity of error</td>
</tr>
<tr>
<td>19. Classify errors into types</td>
<td>24. Consider Liangdian (advanced-level lexical and syntactic resources)</td>
</tr>
<tr>
<td>20. Edit phrases for interpretation</td>
<td>25. Assess comprehensibility, fluency, and naturalness</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Judgment Strategies</strong></td>
<td>26. Consider handwriting and answer sheet tidiness</td>
</tr>
<tr>
<td>21. Consider syntax or morphology</td>
<td>27. Consider spelling and punctuation</td>
</tr>
<tr>
<td>22. Consider lexis</td>
<td></td>
</tr>
<tr>
<td>23. Consider frequency and gravity of error</td>
<td></td>
</tr>
<tr>
<td>24. Consider Liangdian (advanced-level lexical and syntactic resources)</td>
<td></td>
</tr>
<tr>
<td>25. Assess comprehensibility, fluency, and naturalness</td>
<td></td>
</tr>
<tr>
<td>26. Consider handwriting and answer sheet tidiness</td>
<td></td>
</tr>
<tr>
<td>27. Consider spelling and punctuation</td>
<td></td>
</tr>
</tbody>
</table>

*Figure 13. Coding scheme for TAPs and SRs*

To arrive at this coding scheme I first kept the codes I thought were relevant to my data and removed the ones that were not applicable. I then made revisions to the following codes:
1) Renamed the code *Assess quantity of total written production* first to an *in vivo* code *Observe text length*; later merged this code into *Observe layout* under Language Focus and Interpretation Strategies, as this behaviour often occurred when raters were making comments about the essay layout.

2) Renamed the code *Consider use and understanding of source material* to an *in vivo* code *Assess relevance to central topic*.

3) Renamed the code *Read or reread composition* to *(Re) Read or translate composition*, because many rater participants translated essays into Chinese when they were reading and interpreting those essays.

4) Broke down several long codes into shorter ones. For example, *Decide on macrostrategy for reading and rating; compare with other compositions; or summarize, distinguish, or tally judgments collectively* was a stand-alone code in Cumming et al.’s framework. I broke it down into three codes to describe my data more clearly.

Finally, I added five new codes:

1) To account for raters’ use of the rating scale, a code *Refer to or interpret rating scale* was added under Self-Monitoring Focus and Interpretation Strategies, following Barkaoui (2008).

2) Added an *in vivo* code *Assess major details* under Rhetorical and Ideational Focus and Judgment Strategies, because all the rater participants frequently commented on whether or not essays covered the major details as required by the writing task.

3) Added a code *Consider legibility* under Language Focus and Interpretation Strategies to account for raters’ comments on whether or not test-takers’ handwriting created reading difficulties.
4) Added two *in vivo* codes *Consider Liangdian* and *Consider handwriting and answer sheet tidiness* under Language Focus and Judgment Strategies. The former was to account for some raters’ comments on whether or not an essay contained Liangdian (shining points), which referred to advanced-level lexical and syntactic resources test-takers used. The latter was to account for raters’ comments on whether or not test-takers’ handwriting was proper and their answer sheets looked neat.

A list of the codes with examples from this study are presented in Appendix I.

**Interview data.** Interview data (including follow-up interview data after SRs) were analyzed using the open coding technique. I analytically examined interview transcripts by asking questions and making comparisons within one transcript, of transcripts within the same participant group, across different participant groups, and by participants sharing the same activity-related settings. This coding process involved using methods such as initial coding, structural coding, descriptive coding, *in vivo* coding, and values coding (Saldaña, 2009). Initial coding involved breaking down the data into discrete segments to examine and compare for similarities and differences. Structural coding involved matching interview questions to the study’s research questions. Descriptive coding involved summarizing the basic topic of a passage in the transcript using a word or short phrase. *In vivo* coding involved using participants’ generated words to reflect their voices. Values coding was used to reflect participants’ values, attitudes, and beliefs, to represent their perspectives.

**Document data.** My analysis of document data focused on the content of documents. Content analysis is a strategy attending to documents as a container of information for understanding some aspect of social practice and meaning (Miller & Alvarado, 2005). I conducted thematic analysis to analyze the content of the document data.
Axial Coding

Axial coding involves constant interplay between inductive and deductive thinking. This coding technique also uses the analytic procedures of asking questions and making comparisons, but is focused on discovering relationships between categories and relating subcategories to a category. During axial coding, I compared the open coding results of each data set and asked questions to identify the relationships between and across their categories and subcategories, to link, merge, and develop the categories. Following Strauss and Corbin’s (1990) coding paradigm, I asked questions about the following relationships: (a) conditions, which refer to the events that lead to or influence the occurrence of a phenomenon of interest; (b) context, which represents the set of conditions within which the action/interaction strategies related to a phenomenon take place; (c) action/interactional strategies, which describe how participants manage, carry out, and respond to a phenomenon; and (d) consequences, which refer to outcomes of action/interaction. To facilitate this axial coding process, I further analyzed the documents using a context analysis strategy. Context analysis is a strategy that approaches documents as elements in the larger field of social activity, with socially situated meanings (Miller & Alvarado, 2005). Following Miller and Alvarado, I worked on document data, complemented by other data sources, to examine the intentions and meanings indicated in the documents, and the roles, uses, and functions these documents played in the NMET ERA in this study, with the purpose of helping to identify the relationships between the categories.

Data Integration

After my data analysis, I integrated the findings generated from the axial coding by using the CHAT framework. I mapped my findings onto different components of the two interacting layers in the CHAT framework: the cognitive-layer data (i.e., TAPs, SRs, interviews) to describe
raters’ goal-directed actions; and the social-layer data (i.e., interviews and documents) to describe the collective object-oriented activity. This collective activity, in the form of an activity system, was made up of the central activity and interrelated practices at the two settings (rating centre and high schools), respectively. I then combined the findings about the two layers into a visualized set of CHAT triangles, and drew on the findings to identify systemic tensions within the activity system and interactions between the central activity and its relevant practices in the two settings. This data integration enabled me to generate a comprehensive picture of the relationships between the cognitive and social layers of the NMET ERA, and between the central activity and its interconnected practices in the two settings.

Enhancing Validity and Reliability of the Study

Validity and reliability in qualitative research concern the trustworthiness and rigor of the study (Merriam, 2009). According to Merriam, validity refers to the congruence of research findings with the holistic, multidimensional reality, while reliability refers to whether the results are consistent with the data collected.

Beyond the data collection, analysis, and integration procedures documented earlier, I employed strategies to further enhance the validity and reliability of this case study. I kept the chain of evidence (Yin, 2009), or the audit trail, which enabled me to identify and track sources of evidence and data leading to a particular conclusion or statement. To identify my chain of evidence, I developed a digital database for this study to facilitate the organization and storage of my collected data. A field journal in verbal (during data collection) and written forms was also kept throughout the entire process (prior to data collection through to the final draft), to document my thoughts and interpretations of the findings. I used memoing throughout the coding process to track changes in my interpretations of the data. In addition to the chain of evidence, I
conducted member checking after data analysis and integration, by summarizing and translating my major findings and the CHAT analysis into Chinese, and sharing the results with my participants for feedback, to ensure my interpretations captured their perspectives. The supervisory committee reviewed my coding scheme, interpretations, and conclusions.

In addition to the above strategies often used in qualitative research, I also examined the effects of two data collection instruments, in terms of the methods themselves and the data collection time, to enhance the validity and reliability of this study. The method effects of the instruments (TAPs and SRs) on the collected verbal protocol data were examined by interviewing participants about the perceived influences of these procedures on their rating performance. The effects of TAPs were examined by asking participants about differences in their rating processes and results between producing TAPs in this study and in the operational NMET rating sessions. One effect on rating behaviour admitted by almost all the raters was that producing TAPs made them rate at a lower speed and consequently read essays more closely than in the operational rating sessions. There were other individual variations in terms of admitted TAP influences, such as assigning lower essay scores by one or two points because raters were reading more closely; difficulty verbalizing their thoughts, as speech could not fully capture their thinking; disruption to raters’ thinking processes due to verbalization; and influences of conducting TAPs on raters’ rating behaviours and sequences. For example, one rater reported that he translated the essay to facilitate his reading and verbalization, and he also delayed his score band range decision until he finished reading the entire essay, while he rarely did so in operational sessions. The other rater reported that the essays used in TAPs were paper-based which made him pay less attention to handwriting and answer sheet tidiness, compared with the OSM system used in operational sessions. He explained that reading essays onscreen made the
effect of student handwriting and tidiness of answer sheets more noticeable. These TAP method effects were consistent with the limitations of reactivity and veridicality of TAPs identified in previous literature (e.g., Barkaoui, 2011b; Lumley, 2005; Smagorinsky, 2001; Swain, 2006). The interactive and social nature of TAPs was also identified in this study. Two raters felt that verbalizing their thoughts reminded them of their teaching practice where they read and commented on student essays in front of the students. Another rater expressed his awareness that the researcher was going to listen to his TAP recordings. Such interaction confirmed the socially situated nature of TAPs (Barkaoui, 2011b; Smagorinsky, 2001; Swain, 2006).

The effects of SRs admitted by all raters included (a) allowing them to elaborate on what they could not verbalize during TAPs; and (b) making them read more closely and pay attention to more details in the essay, hence capturing their thoughts about what they did not notice when producing TAPs. As a result of the second effect, two raters lowered three essay scores by three or four points compared to what they had assigned earlier during TAPs. In addition, when producing SRs, due to the researcher’s presence and the researcher’s act of asking questions, some raters (c) showed more interaction with the researcher, such as one rater asking the researcher questions while he was recalling. The first effect confirmed that SRs were useful in supplementing TAPs for capturing raters’ thoughts; while the second and third effects suggested raters’ interaction with the researcher’s research activity, and the limitation of SRs in filtering raters’ thoughts and reporting what they thought the researcher wanted them to say.

These identified method effects showed the limitations of using verbal protocols (TAPs and SRs) in probing in raters’ thought processes. However, when accompanied with information about participants’ perceived differences and influences, these methods helped me to infer raters’ thought processes in the operational NMET rating sessions, which would otherwise be difficult.
to replicate under research conditions. In this sense, asking about the perceived influences of SRs and TAPs enhanced the validity and reliability of this study, and supported the use of these two instruments to better serve my research purpose.

Besides method effects, time lag effects on the TAPs collected in the first round were also examined by comparing the lengths of their transcripts and corresponding codes with those collected in the second round (since the first round data were collected six months after the operational session). No noticeable differences were identified regarding those two aspects. This can be explained by the fact that raters reported considering the 9-day intensive NMET rating experience as professional “training”, after which the NMET rating criteria were internalized and applied in their everyday teaching and school rating practices. Such follow-up practices may have enhanced the training effect, hence explaining why after six months raters did not demonstrate substantive differences regarding their rating focuses and sequences.
Chapter 4 Research Findings

Guided by the cultural-historical activity theory (CHAT) framework (Engeström, 1987, 2001), this study examined essay rating activity (ERA) situated within the sociocultural context of Gaokao. This situated practice included both the cognitive and social layers of the National Matriculation English Test (NMET) ERA. In this chapter, I organized the results in two sections.

In Section 1, I present findings about the cognitive layer (individual’s goal-directed actions) of the NMET ERA, based on TAP, SR, and interview data. I describe eight raters’ goal-directed decision-making actions when rating NMET essays. The findings are grouped in three themes: rating focuses, rating sequence, and influential factors. Given that this study focused on a small sample of raters and a limited number of essays to elicit raters’ thinking processes, it became important to consider these raters’ goal-directed decision-making actions in the sociocultural context where these actions took place. Moreover, it needs to be pointed out that although the influential factors are reported separately in this section, they cannot be fully understood without considering the sociocultural context. I first report these factors in the section on the cognitive layer, because I see they relate the cognitive to the social layers. Most of these identified factors were observed by some previous cognitive-approach studies. Including them in my results of the cognitive layer may make it easier for the reader to understand, as the three themes of the cognitive layer are very similar to findings from a cognitive perspective. Reporting these factors first in the cognitive layer and then situating them in the social layer can show how including a social layer can help to better understand these factors. Later, when I describe and explain the social layer and how these two layers connect, the difference between the cognitive and CHAT approaches to essay rating research becomes more obvious, and thus easier to understand the value of examining the social layer of the ERA.
In Section 2, I present findings from the interviews and documents to describe the social layer (collective object-oriented activity) of the NMET ERA. After establishing the background of Gaokao rating, I depict the sociocultural context of the activity at the two settings (rating centre and school) from participants’ perspectives, which are organized in three themes: (a) “sacred mission with grave responsibility”, (b) “stressful”, and (c) “beneficial”. These identified themes and their related categories not only contribute to the understanding of the sociocultural context of NMET rating, but also provide a more situated understanding of the influential factors observed on the cognitive layer.

Section 1: Raters’ Goal-Directed Decision-Making Actions

This section is organized around three emerging themes: rating focuses, rating sequence, and factors influencing decision-making. These themes were based on TAP and SR data while inferences on raters’ decision-making actions in the operational NMET rating sessions were identified by drawing on raters’ interview data.

Rating Focuses

Raters attended most frequently to three aspects of writing when rating NMET sample essays: (a) content coverage, (b) language quality, and (c) handwriting and answer sheet tidiness. Translated quotations from participants used the following transcription conventions of verbal protocols: Text that is underlined indicates a direct reading from or is an interpretation of the rating scale, while text in UPPERCASE indicates direct reading from an essay (see Appendix F for a full list of transcription conventions).

Content coverage refers to raters’ notions of the major details and relevant content that an essay should include. Given that the NMET writing task is guided by a prompted communicative situation, the major details are based on requirements in the writing prompt and are evaluated in
the rating scale. For instance, the major details for the 2014 task were indicated in the prompt and stated in the rating scale, including expressing interest in a six-week program, asking about details such as the start date, weekly class hours, class size, fees, and accommodation, and ending with an appropriate sentence. The major details for the 2015 task included introducing the theme of the newspaper section, explaining aspects to be covered in the article, text length, and an expected due date, and concluding with an appropriate sentence. Here are two examples from the 2014 TAPs:

“The first major detail that needs to be mentioned in the first paragraph is **what you want to do when you are there**. Then [the second major detail] “**to take a six-week course**”, em, well, [is] present in the second paragraph.” (Ri-juan, attending to two details)

“Later, in the second paragraph when [the student] wrote the major details, he rambled about this and that, going off the topic.” (Ri-zhen, attending to content relevance)

*Language quality* consisted of various linguistic features, including (a) syntax or morphology; (b) lexis; (c) frequency and gravity of errors; (d) comprehensibility, fluency, and naturalness; (e) cohesion and coherence; and (f) spelling and punctuation. Here are a few examples of how raters were attending to some of these features. This is when Ri-yue was attending to lexis used in one essay: “**WHEN I SURFED...I FOUND...WHICH MADE ME EXCITED, . . .DECIDED TO PARTICIPATE. . . “PARTICIPATE” this [word] should be counted as advanced vocabulary.” Here is another example when Ri-xia was assessing frequency and gravity of errors included in an essay: “He [the student] made too many errors, including several grave errors, as well as several minor errors in details.” When Ri-zhen was assessing fluency of a sentence in one essay, he commented, “Let me see, **BESIDES, . . .GET ALONG WITH OTHERS TO IMPROVE OUR ENGLISH...HOW MUCH I HAVE TO PAY.** This sentence reads
Handwriting and answer sheet tidiness refers to legibility and neatness of students’ written work and completed answer sheets. One rater, Ri-yue, considered handwriting legibility of one student: “The words were written too small; [I] cannot recognize whether he wrote HOW or something else, is it HO or HOW? I could not identify the letter W.” Ri-wei also looked at handwriting: “The handwriting of this essay is neat, making me feel very refreshed.”

To present a profile of those writing aspects that specifically drew raters’ attention, occurrence counts of the three aspects and their related subcategories are presented in Table 8. Overall, three trends arose: (a) content coverage, different features regarding language quality, handwriting and answer sheet tidiness were all frequently considered when raters were reading essays and making their scoring decisions; (b) raters largely paid attention to language quality, and least about handwriting and answer sheet tidiness; and (c) among the various linguistic features raters paid attention to, the most frequently cited was syntax or morphology, and the least frequently cited was cohesion and coherence.

Table 8

<table>
<thead>
<tr>
<th>Writing aspect</th>
<th>TAPs</th>
<th>SRs</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Content coverage</td>
<td>111</td>
<td>48</td>
<td>159</td>
</tr>
<tr>
<td>Language quality</td>
<td>349</td>
<td>205</td>
<td>554</td>
</tr>
<tr>
<td>(a) Syntax or morphology</td>
<td>155</td>
<td>78</td>
<td>233</td>
</tr>
<tr>
<td>(b) Lexis</td>
<td>49</td>
<td>28</td>
<td>77</td>
</tr>
<tr>
<td>(c) Frequency and gravity of errors</td>
<td>58</td>
<td>34</td>
<td>92</td>
</tr>
<tr>
<td>(d) Comprehensibility, fluency, and naturalness</td>
<td>40</td>
<td>30</td>
<td>70</td>
</tr>
<tr>
<td>(e) Cohesion and coherence</td>
<td>22</td>
<td>17</td>
<td>39</td>
</tr>
<tr>
<td>(f) Spelling and punctuation</td>
<td>25</td>
<td>18</td>
<td>43</td>
</tr>
<tr>
<td>Handwriting and answer sheet tidiness</td>
<td>62</td>
<td>45</td>
<td>107</td>
</tr>
</tbody>
</table>

Language quality was the most frequently cited aspect when raters made their judgements, although raters seemed to attend to slightly different linguistic features. Among the six features,
all eight raters referred to (a) syntax or morphology, (b) lexis, (e) cohesion and coherence, and (f) spelling and punctuation; while seven raters paid attention to (c) frequency and gravity of errors, and six reported attending to (d) comprehensibility, fluency, and naturalness.

*Content coverage* was the second-most considered aspect by raters. When reading and examining each essay, raters consistently looked for key words related to the major details and assessed the relevance of details to the writing task. For example, Ri-zhen recalled how he evaluated content coverage in Essay 14-06:

Then I was reading the details. He [the student] mentioned to go somewhere to study for six weeks. I first looked for “six weeks”, then I saw he mentioned it as required. Then he asked questions, which include the major details: start date, check; class size, check; hours per week, check; how much to pay, check.

Here is another example of Ri-ling looking for key words related to two major details while reading Essay 15-21: “Then next, I am quickly scanning the essay, looking for the next major detail, THE ARTICLE [length] IS ABOUT 400 WORDS. AND THE DEADLINE OF HANDING THE ARTICLE [expected due date] IS…”

The third-most frequently considered aspect by raters was *handwriting and answer sheet tidiness*. All eight raters tended to have more favourable impressions about neatly written essays and answer sheets. Four raters stated that poor handwriting created reading difficulties and forced them to make judgements based on what they could decipher and understand. Raters appeared to associate handwriting and answer sheet appearance with the predicated English language proficiency of the writer or qualities of the envisioned writer. Here are examples from different essays:

“This student’s handwriting is very neat, like being printed out. This must be a high
achiever.” (Ri-xia)

“This answer sheet is very neat, I think the quality of this student should be good.” (Ri-xiu)

“If the handwriting is good, it must be written by a student with a careful attitude.” (Ri-dong)

These raters demonstrated high consistency in the writing aspects they focused on (content, language, and handwriting). This seemed to suggest a strong training effect on these raters, which resonated with raters’ comments that they regarded the NMET rating experience as professional training that informed their essay rating practices afterward.

**Rating Sequence**

While attending to the above three aspects in rating, all eight raters followed a general sequential procedure for rating essays, as illustrated in Figure 14. Raters usually followed four steps when rating NMET essays. They first scanned the whole essay or the first few sentences of the essay, attending to the three aspects to form a general impression. Second, they made initial scoring decisions (the scoring band to which the essay should belong) based on their initial impressions. Raters would then continue to read the rest of the essay or skim it again, looking for further evidence primarily on content coverage and language quality. Fourth, raters articulated their final score decisions based on overall impressions of a composition by fine-tuning scores within the band range. Raters occasionally found their initial band range decision inaccurate. In such cases, they would articulate their adjusted band range decision and fine-tune the scoring decision within the revised band range.

---

5 In this dissertation, a score decision refers to an essay score as the results of raters’ assessment decision; a scoring decision refers to a process of raters’ assessment decision.
Figure 14. Rater’s general rating sequence

1. **Scan essay to form initial impression**, based on
   - content coverage
   - language quality
   - handwriting and answer sheet tidiness

2. **Make initial score band decision**

3. **Read essay more carefully**, focusing on
   - content coverage
   - language quality

4. **Articulate final scoring decision, while refining or revising initial band range decision**

Table 9 is an example of the sequence in Figure 14, showing how a rater, Ri-xiu, marked one essay.
### Translated Example of Think-Aloud Protocol Extract to Illustrate the Coding of Rating Sequence

<table>
<thead>
<tr>
<th>Transcript</th>
<th>Code</th>
<th>Rating sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Essay [15-]57: Looking globally, the answer sheet of this essay is not neat.</td>
<td>Articulate general impression, consider handwriting and answer sheet tidiness</td>
<td>1. Scan essay to form initial impression</td>
</tr>
<tr>
<td>[Let me] see if this essay contains the six major details as required [by the rating scale].</td>
<td>Decide on macro-strategy for reading and rating</td>
<td>1. Scan essay to form initial impression</td>
</tr>
<tr>
<td>(Reads the essay silently) It writes CULTURES OVERSEAS MODEL. First of all, this expression is not accurate.</td>
<td>(Re)Read or translate essay, consider lexis</td>
<td>1. Scan essay to form initial impression</td>
</tr>
<tr>
<td>It should not count [as a major detail]; it does not include the major detail of “Foreign Cultures” section.</td>
<td>Assess major details, refer to or interpret rating scale</td>
<td>1. Scan essay to form initial impression</td>
</tr>
<tr>
<td>Also, it writes something like TWO SPECIFIC CULTURAL EVENTS, COSTUMES OR DAILY BASIS ABROAD. These expressions are unclear.</td>
<td>(Re)Read or translate essay, consider lexis</td>
<td>1. Scan essay to form initial impression</td>
</tr>
<tr>
<td>[One major detail] basically fails to be included.</td>
<td>Assess major details</td>
<td>1. Scan essay to form initial impression</td>
</tr>
<tr>
<td>What’s more, (reads the essay) AN AMERICAN CITIZEN CAN WRITE AN ARTICLE TO POST ON OUR PAPER, AND WE WANT YOU TO DISCUSS ABOUT TWO TOPICS: ONE IS AMERICAN CUSTOMES OF NATIONAL HOLIDAYS, AND THE OTHER IS DAILY LIFE OF YOUR HIGH STUDENTS, here DAILY LIFE OF YOUR…OUR DEADLINE IS JULY.</td>
<td>(Re)Read or translate essay</td>
<td>1. Scan essay to form initial impression</td>
</tr>
<tr>
<td>We ask for JUNE 28, he writes JULY 28.</td>
<td>Assess lexis</td>
<td>1. Scan essay to form initial impression</td>
</tr>
<tr>
<td>And regarding the word limit, he writes AS YOU CAN IN AROUND 400 WORDS.</td>
<td>(Re)Read or translate essay</td>
<td>1. Scan essay to form initial impression</td>
</tr>
<tr>
<td>Generally speaking, this essay is not good.</td>
<td>Articulate general impression</td>
<td>1. Scan essay to form initial impression</td>
</tr>
<tr>
<td>[I] can place it to Band Two between 6 and 10 points, [even Band Two] may be too high for it.</td>
<td>Articulate or revise scoring decision</td>
<td>2. Make initial score band</td>
</tr>
</tbody>
</table>
As shown in Ri-xiu’s example, raters tended to first scan the essay using various interpretation and judgement strategies in Step 1. They kept comparing essay details with the major details required by the rating scale, while attending to language quality and handwriting. Based on these three aspects, they formed an initial impression of the essay quality. Step 2, raters made an initial score band decision based on their impression in Step 1. Step 3, raters scanned the essay again, adopting judgement strategies while attending to further details about content and language. Step 4, based on the overall impression of the essay, raters refined or revised their score band decision before articulating a final score decision.

The rating sequence above was identified based on the transcribed TAP data, and confirmed with SR interview data when raters reported their rating sequence. For example, during the SR, Ri-wei explained the procedures he followed when rating NMET essays:

When I rate an [NMET] essay, I start reading it from the beginning to the end, examining various aspects. For example, whether or not it includes all the major details, its grammar [is correct], its sentences are natural, it includes connectives and advanced syntactic
structures, and the handwriting [is neat]. Then I decide on the score band [the essay belong
to] and assign an appropriate score [within the band]. I would consider all those aspects
before assigning a final score.

Ri-yue summarized a similar rating sequence, “[I] scan [an essay] and decide on a score band
[the essay belong to]; then [I] carefully read [the essay] again [before assigning a final score].”

I now elaborate on the four steps presented in Figure 14.

Step 1: Scan essay to form initial impression. Raters began marking an essay by first
scanning the partial or entire essay to generate an initial impression. This impression was formed
by attending to the three rating focuses and by adopting various interpretation and judgement
strategies. Handwriting and answer sheet tidiness was often considered at this stage. Ri-dong’s
think-aloud began as follows: “Essay [15-]57. [My] general impression [of this essay] is [its] handwritting is poor, with multiple places crossed out.”

Raters then continued to look for key words relevant to the required major details while
assessing relevance of additional details to the central topic. Meanwhile, raters also attended to
various linguistic features to assist their interpretation of the essay and judgment of language
quality. Here are two examples. In the first, Ri-ling assessed a major detail and the cohesion of
an essay simultaneously: “The second major detail is present, INTRODUCE THE CUSTOMS
OF AMERICAN FESTIVALS, followed by a good connective AS WELL AS.” In the second,
Ri-wei assessed a major detail and attended to syntax when marking an essay:

The second paragraph mentioned studying six weeks, but did not emphasize where to;
then it continued, I WANT TO KNOW. Here [the student] wanted to use [a structure of]
link verb plus a predicative, but ended up with [a structure of] link verb plus a verb; this
sentence is incorrect.
**Step 2: Make initial score band decision.** Based on their initial impression of an essay, raters tended to make an initial scoring decision about the score band. For example, after scanning the entire Essay 14-04 for the first time, Ri-juan articulated his score band decision, “so after first reading, [I think] this essay should be placed below Band Four, that is, below the pass/fail [cut-off score]”. Similarly, Ri-ling articulated his score band decision after scanning the first few lines of Essay 15-35: “This essay . . . based on my first impression, should be [placed] at the higher end of Band Three or the lower end of Band Four; it should be at this level.”

**Step 3: Read essay more carefully.** After deciding on the score band of an essay, raters continued to read the essay more carefully or skimmed the essay again, while gathering additional evidence on content coverage and language quality to assist with their final scoring decision. Ri-zhen skimmed the essay after articulating his initial score band decision:

[Essay 14-12] should be [placed] at least at Band Four. Let me see, in addition to BESIDES, [the student also used] GET ALONG WITH OTHERS TO LEARN ENGLISH, HOW MUCH I HAVE TO PAY . . . This sentence reads ok, [it is] fluent. The sentences are fluent. The language [this essay used] is fine. SEEN YOUR AD IN THE NEWSPAPER, I’D LIKE TO KNOW SOME-. . . SIX-WEEK, this is the first major detail, present; the second major detail, the third major detail, the fourth major detail, HOW MUCH YOU KNOW . . . the major details are all included.

**Step 4: Articulate final scoring decision, while refining or revising initial band range decision.** After gathering more evidence from Step 3, raters usually articulated their final scoring decision within the band. Using Ri-yue’s think-aloud as an example of refining a scoring decision: “It [Essay 14-12] should be placed at Band Four, that is, between 16 and 20 points. It should be [placed] at the higher end of Band Four, so it should be assigned 19 points”.

112
However, sometimes raters may find their initial score band decision inaccurate based on evidence they collected from Step 3. In such cases, raters would articulate their scoring decision by adjusting the initial band range decision and then fine-tuning within the revised score band. Here is one example of Ri-ling marking Essay 15-35. His first impression of the essay was upper Band Three or lower Band Four. But as he continued reading, he noticed the writer used some smooth transitions and ended the letter with a good closing sentence, which he thought performed the communicative task well. The writer’s handwriting also seemed to leave a positive impression. After Ri-ling finished reading the essay, he reconsidered his score decision: “[This essay] should deserve a Band Five. The final score I assigned to it is 21 points”.

In another example, Ri-wei lowered his initial scoring decision on Essay 14-10 based on content (absence of one major detail):

This [essay] could be a low Band Three. [I] thought about assigning it 11 or 12 points, but considering its absence of a closing sentence…so it has…not completely fulfilled the required task, therefore [I] lower it by one band, assigning it 10 points.

In terms of rating sequence, these raters showed convergence in the sequential procedure they followed when rating NMET sample essays. Such results again echoed the speculation about a strong training effect they acquired from the NMET operational rating sessions.

**Factors Influencing Decision-Making**

Factors that influenced raters’ decision-making are grouped into five categories: (1) institutional requirements; (2) high-stakes consequences to students; (3) saving “face”; (4) prior teaching and rating experience; and (5) advice from colleagues. I point out that although the impact on rating results (i.e., scores) is beyond the scope of this study, raters cited the impact of these five factors on their final score decisions in their interviews. As score decisions are part of
decision-making processes, I include this portion of information in the reported findings. In this subsection I describe these factors, in terms of how they impacted raters’ decision-making (i.e., rating focuses, rating sequence, and score decisions).

**Institutional requirements.** Institutional requirements were raters’ dominant consideration when rating NMET essays, derived from documents illustrating rating scale and specifications, rater training, and rating quality indicators in the OSM system.

**Rating scale and specifications.** All eight raters said NMET rating scale and specifications exerted the biggest impact on their decision-making. Those documents were “framework documents” (Di-nan) that established the fundamental rating principles and criteria for NMET raters. The documents guided raters on aspects to attend to, the rating sequence to follow, and score band decisions. For example, Ri-xiu described, “My major strategy was to closely follow the rating specifications, looking for the required major details [in the essay], and then decide on its appropriate score band.” At the same time, rating scale and specifications were cited by every rater during their TAPs and SR interviews. When rating essays, they kept attending to the major details stated in the scale, assessing content relevance to the assigned topic, considering different linguistic features indicated in the scale, and summarizing and comparing their impressions with band descriptors to make or revise their score band decisions.

**Rater training.** The second source of institutional requirements was rater training. Rater training included formal training raters received prior to the operational rating session and informal training during the session. Rater training complemented rating scale and specifications by norming raters’ interpretation and application of the NMET rating criteria specified in those documents. As Ri-dong commented, “The rating specifications [document] were our rating guide; the entire training process [helped me] internalize this guide.” Anchor essays and team
leaders were identified as factors that facilitated this norming process.

Six raters referred to the influence of anchor essays on their decision-making. These raters used anchor essays (and scores assigned to those essays by expert raters) as rating criteria exemplars (i.e., illustrations of how the rating criteria should be applied to rating specific essays). At the beginning of training, raters used the fit between the scores they assigned and experts’ scores as an indicator of their understanding of the rating criteria and to inform the adjustment of their rating severity. For example, Ri-ling described: “[At the beginning of the training session] I was concerned I had not accurately grasped the rating specifications and the scores I assigned were too far from the scores expert raters assigned.” Ri-wei observed the difference between the scores he assigned to anchor essays and those by experts and concluded that he was more lenient. “Learning that, when deciding on final scores, I would lower my scores to be closer to experts’ scores.” During the operational rating session, raters might go back to anchor essays when they encountered rating difficulties. They studied features in anchor essays within the same score band as the essay they were reading at the time, to help assign a reasonable score.

In addition to anchor essays, seven raters stated that their team leaders facilitated this norming process which had an impact on their decision-making. Some viewed their team leader as an authoritative instructor who explained and clarified the rating criteria, guiding raters to interpret and apply the rating criteria in the same way. “Otherwise, we cannot rate essays [accurately]. Each rater has his own interpretation of the [rating] criteria” (Ri-juan). Besides communicating with team leaders during the formal training session, raters also turned to their team leaders for clarification and elaboration when they encountered difficulty in applying the rating criteria to specific essays. For example, Ri-dong found it was particularly difficult for him to distinguish essays of 8 to 12 points. He discussed this with his team leader, “He told me what
features essays worthy of 7 or 8 points, of 9 or 10 points, and of 11 or 12 points. After he pointed that out, the specifications were much clearer to me and easier to apply.” There were also times when raters encountered difficulties (e.g., uncertainty about a score decision, having rating errors) and would rely on their team leaders’ expertise to resolve an issue. For example, Ri-yue described a time when he was uncertain about a final score decision, “I ran into a well-written essay, but was not sure if I should assign 22 or 23 [out of 25 points], or even 24 or 25. Then I consulted our team leader and asked him to decide.” Raters would also consult their team leader when they had scoring disagreements with other raters. In cases like this, the team leader was both a judge and an instructor. For instance, at the beginning of the operational session, Ri-yue, noticing a consistent discrepancy beyond the acceptable level between his scores and the second raters’ scores, turned to his team leader to see whether those rating errors were his or of other raters, “He [Ri-yue’s team leader] examined those high discrepancy essays and told me some of them were my errors. Then he analyzed those essays with me and gave me some advice on how to address my rating errors.” It was noted that the impact of team leaders in those two roles was reported specifically by first-time NMET raters. One plausible reason was, according to Ri-dong, that the rating specification descriptors were “too vague to master.” So the first-timers consulted their team leaders more often about how to interpret and apply the rating specifications when encountering specific types of essays.

**Rating quality indicators.** The third source of institutional requirements came from rating quality indicators in the OSM system. All eight raters reported rating quality indicators (e.g., valid rating rates, and serious rating error rates\(^6\)) as factors that influenced their score decisions.

---

\(^{6}\) These two types of “rates” both refer to the computed percentage of one rater’s certain type of ratings compared with the rater’s total number of ratings. When a score a rater assigned to an essay is within the acceptable level (i.e., no greater than four points) when compared with the score assigned by a second, or in some cases third, rater to the
Raters used these indicators as references or checks on whether or not they had been deviant from the norm, which affected the final scores they assigned. Ri-wei described, “Sometimes I would consider how other raters might think and assign scores, and tried to assign scores close to theirs . . . because we were required to keep serious rating error rates low.” Similarly, when Rizhen’s serious rating error rate increased, he wondered if he had adopted different criteria than other raters, and was concerned about the scores he assigned in relation to other raters, “When I read the next essay, I could not help thinking what scores other raters might assign to it . . . and ended up with a score in the middle, not too high or too low”.

**High-stakes consequences to students.** The high-stakes consequences to students were a second factor that six of the eight raters cited as an influence to their score decisions. The impact of this consideration has twofold meanings. On one hand, implications of their score decisions to individual students tended to make raters’ final score decisions more lenient. For example, Ri-wei described, “I was always thinking about how Gaokao determines their fates, so I tended to be more lenient. Compared to the second raters, I often assigned slightly higher scores.” On the other hand, such awareness of the high-stakes consequences appeared to heighten raters’ sense of responsibility and dedication to be fair and impartial to all students, which restrained the extent of their leniency. As a result, the six raters all stated that they tended to be more lenient, but not too generous if they thought it would threaten the fairness to other students. “This is Gaokao. If I can assign one more point, I certainly will. But I have to be fair and objective; otherwise it is not fair to other students” (Ri-xia).

**Saving “face”**. Saving “face” was a third consideration cited by seven of the eight raters. Saving “face” here means that a rater tried to avoid being judged as less competent than other
raters, in order to claim deference for himself or herself from others. Information about individual raters’ statistics (e.g., number of completed ratings, and those rating quality indicators) were often shared within each rating team. Besides seeing various quality indicators as institutional requirements for rating quality control (as explained in the first category), raters also felt those quality indicators posed potential threats to their “face”, by influencing how others would judge them, hence creating pressure on them. Ri-xiu explained: “When you saw other raters’ [valid rating rates] were high and your own was low, you started to question [yourself] . . . you certainly felt embarrassed.” This pressure made him pay more attention to his own statistics, “When your statistics [valid rating rate and number of ratings] went up, you wanted to keep the record or make them even higher; when they dropped, you started to think why mine dropped and how I could improve them.” Consequently, to avoid losing “face”, raters attended to maintaining or improving their own statistics, and one strategy some raters used was to consider the possible score a second rater might assign to the same essay and decide on a score in between. Ri-ling admitted when he had rating difficulties and was uncertain about essay scores, he would consider the influence of assigning certain scores on his valid rating rate. “Others would look at your valid rating rate. So you need to pay attention to it, and subconsciously ended up with some central scores” (Ri-dong). Ri-zhen described that when he saw his serious rating error rate went up, “you would subconsciously start to think about the highest and lowest scores this essay can be assigned. If you did not want it to go to a third rater, you would assign a score in the middle.”

**Prior teaching and rating experience.** The fourth factor that influenced raters’ decision-making was their prior teaching and rating experiences.

**Prior teaching experience.** All eight raters used their prior teaching experience to judge
essay language quality (when attending to the range and accuracy of lexical and syntactic resources students used) and to envision students’ personal situations (e.g., proficiency level, attitude).

Raters considered successful use of the lexical and syntactic resources that high school students were expected to master as evidence of students’ high English proficiency levels. Those resources came from high school curricula and raters’ experience with training their students for NMET writing test preparation. Raters often referred to those lexical and syntactic resources as “advanced-level vocabulary and sentence patterns” or “shining points.” When raters were attending to the language quality of an essay, they would compare the lexical and syntactic resources the student used with “advanced-level vocabulary and sentence patterns”. For example, while first scanning Essay 14-10, Ri-zhen noticed that in the last paragraph the writer used a phrase included in the high school curriculum, “(The student) used the expression ‘last but not least’ which he has learned in high school”. Similarly, Ri-ling commented on an error in the last paragraph of Essay 15-21:

This essay ends very well, I WOULD APPRECIATE IF YOU HELP ME. But the syntactic structure it used [here] was not completely aligned with the classic syntactic structure we ask our senior high students to remember, that is, something is missing here.

When making final score decisions, raters often took into consideration the number and accuracy of “shining points” included in an essay, as Ri-dong did when summarizing his judgements and articulating final scoring decision: “This [Essay 15-53] should belong to Band 4. It also used some simple connectives, which is a shining point in the essay. [It] should be assigned 17 points.”

Raters also relied on their teaching experience to envision students’ personal situations—
such as their proficiency levels and attitudes—to assist with scoring decisions. Several first-time raters reported that at the beginning of the rating session they often compared the proficiency levels demonstrated in NMET essays with their own students’ proficiency levels to help them make score decisions. Ri-wei reflected on a technique he used to form score decisions, when acting as NMET rater for the first time: “I kept thinking about which of my students’ proficiency level can be matched with the essay I was rating, imagining it was written by my own student, then I would assign a slightly higher score”. Ri-xiu also drew on his teaching experience to predict students’ proficiency levels according to the number of grammatical errors in the essays before he made final score decisions. However, raters reported that such influences faded during the rating session as they gradually mastered the rating criteria. Ri-wei asserted that this consideration was no longer an influence on score decisions in his second-year of NMET rating. Raters also drew on their teaching experience to predict students’ attitudes and proficiency levels based on their handwriting and answer sheet tidiness. They believed there was a relationship between students’ handwriting and their attitudes or proficiency levels.

**Prior rating experience.** Prior rating experience was also mentioned as an additional influencing factor by two returning raters. Both Ri-ling and Ri-wei attributed their speed rating essays and confidence about final score decisions to their prior NMET rating experience. In 2015, Ri-wei’s second year of participation, he said: “Right after reading an essay once, I had a clear idea about in which score band this essay should be placed and how many points I should assign. I felt I was faster and more accurate than last year.”

**Advice from colleagues.** Advice from teaching colleagues with prior NMET rating experience shaped three first-time raters’ rating sequences and score decisions. For two of the raters, this influence only surfaced at the beginning of the operational rating session. Ri-dong’s
teaching colleague Ri-ling gave him some rating strategies to help him read essays more efficiently: “He [Ri-ling] suggested me to read an essay by skimming for key words [instead of reading word by word] . . . at first I thought that was impossible. Later with more practice, I found myself able to do it”. When Ri-wei served as NMET rater for the first time in 2014, his valid rating rate was lower than others at the beginning of the session. He communicated with his teaching colleague Ti-zhong, one of the team leaders, who examined the essays where his scores were highly different than the second raters’ and found that Ri-wei’s scores tended to be higher. Ti-zhong then suggested that he be less lenient: “He told me, [for example,] if I want to assign 18 points, I could reduce by two or three points.” For Ri-xiu, it was his prior communication with colleagues that shaped his rating practices. At his school, teachers who participated in NMET rating would share their experience with other teachers about how the NMET rating criteria were applied that year. Based on his colleagues’ experience, the teachers at his school adopted a localized version of NMET rating criteria when rating their students’ essays. He explained, “When we mark student essays, we assign certain points to each major detail of content coverage. If there is one major detail missing, we deduct certain points [from total points]”. This school-based rating practice he learned from colleagues with prior NMET rating experience influenced Ri-xiu to adopt the same approach (i.e., point deduction) when making final score decisions on NMET essays.

The results of these influencing factors showed that, although these raters’ similar rating focuses and sequential procedures seemed to suggest a strong training effect, their decision-making actions in fact involved interactions between various contributing factors, far more complicated than the assumed training effect. The investigation of rating focuses and sequence alone seems inadequate to capture such complexity. To further understand these raters’ decision-
making actions, it is important to situate their actions within a wider sociocultural context where those actions took place.

**Section 1 Summary**

This section presented findings from eight NMET essay raters’ think-aloud protocols, stimulated recalls, and interviews. In relation to raters’ goal-directed decision-making actions when they rated NMET essays, three themes emerged from the data: rating focuses, rating sequence, and factors that affected decision-making. Within rating focuses were three categories: content coverage, language quality, and handwriting and answer sheet tidiness. These were the three major aspects of writing raters attended to when assessing NMET essays. One general rating sequence that raters followed to assess NMET essays was identified: they first decided on a score band based on an initial impression, then fine-tuned their assessment to obtain a final score decision. The considerations and factors that seemed to influence raters’ decision-making actions were grouped into five categories: institutional requirements, high-stakes consequences to students, saving “face”, prior teaching and rating experience, and advice from colleagues.

Raters cited documents illustrating rating scale and specifications, rater training, and rating quality indicators in the OSM system as institutional requirements that influenced their decision-making in terms of rating focuses, rating sequence, and score decisions. With regard to high-stakes consequences to students and saving “face”, raters described that these two considerations impacted their score decisions. The raters described that their prior teaching and rating experience also influenced their decision-making in terms of rating focuses, rating sequence, and score decisions. Finally, raters reported that advice from colleagues impacted their rating sequence and score decisions.
Section 2: Raters’ Object-Oriented NMET Essay Rating Activity

The previous section highlighted the various themes about raters’ goal-directed decision-making actions that arose from TAP, SR, and interview data. That section did not, however, situate the findings within the sociocultural context where these raters’ collective NMET ERA took place. Due to the situated nature of individuals’ goal-directed actions in relation to the collective object-oriented activity revealed by the guiding CHAT framework, the collective NMET ERA in this section is described in terms of its broader sociocultural context and the participants’ perceptions of NMET ERA in which they participated. I provide a detailed description of how various themes and categories appear within the broader sociocultural context of Gaokao rating in China and the NMET ERA within the two settings (rating centre and high schools). These findings allow me to gain a contextualized understanding of the findings I reported earlier on the cognitive layer.

Setting the Scene: Gaokao Rating in China

Description of the broader sociocultural context of Gaokao rating in China is grouped into three categories: social recognition of Gaokao, importance of Gaokao to test-takers and their families, and social attention to Gaokao and its rating.

Social recognition of Gaokao. Gaokao was cited by two-thirds of the participants as a fair and reasonable means of selecting talented individuals for university admission, irrespective of socioeconomic status. Ri-dong described Gaokao as “a heritage of the 2,000-year history of the Imperial Examinations to select the talented in China”, and Ti-rong recognized it as “a relatively fair opportunity for students who do not come from a privileged background or were not born in a rich family”. These assertions were made based on two considerations. First of all, participants felt strongly that Gaokao provided test-takers with a fair competition platform, because
university admission decisions were made solely based on Gaokao scores. Considering social realities in today’s China (e.g., disparity between the rich and poor, unbalanced development in urban and rural areas, presence of inequality and malpractices in the society), “Gaokao is so far the most reasonable and relatively fair means of selecting the talented individuals” (Ti-zhong). Another reason why participants recognized Gaokao as fair was because test-takers at lower socioeconomic status may have an equal chance to succeed and climb the social hierarchy. Dianan cited an example of test-takers from underdeveloped rural areas. “Through Gaokao, their life can truly be changed. Even without any privileged background, if you succeed in Gaokao you can still go to a good university, which leads you to a good job and a good life.”

Due to the recognition of Gaokao, some participants commented on the Gaokao slogan “Beat the tall, rich, and handsome; beat the second generation of the officials!”, which means “to outperform the rich and powerful” by taking Gaokao. Ri-ye told me that although this slogan was overstated, it contained some truth. He even printed it out to motivate his daughter when she was preparing for Gaokao: “I am a teacher; I have no power or money. Without Gaokao, my child could not compete with the children from the rich and powerful family and be successful.”

**Importance of Gaokao to test-takers and their families.** In spite of the rapid economic development in China and the recent rise in the number of Chinese students studying abroad, a lot of participants stressed that Gaokao remained the only path to university for the vast majority of Chinese students, bearing high stakes to both test-takers and their families.

Several participants from rural families shared their own experiences about how their life courses had been altered after success in Gaokao; and some shared stories about their students from poor families who worked extremely hard in the hope of succeeding in Gaokao and, thus, changing their lives. They commented, “Gaokao is a threshold to rural children. Stepping over
this threshold, these children could leave rural areas and free themselves from the kind of life their parents have lived” (Ci-ping). “To children from poor families, Gaokao is their saviour. Without Gaokao, how can they get out and see the outside world? . . . I told my students, Gaokao is their opportunity for rebirth” (Ti-ren).

Gaokao is a life-changing opportunity for all students, not just those from rural areas and poor families. Students’ Gaokao performance determines which university they can enter, in turn shaping future success and job opportunities. Pi-shi cited the difference in job opportunities for graduates from Tiers One and Two universities, with the ones from renowned Tier One universities being favoured, “Some employers may reject you if you do not have a bachelor’s degree from a Tier One university”. Because of such high-stakes consequences to test-takers, half of the participants commented on the Gaokao slogan, “One point raised, 1,000 competitors eliminated”. They agreed that this slogan sounded cruel, but was true. Due to such a fierce competition in Gaokao, participants declared that every point was important and may result in a different life. “Sometimes with one or two points’ difference, you may or may not get into a Tier One university” (Ri-ling).

In addition to the importance of succeeding in Gaokao to test-takers, many participants cited its importance to test-takers’ families. Perception was widespread among parents that if their child could enter a good university, that would bring glory to their family name and change the fate of future generations. “Parents hope their children can be more successful than themselves. . . . They believe if their children can enter Qinghua [University] or Beida [Beijing

---

7 In China, universities are categorised under Tiers for the purpose of university entrance via Gaokao. Tier One universities usually are well-funded and enjoy good reputation. Tier Two universities include the bulk of four- and three-year universities and colleges across China. These two tiers make up the majority of public universities and colleges in China (Commonwealth of Australia, 2009; D. Wang & Gao, 2013).
University]⁸, their whole life and the fate of their family’s future generations will be changed” (Pi-shi). Several participants shared personal experiences to demonstrate Gaokao’s importance to test-takers’ families. Di-nan told me when he was a student his parents held high hopes for him to succeed in Gaokao and live a good life, “My parents frequently talked about others’ children going to universities in Beijing and Shanghai, working in big cities, or going abroad. They instilled these thoughts in me.” These parental expectations made him aware of the family honour related to success in Gaokao, “I told myself I should study hard and someday I would go to an even better city and return to my hometown in glory”. Having made it, his parents were proud of him and he became the trophy child in other parents’ mouths. “Gaokao is important because it not only benefited ourselves, but also provided an opportunity to reward our parents” (Ti-jian). That explained why some participants had strong feelings toward the Gaokao slogan “I am crazy about Gaokao; my goal is to enter a key university to show my filial piety to my parents”. These participants concurred that it was in line with the Chinese virtue of filial piety as a means of showing gratitude to their parents for the parents’ years of upbringing and support.

Social attention to Gaokao and its rating. Not surprisingly, Gaokao and its rating attracted wide attention every year during Gaokao season. “Millions of households have a bearing on Gaokao” (Ri-wei). “The whole country attaches a lot of importance to Gaokao. Whenever Gaokao is coming up, you hear a lot of reports about how the city and police are ensuring its secure and smooth administration” (Ci-lei). In addition to countless news reports on various aspects related to Gaokao every year, the broad social attention to its rating work on different Gaokao subjects can be seen in the volume and contents of relevant news reports. Over three Gaokao rating periods (June of 2013, 2014 and 2015), Xinhua, People’s Daily, and

---

⁸ Qinghua University and Beijing University are among the top universities that are highly reputed in China.
Sina.com.cn published a total of 230 news reports about Gaokao rating. These reports were grouped into five general topics: (a) Gaokao rating preparation and progress updates across various provinces and regions (103 of 230 reports, 44.8%); (b) measures taken by provincial rating centres to ensure test security and rating quality (65 reports, 28.3%); (c) visits by test-takers, parents, news reporters, members of the People’s Congress, and provincial government leaders to rating centres made 43 headlines (18.7%); (d) summary of test-takers’ performance and suggestions for future test-takers by rating centre administrators (41 reports, 17.8%); and (e) qualifications and selection of raters (22 reports, 9.6%). These 230 news reports introduced a wide range of stakeholders (e.g., Gaokao rating administrators and staff, raters, government officials, test-takers, parents, news reporters) involved in Gaokao. They formed a picture of how Gaokao rating held nationwide attention and described the background of the importance of Gaokao rating in Chinese society.

**NMET ERA at the Two Settings**

Participants’ perceptions of NMET ERA are grouped into three themes: “sacred mission with grave responsibility”, “stressful”, and “beneficial”. Findings concerning the first two themes were contributed by participants at the rating centre setting, the third by participants at the school setting. Although these themes are described sequentially below, it is important to note that these perceptions were interconnected and sometimes occurred concurrently during participation in the NMET ERA.

**“Sacred mission with grave responsibility”**. Most participants, including Ti-ren, considered it an honour to take part in the NMET ERA. Ti-ren described the moment when he was trusted with the role as a team leader: “My supervisor said to me, ‘This is a heavy job with weighty responsibilities. Please take it carefully and seriously.’ Hearing this, I suddenly realized I
was assigned a sacred mission with grave responsibility.” When Ri-xiu first saw the slogans and austere security guards at the entrance to the rating centre, right away he felt NMET rating as an extremely sacred task. Ri-dong said once he sat in front of the computer and started rating, he felt the onus on himself and became serious. “The slogan ‘Iron shoulders take on morality and justice’ describes exactly how I felt” (Ri-dong).

This perception was closely related to raters’ awareness of the potential high-stakes consequences of their assigned scores to students, which aroused their sense of responsibility. Many participants admitted being well aware of possible consequences to students’ future life choices, “This is Gaokao . . . every point counts” (Ci-ping); “If I assigned one or two extra points, a student may pass Tier One or Tier Two cut-off scores” (Ri-ling); “So many children’s fates are in your hands. Because of one or two points you assigned differently, a child’s fate can be changed” (Ri-zhen). These thoughts heightened participants’ sense of responsibility and dedication to be fair and impartial. “I was heavily burdened with the sense of mission and responsibility, I kept telling myself I could not make any rating error or be biased” (Ri-zhen).

Bearing in mind the high-stakes consequences and shouldering a sense of responsibility, participants contended that they were cautious when rating essays and strived to assign reasonable and fair essay scores to the extent possible. “I was much more careful and cautious [when rating NMET essays] than when I rated my students’ essays” (Ri-zhen). “I should assign fair and reasonable scores for students. If I had not rated properly, their promising future would have been dimmed . . . that would make me feel sorry” (Ri-xiu). Meanwhile, participants such as Ri-ling tried to look for positive elements in essays and assign slightly higher yet reasonable scores: “It was different than rating students’ essays at school, where the purpose is to correct students’ errors. When rating NMET essays, my goal was to avoid putting students in an
unfavourable situation and try to assign slightly higher scores”, adding “many teacher raters expressed similar considerations”.

Consequently, participants felt relieved and proud at the end of NMET rating session when they completed the rating task satisfactorily, according to their rating quality indicators in the OSM system. But they were stricken with anxiety, remorse, and restlessness when they became aware of the rating errors they made or achieved low rating quality indicators. Ci-ping reported that he was very nervous about his valid rating rate, “I kept looking at mine [on the projector screen], afraid that it went down. If [my] valid rating rate was low, I was worried I did not assign reasonable scores for students.” When Ri-juan noticed his valid rating rate was going down, he felt remorseful. “Although after third ratings those errors were corrected, I still blamed myself.” Ci-hai told me once he assigned a low score to a disputable essay and realized that that essay should have been assigned a higher score after he later discussed it with an expert rater, “I did not sleep well that night; I felt very sorry [for the student]”.

“Stressful”. Participants considered it an honour to take part in the NMET ERA but also a stressful experience. Team leaders noticed that raters became nervous once the pilot rating session ended and the operational rating session began; raters were not used to the institutional constraints. In the first few days, Ri-wei’s team leader kept updating the team on the rating volume of the fastest raters in their team and the total rating volume of their team relative to other teams. “Everyone pressed forward and feared to lag behind others. Some teachers continued rating without taking breaks” (Ri-wei). In the OSM system, there was a bar set on the valid rating rate indicating satisfactory rater performance. Raters felt anxious and stressful when they saw their valid rating rates declining and approaching the bar level. Some raters felt pressured when they heard that raters with unsatisfactory performance would be summoned to
talk with the quality inspection team leader. Ri-ling told me, “Many first-time raters struggled over how many points they should assign, worrying that they were not doing well. . . . Some were so stressed that even in their dreams at night they were still rating essays”.

Such stress and pressure were caused by a chain of challenges and tensions. One common challenge facing raters was how to apply the rating scale and specifications properly over thousands of different essays. Different participants reported that essays at the middle ranges or at the extreme ends were difficult to score based on the stated rating criteria; the rating criteria apparently failed to cover all the eventualities. For example, some band descriptors were vague, leading Ri-xiu to complain, “Descriptors such as few and some were really giving me a headache”. Ri-dong described the difficulty he had with the rating specifications concerning essays scored between 8 and 12. “Essays at this range all feature long, messy sentences. It was very difficult to make decisions. Sometimes you felt like assigning 7 or 8 points, sometimes you felt 11 or 12 points was also legitimate”. This challenge of “theory [i.e., the stated criteria] versus practice [i.e., applying the criteria]” (Ri-dong) often led to another challenge facing many raters—balancing the two institutional requirements of rating accuracy and rating speed. The challenge caused by the ambiguous rating criteria posed a potential threat to raters’ rating accuracy. Their rating load was heavy, yet the requirements for rating accuracy were high. As a result, raters found maintaining both accuracy and speed challenging. “If you want to be accurate, you have to spend more time on each essay. Then your rating speed is lowered. If you rate fast, you would start to worry your scores are not accurate” (Ci-su). This second challenge further led to a sense of external (meeting the institutional requirements) and internal pressure (self-imposed pressure to rate fast and accurately, because raters perceived failing to do either as a potential threat to their “face”) among raters. “If the number of ratings you completed is small,
that makes you look bad” (Ci-hai). “If you saw on the projector screen others’ valid rating rates were so high and in green⁹, and your own was in red, you would feel very embarrassed” (Ri-dong). Some raters considered it a “disgrace” (Ci-lei) if they were asked to talk with the quality inspection team leader. As a result of such pressure, “you would keep thinking about how to increase it [your valid rating rate]” (Ri-juan).

To alleviate the stress, raters often relied on themselves first, before consulting team leaders or discussing their ratings with colleagues. They would (a) reread rating specifications more carefully and draw on prior teaching or rating experience in an attempt to figure out the fine differences between band descriptors; (b) closely examine anchor essays and constantly compare essays against each other to adjust scores; or, occasionally, (c) speculate on the highest and lowest possible scores the second raters might assign to specific essays before assigning safe scores to avoid risking lowered valid rating rates. If raters could not solve difficulties by themselves, they would ask team leaders to further clarify the rating criteria, identify types of their rating errors, or assist in making score decisions. “More often than not, team leaders were kept busy” (Ri-dong); so raters, especially some first-timers, frequently chose to discuss their ratings and look for solutions with their colleagues. Raters found it helpful to talk with other raters and learn from them how to deal with rating difficulties, consult experienced raters about strategies to rate fast and accurately, or discuss scores assigned to specific essays when they felt uncertain. Most raters felt the pressure subsided with this combination of strategies.

“Beneficial”. Most participants from the four high schools considered participating in Gaokao rating beneficial to a teacher’s career, which has twofold meanings: creating incentives for teachers in these schools to participate, and describing how teachers felt after participation.

---

⁹ If raters’ valid rating rates were above the bar of satisfactory performance in the OSM system, their statistics were displayed in green; otherwise, displayed in red.
Participants held two incentives to participate in Gaokao rating: First, to experience Gaokao rating, which was considered essential to a full teacher’s life; and second, it provided a professional development opportunity. Many participants stressed that “a high school teacher’s experience would be incomplete without experiencing NMET rating” (Ci-hai). Ci-lei explained, “If you have taught for many years but have never experienced NMET rating, when your students ask you questions about it, you would not know how to answer them.” Ti-zhou thought that without experiencing NMET rating, “Your understanding [of NMET] would be incomplete and what you can teach your students would be incomplete.” Furthermore, most participants asserted that such rating experience could be a valuable professional development opportunity for teachers. Participants claimed that their incentive to participate in Gaokao rating was to obtain first-hand information about how the rating criteria were applied and how students responded to the writing task, in the hope of getting a clearer idea “about the trends of NMET and towards which direction I should guide my students and make efforts” (Ri-xiu). Participants also valued the NMET rating experience for the opportunity to communicate with teachers from other schools, which could be “an eye-opening experience” (Ri-juan).

This perceived importance of participating in Gaokao rating was highly related to the impact of Gaokao on high school education, and the significance of Gaokao to these high schools and to their teachers. It was widely acknowledged that Gaokao has shaped the direction of high school education in China. “Gaokao is a guiding wand: We take whichever direction it points” (Ci-lei). Participants cited that Gaokao dictated the contents and methods of teaching and assessment format in their schools. As Pi-xue commented, “High school teaching, to a large extent, is to prepare students to succeed in Gaokao”. This considerable impact of Gaokao on high school education was associated with social recognition of Gaokao and high-stakes
consequences of Gaokao performance for test-takers and their families. These three aspects concerning the impact of Gaokao determine the critical importance of Gaokao to high schools.

Gaokao was compared to “life to high schools” (Pi-li). Gaokao performance determines whether a high school survives or not: It is used as a measuring stick to predict a school’s success in attracting new students. “The society judges a high school by its Gaokao performance” (Pi-li). Every year after the release of Gaokao results, a popular social topic is to discuss and compare different schools’ Gaokao performance. “People will ask how many students in your school have been admitted into Qinghua [University] and Beida [Beijing University], and how many have reached the Tier One cut-off score” (Ri-dong). This practice allowed schools to predict student enrolment based on their Gaokao performance. Pi-shi explained, “It is very noticeable that if our school does well in Gaokao this year, our new student enrolment would turn out well; if we do poorly, our enrolment would turn out poorly.” He continued to explain that poor Gaokao performance in one year may lead to poor performance the next, creating a vicious cycle that may threaten the survival of a high school. This is analogous to school evaluation practices under the former No Child Left Behind act in the United States, where schools failing to make annual academic progress in standards-based assessments would face escalating punishment that ultimately threatened the schools’ survival (McNamara & Roever, 2006). This similar connection pushed the schools in this study to attach tremendous importance to Gaokao performance and also passed down the pressure onto their teachers.

Students’ Gaokao performance was often associated with recognition of a teacher’s capacity and teachers’ self-esteem. These four schools all assigned Gaokao Zhibiao (“Gaokao target”, the targeted number of students who pass Tiers One and Two university cut-off scores) to their Grade 12 teachers. Achievement of Gaokao Zhibiao is a well acknowledged indicator of a
teacher’s teaching and work capacity. Ri-ling stated, “The principal cares about it, and teachers also keep an eye on each other’s [Zhibiao achievement]. . . . Students also care about it”. He continued to explain, “Our leaders would use it to inform their decisions on whether or not a teacher will be put in an important position”. The high importance attached to Zhibiao achievement led participants to associate their self-esteem with their students’ Gaokao performance. A strong Gaokao performance by students would boost their teachers’ sense of achievement and self-esteem (having a higher opinion of themselves). Conversely, if participants failed to meet their Zhibiao, “I would feel losing my ‘face’” (Ci-lei) and “be very low for several days, unwilling to stay where there are crowds of colleagues, and try to avoid these topics. It makes me less confident with a lowered self-esteem” (Ri-ling). As students’ Gaokao performance affected social recognition and teacher self-esteem, participants had to face external pressure to reach Gaokao Zhibiao and to compete with their teaching colleagues. This external pressure further built up self-imposed pressure to improve student Gaokao performance.

Teachers had the goal of improving students’ Gaokao performance, and participation in NMET rating was considered by most participants as beneficial for this purpose. Pi-li observed, “Going through the rigorous Gaokao rating [training] has helped the teachers improve themselves and their teaching [-to-the-test] practices. . . . In our school they are playing a leading role in guiding test preparation training practices and boosting Gaokao scores.” Although all teacher participants recognized the benefits of participating in NMET rating, only participants from three schools gained solid school support; the principal of the fourth school held a reserved attitude towards supporting teachers to participate. This fourth school excelled in Gaokao performance and outperformed its neighbouring schools but the school’s principal thought that teacher participation in Gaokao rating had limited impact on their school’s Gaokao performance,
hence he felt it was not worth much of the school’s attention. However, he was still willing to support teachers in participating in NMET rating out of the political demand from superior authorities. In spite of this varied level of school support for participation in NMET rating, most participants held positive attitudes towards this professional opportunity. Participants also described Gaokao rating experience as beneficial after experiencing it in terms of its potential “positive” impact on their teaching practices. Although participants felt stressful when rating NMET essays, most participants started to feel that participating in the NMET ERA was a beneficial experience overall. To illustrate this transition in his perception, Ridoong cited one line from a movie The Shawshank Redemption, forming an analogy between prison and his experience with NMET ERA: “At the beginning, you cursed it and hated it. But at the end, you find this has become part of your life. At this moment, you find yourself cannot part with it and have been attracted to it”. Despite the stress he felt at the beginning, Ridoong enjoyed this rating activity by the end of the rating session as he had learned a lot from this experience. Ridoong’s perception was shared by other participants. Tizhong also wrote in his own reflections, “By the end of the rating session, many teachers told me they thought the experience was rewarding. . . . No matter who participated in NMET rating, the benefits you have received are invisible and cannot be calculated with figures.”

This “rewarding” perception was associated directly with participants’ perceived potential impact of NMET rating experience on their teaching practices, by which they meant teaching-to-the-test practices that aimed for higher NMET scores. After rating thousands of essays from around the province over 9 days, these participants felt they had obtained rich, first-hand information about good and poor NMET essays. In addition, they also had the opportunity to exchange information with teachers from other schools about their teaching practices. By the
end, all the participants from high schools believed that what they learned from the NMET rating experience would exert or had exerted “positive” impact on their teaching-to-the-test practices. As experienced raters and team leaders, Ti-zhou and Ti-zhong would apply what they learned from NMET ERA participation each year to their classrooms. Ti-xian reported that he could find common problems his students had after rating NMET essays, “This makes my teaching [for NMET preparation] more targeted”. Like Ti-zhou, Ti-zhong, and Ti-xian, many participants reported a “positive” influence on their teaching-to-the-test practices after experiencing NMET rating. In this sense, the NMET rating experience seemed to have enhanced participants’ teaching-to-the-test practices, which often leads to an undesirable narrowing of the curriculum and is considered as negative washback in the washback scholarship (Alderson & Wall, 1993; Au, 2007; Cheng, Sun, & Ma, 2015; L. Qi, 2005).

This perceived benefit matched with participants’ incentive of learning from this rating experience to help improve their teaching-to-the-test practices at school. To facilitate their professional learning, they adopted two strategies. First, raters summarized and categorized frequent errors of strong and weak NMET essays, and reflected on how to better train their students to boost students’ NMET essay scores. For example, when rating NMET essays, Ri-zhen noticed essays that were organized into paragraphs and only contained relevant details tended to receive higher scores; he also enjoyed reading essays written in plain language but with proper use of transitions. He later translated these thoughts into his teaching practices, and taught his students strategies on how to write proper NMET essays, “In this way, they will make few errors and their essays contain shining points, which will catch raters’ eyes and tend to receive higher scores.”

Participants also used the opportunity to communicate with teachers from other schools,
through face-to-face conversation, social media (e.g., WeChat, a popular texting and voice messaging communication service in China), or shared documents. During breaks or after work at the end of the day, like many raters, Ri-juan would discuss certain topics with other teacher raters. “I asked them how their students responded to certain types of writing tasks. We often found differences between students from different schools, sometimes it was due to differences in teaching philosophy” (Ri-juan). Besides talking face to face, Ti-zhong’s, Ti-jian’s, and Ti- ren’s teams created their own WeChat group during the rating session, where raters were often involved in heated discussions after work. “More often than not, we exchange ideas and talk about our reflections on our teaching practices” (Ti-zhong). Teachers’ communication was also achieved through the use of shared documents. In the past few years, at the end of the NMET rating session, one team leader was responsible for giving a talk to all the raters. His presentation included a summary of test-takers’ frequent errors and a reflection on how to train students on the NMET essay. Some participants kept a copy of his talk slides and incorporated his experience into their own teaching and test preparation.

Consequently, teacher participants (both first-time and returning NMET rating participants) were all able to relate their NMET rating experience to their teaching-to-the-test practices. The first-time raters reported a change in their understanding of the NMET (and English) writing instruction. Through participation, they gained a “deeper” understanding of the NMET writing test construct. They were able to develop specific teaching plans that better accommodated their students’ learning needs and were potentially more effective in helping students to improve their NMET scores. For returning raters, they found repeated participation still contributing to their teaching-to-the-test practices.

Prior to participation, some first-timers seemed to be unclear about the NMET writing test
construct and claimed not to understand the differences between the NMET scoring bands until after the NMET rating experience. For example, Ri-xiu described how he used to rate his students’ essays, “I used not to place students’ essays to different score bands. I assigned certain points to each major detail and deducted points accordingly for lack of major details. I never thought about placing essays to different bands”. Therefore, it was not surprising that many participants claimed they used not to know how to teach NMET essay writing effectively. Their prior teaching strategies included asking students to “make sentences with given sentence patterns or phrases, translate [Chinese] sentences [into English], and describe what you [students] have done today” (Ri-dong), without any specific purposes or goals to achieve. Some participants trained students through writing NMET mock writing tests, and would provide students with corrective feedback or “sample essays” (usually provided by mock test designers for reference) for students to study by themselves. The feedback rarely explained what was good and not good about students’ writing, nor did it provide constructive feedback on how to improve their essay writing skills. Due to this lack of guidance and support, in spite of frequent essay writing practices, their students’ writing skills received little improvement: “Students seemed not to know how to write good essays” (Ci-su).

Through participation, these first-timers seemed to have gained a “deeper” understanding of the NMET writing test construct. Not only did they better understand the fine differences between score bands and later were able to continue applying the NMET rating criteria to assess their students’ essays; by studying and summarizing features of good and poor NMET essays, they gained a better understanding of what makes a “good” NMET essay (i.e., essay that tends to receive a higher score). According to participants, the features of a good NMET essay included: neat handwriting, covering all major details without irrelevant details, writing in clear and
concise language, organizing into paragraphs, using transitions, and advanced-level vocabulary and sentence patterns. Based on the NMET writing test construct identified by L. Qi (2007), most of those features match with the intended test construct, except for neat handwriting that goes beyond legibility concerns. Moreover, these features focused mostly on linguistic features, with no explicit consideration of communicating ideas with the supposed reader as required in the writing task (part of the intended construct for NMET essay writing). However, by suggesting students to pay attention to those features, participants held an instrumental motivation and attempted to train students how to effectively communicate messages to the imaginary NMET raters rather than the supposed reader, so that students’ essays were more likely to receive higher scores. As Ri-xiu reflected on his NMET rating experience, “when you [the rater] are faced with so many essays to mark [with time constraints], it is impossible for you to read word by word. Under such circumstances, how should our students write?” This experience prompted him to teach students how to impress raters, “[If] your essay reads well from beginning to end, putting the rater in a good mood, then your score cannot be low.” Consequently, participants not only taught their students to attend to those linguistic features; some started to teach students how to plan before writing, such as analyzing the writing task and major details to be included, highlighting key points with appropriate details, and organizing the essay into several paragraphs with the help of transitional devices. Therefore, the NMET rating experience helped participants to acquire a “deepened” (both positively and negatively) understanding of the NMET test construct.

In addition, this rating experience helped the first-time rater participants to better understand their students’ needs and develop more targeted teaching plans accordingly. Reading several thousand essays written by students provincewide allowed participants to gain a better
idea of the strengths and weaknesses shared by high school students. Based on such information, participants were able to develop specific plans guiding their students to improve English essay writing skills and subsequently NMET essay scores. For example, Ri-dong said, “After rating so many NMET essays, I find many students cannot write complete sentences, or cannot accurately use, what we think, very simple expressions”. The NMET rating experience inspired him to come up with a three-step plan for writing skill training. “It would be more effective if I start from the very basic and [train my students to] do it step by step” (Ri-dong). Due to all these gained benefits, participants all felt positive about their first-hand NMET rating experiences. Ri-xiu compared with his previous discussions with his teaching colleagues who participated in NMET rating before and commented, “It felt different when I experienced it on my own.” After this rating experience, he believed his teaching-to-the-test practices would be more targeted and purposeful, “I have a clearer understanding of towards which direction I should train my students to prepare for NMET writing”.

In addition to the benefits reported by the first-timers, returning NMET raters still found the NMET rating experience beneficial. Ri-ling’s NMET rating experience each year added to his collection of examples to use when training students, thus making his instruction more credible to students. Every year, Ti-zhou benefited from the opportunity to learn test preparation training strategies from teachers at other schools. “You can learn a lot of different things from teachers from different places.”

On the whole, NMET rating experience seemed to have brought about not just negative washback, but also positive washback on participants’ teaching practices. The negative washback included participants’ enhanced teaching-to-the-test practices that potentially led to the narrowed interpretation of NMET test construct, overemphasis on unintended test construct, and
instrumental interpretation of communicative writing purposes. On the other hand, NMET rating experience may have potentially exerted positive washback, with participants’ deepened understanding of the test construct and development of English writing instruction plans that better accommodated their students’ learning needs.

Besides this direct impact on teaching, NMET rating experience could also bring positive impact on teachers’ self-perception. With NMET rating experience, participants felt students would be more willing to take their teachers’ feedback seriously, because “they will write NMET in the near future and hope to obtain high scores” (Ti-xian). Furthermore, many participants claimed to have more authority over students when teaching and training their students to prepare for NMET. This enhanced sense of authority further contributed to participants’ sense of pride. In addition, NMET rating experience may also be seen as an advantage for teachers who have participated. Ri-dong explained, “Not every teacher at our school gets this [NMET rating] experience”. In others’ eyes, NMET rating was a mysterious, sacred, and privileged mission. “If you have participated once, . . .this experience would be a halo [around you]” (Ri-dong). Hence, Ri-ling commented, “Whoever [is recommended by school leaders and] comes [to the NMET rating session] must have a sense of pride.” As a result of both the benefits of enhanced self-perception and perceived “positive” impact on their teaching, most participants expressed their willingness to participate again in the future.

Section 2 Summary

In this section I presented findings of the social layer (raters’ object-oriented activity) of NMET ERA based on interview and document data. These findings provided a detailed description of the broader sociocultural context of Gaokao rating in China and the two settings (rating centre and schools) in which the NMET ERA was situated. This social layer added to the
previous cognitive layer (Section 1) by situating the NMET raters’ goal-directed decision-making actions against the background of raters’ object-oriented activity within the sociocultural context of the Chinese education system. In the following chapter, I will fully interpret and analyze the findings from both layers through the CHAT lens to illustrate how adopting a CHAT view can provide a deepened understanding of the research phenomenon.
Chapter 5 Discussion and Implications

This chapter discusses the major findings and implications of this study. I first revisit the purpose of the study and the findings I obtained. Then I synthesize the major findings through the lens of CHAT. Based on this synthesis, I answer my study’s three research questions by discussing and connecting the study findings with the existing literature. Finally, I discuss the implications of my study, its limitations, and directions for further research.

Summary of Findings

This qualitative case study examined the NMET essay rating activity (ERA) as a situated practice, with the cognitive (goal-directed actions) and social (object-oriented activity) layers of the NMET ERA within the sociocultural context of Gaokao. The analysis of the cognitive layer revealed that NMET rater participants tended to focus on three aspects of writing (content coverage, language quality, handwriting and answer sheet tidiness) and followed a sequential rating procedure when rating NMET essays (i.e., first deciding on a score band range based on an initial impression, then refining a score decision within the band range). Meanwhile, their scoring decisions were influenced by five factors:

- institutional requirements (rating scale and specifications, rater training, rating quality indicators in the on-screen marking [OSM] system);
- high-stakes consequences of raters’ ratings to student writers;
- saving “face” so as to stay close to others’ ratings;
- prior teaching and rating experience; and
- advice from colleagues.

Among these factors, saving “face” was highly pronounced and pervasive in this context, which is an indigenous concept that is frequently used to define appropriateness of interpersonal
relationships and social behaviours in China (K. Hwang, 1987). These findings from the cognitive layer were then situated in findings of the social layer in the context of where NMET raters performed their ERA at the rating centre and their related practices at the high schools where they taught. Findings from the social layer showed that, as one of the highest-stakes tests in China, Gaokao has a considerable impact on Chinese society, and the rating administration and results often draw nationwide attention. In this context, NMET raters held mixed feelings towards their NMET experiences. On one hand, they thought the NMET ERA they participated in was a sacred mission with grave responsibility, and were under high stress due to the challenge of meeting the rating centre requirements and the pressure to stay within the interrater agreement of their peers to save their “face”. On the other hand, raters thought their NMET ERA experience was beneficial to their teaching practices, where one of their priorities was to improve student NMET performance. Findings from the social layer contributed to the understanding of the sociocultural context of NMET ERA under study. Findings from the two layers together indicated that interpreting raters’ decision-making actions against the context where their ERA took place can help make better sense of NMET raters’ essay rating.

**CHAT Analysis of Findings**

To further understand the findings reported in the previous chapter within a larger sociocultural context, I interpret the study findings by mapping them onto a CHAT analysis. From the lens of CHAT, the NMET ERA was a result of interactions between the two activity-related settings (schools and rating centre). This activity consisted of three components: the context prior to raters’ participation in NMET ERA (at high schools), the central NMET ERA (at rating centre), and the consequences following raters’ participation (at high schools). I conducted this CHAT analysis to better understand interactions within the central activity and between the
two settings. This section presents the three components chronologically (i.e., preparticipation context, central NMET ERA, and postparticipation consequences).

**Preparticipation Context (at High Schools)**

The group of teachers from the four high schools who volunteered to be NMET essay raters were the key participants of the central NMET ERA; and not all the teachers at their schools were willing or provided with the opportunity to participate. Therefore, it is important to understand what constituted the preparticipation context at these high schools and what motivated the participating teachers.

These four high schools were situated within the context of Gaokao’s impact on high school education in China where student Gaokao performance, to a large extent, determines the success of a high school and a teacher’s career. Gaokao exerts such an enormous impact on high school education that Gaokao test syllabi are treated as curricula and influence classroom teaching and learning, which is a phenomenon discussed in the literature (e.g., Cheng & L. Qi, 2006; Liu, 2010; L. Qi, 2007). Correspondingly, the success of these high schools rested on student Gaokao performance. L. Qi (2005) also reported that the selection role of Gaokao was so strong that high schools in China were commonly judged by students, parents and local education departments according to the school graduates’ Gaokao performance. In addition, McNamara and Roever (2006) reported an analogous influence of the No Child Left Behind act in the United States, where there was an all-or-nothing proposition for schools called Adequate Yearly Progress (AYP). AYP stipulated the accumulative yearly proficiency quota of a school for its entire student body to meet the state performance standards in prescribed academic subjects. A school continuing to not meet AYP would end up facing an escalating series of sanctions, the worst of which could be school restructuring. In my study, the evaluation pressure from
education departments was found not an overriding concern by the principals as identified by L. Qi (2005). This was probably because of the recent school evaluation reform by the Ministry of Education (MoE), which aimed to shift from an exam-oriented to a well-rounded education program. Accordingly, MoE promoted a change in school evaluation criteria from student Gaokao performance toward efforts and achievement in fostering student comprehensive qualities (MoE, 2013). Compared with pressure from superior administrative departments, these four high schools were faced with more evaluation pressure from students and parents, and from competition with other schools in attracting high-achieving students.

Due to the notable impact of Gaokao on education and high schools, a major task for these high schools was to support teachers in the goal of improving student Gaokao performance. Student Gaokao performance was important to a teacher’s career, related to recognition of teachers’ capacity and their sense of self-esteem and “face”. Teachers’ responsibilities included, in addition to completing daily teaching responsibilities, training students for better Gaokao performance and meeting Gaokao Zhibiao (the targeted number of students meeting Tiers One and Two universities’ cut-off scores) set by the high school administrations to teachers. In spite of those requirements, the teaching resources in those schools seemed insufficient in supporting these teachers’ English (and NMET) writing instruction. Many participants contended that prior to their participation in NMET rating they were not clear about the NMET writing test construct and did not know how to teach students English writing effectively, nor did they have good strategies to improve students’ NMET essay scores. These participants faced both external and internal pressure to improve students’ NMET performance and meet the Zhibiao set by their schools. Participants claimed that it was a common practice for principals, teaching colleagues, and students to judge a teacher’s capacity by his or her students’ Gaokao performance. Zhibiao
achievement was considered an indicator of a successful teacher. Hence teachers worked hard and “compete[d] with each other” (Ci-hai) to improve student Gaokao performance and their Zhibiao achievement, which was a matter of self-esteem and “face” to them. This was consistent with L. Qi’s (2005) findings that Chinese high school teachers felt their students’ Gaokao performance would affect their professional life in various aspects, such as their status among school administration and colleagues, promotion opportunities, popularity with students and parents, their self-evaluation and sense of achievement. In this study, the inadequacy of available teaching resources required teachers to use additional resources to alleviate pressure from school goals for student Gaokao performance. Meanwhile, there was political demand from local departments of education, asking their schools to select experienced teachers and support their participation in NMET rating. The school principals held varied attitudes towards potential benefits of teacher participation to improving their schools’ Gaokao performance. Nevertheless, these teachers deemed participating in NMET rating as a professional development opportunity, which was their “only access to authentic NMET essays and makes up for the shortage of [NMET preparation] resources” (Ri-ling), hoping to become familiar with NMET trends and how they ought to guide their students accordingly.

As a result, these qualified teachers volunteered to be NMET essay raters because, among other reasons, they believed that NMET rating experience may help to improve their teaching-to-the-test practices and increase students’ NMET scores, potentially helping to achieve the Gaokao Zhibiao assigned to them. Most participants claimed that experiencing NMET rating was an essential step in their teaching careers.

This preparticipation context at high schools accounted for the subject (i.e., the volunteer teachers) of central NMET ERA who brought their teaching experience in the tool component,
and their incentive of professional development in the object component of central NMET ERA. This analysis indicates that in this NMET context the influence of raters’ teaching on their essay rating extended beyond the potential influence of raters’ teaching experience on essay rating, as discussed by previous studies (e.g., Barkaoui, 2010a; Cumming, 1990; Eckes, 2008). The influence also included raters’ object, or incentive, of professional development to participate in NMET essay rating, which has attracted little attention in previous research.

**Central NMET ERA (at Rating Centre)**

At the rating centre, the central NMET ERA system is represented in Figure 15. This activity system captures the participation of the 13 teachers who came from the four high schools and volunteered as NMET essay raters. Therefore, the subjects in this activity system are these 13 NMET essay raters from those four schools, including the eight essay rater participants and their five teaching colleagues with prior NMET essay rating experience. There was a wide range of mediating tools available to these raters when they rated NMET essays, including

- resources provided by the rating centre, such as NMET essays, NMET writing tasks, rating scale and specifications, rater training (including anchor essays and team leaders), and rating quality indicators in the OSM system; plus
- factors that affected raters’ decision-making, such as possible scores assigned by a second rater, teaching experience, prior NMET rating experience (applicable to returning NMET raters), and advice from teaching and rating colleagues.

Raters intended to achieve three objects while participating in this activity: (a) *rating with high accuracy and speed*, as expected by the rating centre and related to raters’ concern of saving “face”; (b) *being accountable for student writers*, due to their awareness of high-stakes consequences of their ratings; and (c) *acquiring professional development*, which was one of the
Tools:
Essays; writing task; rating scale and specifications; training (e.g., anchor essays, communication with team leaders); rating quality indicators; possible scores assigned by a second rater; teaching and rating experience; advice from colleagues

Subject:
NMET essay raters

Object (Goal):
- Rating accurately and fast
- Holding accountable for student writers
- Acquiring professional development

Outcome:
- Maintaining rating accuracy and speed that met institutional requirements and ensured accountability
- Collecting information to inform future teaching and test preparation training practices

Rules:
(Institutional)
- Rating essays based on the rating criteria
- Maintaining high rating accuracy and speed

(Sociocultural)
- High-stakes nature of Gaokao to student writers
- Saving “face”

Community:
- Peer essay raters
- Team leaders
- Directors

Division of labour:
- Raters marking essays by following institutional rules
- Raters coping with rating difficulties not covered by training and rating scale and specifications
- Directors and team leaders providing support

Systemic tension:
a) Rating essays based on inadequate rating criteria, while maintaining high rating accuracy and speed (i.e., tension within rules)
b) Rating essays based on inadequate rating criteria, while maintaining high rating accuracy and speed (i.e., tension between rules and object)
c) Trying to maintain rating accuracy and speed, while following rules and coping with difficulties
d) Following rules and coping with difficulties, while trying to treat all students equitably
e) Dealing with insufficiently provided resources, while trying to maintain rating accuracy and speed
f) Struggling to maintain rating accuracy and speed, while keeping “face”

Figure 15. Central NMET ERA system at rating centre
subject’s claimed incentives to participate.

As Figure 15 indicates, four rules guided these teacher raters in this central activity. These four rules included two institutional requirements and two sociocultural rules in this NMET ERA context. The two institutional rules included that raters (a) need to follow the rating scale and specifications when rating essays; and (b) should maintain high rating accuracy and speed. Similar institutional rules can be found in previous literature on essay rating (e.g., Lumley, 2005). The two sociocultural rules included (c) the high-stakes nature of Gaokao to test-takers; and (d) saving “face”. High-stakes consequences of rating results to test-takers have been reported as occasional concerns by some raters in high-stakes essay rating (e.g., Baker, 2010; Lumley, 2005; Smith, 2000). The high test stakes of Gaokao, however, were reported to be a pervasive concern in this NMET context that affected all the raters, hence deemed as a sociocultural rule here. The second sociocultural rule of saving “face” was highly pronounced in this NMET context. “Face”, or a person’s social self-image, not only is a means of social interaction through which social relations are formed, but also functions as an objective of social relations (X. Qi, 2011). “Face” reflects a personal view of one’s accomplishments and social recognition of a person’s position within a social structure (Earley, 1997). In my study, the matter of “face” involved judgements of a rater’s capacities to perform the NMET rating task and of the rater’s social standing relative to other raters, which was related to a sense of the rater’s fulfilment of obligations (e.g., rating accuracy and speed) to him or herself as a member of the NMET essay rating community (X. Qi, 2011). Evaluation of a rater’s “face” involves both internal (self) and external sources (social others; Earley, 1997), and the interaction between the rater’s own perception of his or her social self-image and the perception that other people in the rating community form of the same rater’s social standing (X. Qi, 2011). That explains why
participating raters reportedly kept an eye on their own statistics (e.g., valid rating rate, number of completed ratings) and on other raters’ when evaluating “face”, and they were also aware of being “watched by others” (Ri-dong). “Face” is quantifiable in the sense that gaining or losing “face” amounts to an addition or subtraction from an existing stock of “face”; and “face” can also be threatened or saved in social interactions (X. Qi, 2011). For example, when raters saw that their valid rating rates or rating speed were lower than their colleagues’, they felt their “face” was threatened; when higher than others’, they felt themselves gaining “face”; when their valid rating rates were below the bar or they were summoned by the rating quality inspection team leader due to unsatisfactory performance, they felt their “face” was lost. As “face” is subject not only to external judgement but also to internal assessment, this “face” evaluation often leads to subsequent, varied emotions within “face” holders: When a person maintains or gains his or her own “face”, the person may feel a sense of confidence, pride, and satisfaction; strong feelings of shame and inferiority often accompany losing “face”; when a person saves his or her own “face” from “face”-threatening incidents, the person is likely to feel a sense of relief and security (Goffman, 1967; X. Qi, 2011; Redding & Ng, 1982). In my study, participants’ performance also induced such variation in their emotions: raters felt “proud” (Ri-ling) when gaining “face”; “nervous” (Ci-su) when their “face” was threatened; and “embarrassed” (Ri-dong) when losing “face”. The power of “face” influences not only a person’s self-appraisal but also the person’s behavior, motivating the person to enhance or change his or her “face” states (X. Qi, 2011). In order to enhance or save their “face”, raters would then try to improve their valid rating rates and rating speed: “[I] kept thinking about how to increase it [valid rating rate]” (Ri-juan); “No one wants to lag behind [others] and be the slowest [rater]” (Ri-wei). “Face” was a means through which raters interacted with others in the rating community and saving “face”
oftentimes created an incentive for raters to behave in certain ways; therefore, saving “face” acted as another sociocultural rule in the central ERA. In brief, the four (two institutional, two sociocultural) rules together influenced raters’ decision-making and hence dictated their activity.

The community to which these raters belonged was the essay rating community at the rating centre, including their peer essay raters, essay rating and quality inspection team leaders, and directors at the rating centre. The division of labour within this community was shared responsibilities among participants in this rating community:

- Raters were responsible for marking essays by following institutional rules while coping with rating difficulties not covered by the rating scale, specifications, and rater training.
- Essay rating team leaders’ responsibilities included clarifying and elaborating the rating criteria, standardizing raters’ interpretation and application of those criteria, and making judgements to help raters solve rating difficulties.
- The quality inspection team leader’s responsibility included monitoring rating quality and talking with raters whose performance was unsatisfactory to help them identify rating problems.
- Directors were responsible for providing administrative support for the entire rating activity, ensuring high overall rating quality and completion of the NMET rating task within the required days.

The conditions of this activity system brought forth six tensions within and among different components, as illustrated in Figure 15. Within the component of rules, the two institutional rules were in conflict with each other, forming the first, circular tension (a) between rating essays based on inadequate rating criteria and maintaining high rating accuracy and speed. Raters were required to mark essays based on the stated criteria, despite encountering numerous
eventualities that did not fall within the stated criteria. All participants described their struggles
with matching certain essays to vague band descriptors when they applied the rating criteria.
Raters’ rating accuracy and speed were affected by the inadequately defined rating criteria, which
created “a lot of pressure” (Ci-ping). This tension caused by the conflicting rules further
triggered a second tension (b) with these raters’ object, that is, trying to maintain high rating
accuracy and speed while working with insufficiently defined rating criteria. Stemming from the
first tension (a) described above, all participants claimed that they tried hard to maintain rating
accuracy and speed, which was a self-generated goal associated with institutional requirements
and “face”-saving concerns. This tension generated third and fourth tensions between raters’
division of labour and their two objects—maintaining high rating accuracy and speed, and being
accountable to students—labelled respectively as tensions (c) and (d). Raters’ responsibilities
required them to generate self-developed strategies (e.g., studying the rating criteria more
carefully, examining anchor essays, guessing possible scores assigned by other raters) to cope
with rating difficulties that the stated rating criteria failed to cover. These self-developed
strategies may have weakened raters’ attempts to maintain high rating accuracy and speed and to
be accountable for student writers. Meanwhile, a fifth tension between tools and one of raters’
objects was (e) dealing with insufficient resources provided by the rating centre (e.g., rating scale
and specifications, training) while trying to maintain high rating accuracy and speed. This
tension further heightened the second tension (b) and caused a sixth tension (f) between the
object of maintaining rating accuracy and speed and the sociocultural rule of saving “face”.
Participants expressed their concerns of losing “face” in front of others if they failed to keep high
rating accuracy and speed. These six systemic tensions illustrated, from a CHAT perspective,
why participants felt stressed and bore in their minds the “sacred mission and grave
responsibilities” they took on. The tensions unravelled the rating difficulties and challenges these raters were faced with when rating NMET essays. This analysis showed that in this NMET context, the tensions raters felt were more complex than the tension identified in previous studies (e.g., Lumley, 2005) where raters struggled to balance between their intuitions of essay quality and the institutional rules of following the rating scale and maintaining high rater reliability.

In an attempt to alleviate the above six tensions in Figure 15, raters not only employed different mediating tools provided by the rating centre, but also brought in additional mediating tools to attain their object, both of which accounted for those factors cited by raters that affected their decision-making. Table 10 summarizes how raters addressed each of those six tensions. As the six tensions were concurrent and intertwined, some mediating tools were used as solutions to multiple tensions.

To alleviate the first, circular tension (a) between the two conflicting institutional rules, raters adopted various mediating tools in hopes of discerning the fine differences within and between each score band range while maintaining high rating accuracy and speed. These tools included anchor essays, where they studied essay features and used the scores assigned by expert raters as a source of reference. Raters also consulted their team leaders to clarify vague descriptors or asked team leaders to help make final score decisions. In addition, raters also drew on their prior teaching or rating experience in an attempt to discern the fine differences between band descriptors. Lastly, some raters regarded their rating quality indicators as references to identify the extent to which their scores deviated from the norm, or where most raters seemed to be, using hypothetical scores assigned by a second rater as their point of reference.

The second tension (b), the one between the conflicting institutional rules and the object of maintaining high rating accuracy and speed, was relieved by raters drawing on prior teaching
Table 10

*Tensions and Raters’ Corresponding Solutions*

<table>
<thead>
<tr>
<th>Tension</th>
<th>Solutions</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) Rating essays based on inadequate rating criteria while maintaining high rating accuracy and speed (i.e., tension within rules)</td>
<td>(a) Anchor essays, consulting team leaders, teaching and rating experience, rating quality indicators, possible scores by second raters</td>
</tr>
<tr>
<td>(b) Rating essays based on inadequate rating criteria while maintaining high rating accuracy and speed (i.e., tension between rules and object)</td>
<td>(b) Teaching and rating experience, advice from colleagues</td>
</tr>
<tr>
<td>(c) Rating essays by following the institutional rules and coping with undefined difficulties, while trying to maintain high rating accuracy and speed</td>
<td>(c) Teaching and rating experience, advice from colleagues, rating quality indicators</td>
</tr>
<tr>
<td>(d) Rating essays by following the institutional rules and coping with undefined difficulties, while trying to be accountable for student writers and treat them equitably</td>
<td>(d) Teaching and rating experience, advice from colleagues, comparing essays against each other, rating quality indicators, possible scores by second raters</td>
</tr>
<tr>
<td>(e) Dealing with insufficient resources provided by the rating centre (e.g., rating scale and specifications, training) while trying to maintain high rating accuracy and speed</td>
<td>(e) Teaching and rating experience, advice from others</td>
</tr>
<tr>
<td>(f) Struggling to maintain rating accuracy and speed, while making sure they did not appear less competent than others and their “face” was saved</td>
<td>(f) Rating quality indicators, advice from colleagues, possible scores by second raters</td>
</tr>
</tbody>
</table>

and rating experience, and taking advice from colleagues, in an attempt to compensate for insufficiently defined rating criteria while achieving this object. Raters drew on personal teaching experiences to make judgements on essay language quality, and envisioned writers’ English proficiency levels by comparing them to their own students. Returning raters’ prior NMET rating experience also made ratings “faster and more accurate” (Ri-wei). In addition, raters applied their colleagues’ rating strategies and suggestions (such as rating sequence and
leniency), discussed their uncertain ratings, and learned from others how to deal with rating difficulties and rate fast and accurately.

To reduce the third (c) and fourth (d) tensions, the tensions between raters’ division of labour and two of their objects, raters made use of their prior teaching and rating experience, and colleagues’ advice, trying to deal with rating difficulties while meeting the two objects—keeping their rating accuracy and speed satisfactory and being accountable for students. Some raters also compared essays against each other to ensure their ratings were equitable. At the same time, raters used rating quality indicators to check their rating performance and to ensure they assigned reasonable scores. When they noticed their rating performance was going down, in addition to the aforementioned tools, they also considered possible scores assigned by a second rater to make sure they did not treat student writers unfavourably.

The fifth tension (e), a tension between the insufficient tools and the object of maintaining rating accuracy and speed, was closely related to tension (b) between the same object and the conflicting institutional rules. Hence, raters’ solutions were similar, including the use of prior teaching and rating experience as well as advice from their colleagues to compensate for the inadequacy of stated rating criteria and rater training while making sure their performance remained satisfactory in terms of rating accuracy and speed.

To relieve the sixth tension (f), the tension between the object and the sociocultural rule of saving “face”, raters paid close attention to their rating quality indicators in the OSM system, which were shared information within their rating teams. To ensure their indicators were satisfactory, raters communicated with other colleagues, hoping to learn from their experiences how to rate quickly with accuracy. But also, some raters compromised their own judgements on essay scores and thought about possible scores assigned by a second rater before assigning a
middle ground score that was more likely to be valid. Either way, raters were trying to maintain a good record on rating quality indicators and making sure their “face” was saved.

As illustrated above, the six tensions in Figure 15 were closely associated with each other, therefore leading NMET raters to apply the same tools or resources to solve different tensions, and sometimes different tools for the same tension. I unpacked the above tensions and their corresponding solutions one by one. However, in reality the situation was concurrent, interactive, and mingled. For example, one mediating tool, raters’ personal rating quality indicators (e.g., valid rating rates, serious rating error rates), was shared information within their rating teams. Raters regarded those indicators as a way to identify whether or not their rating performance was deviant from the norm. If their indicators were not good enough, it triggered three types of raters’ concerns. The first concern was that their ratings failed to meet the institutional requirements. Their second concern was the possibility of having assigned inequitable scores to students. The third type of concern was that they may look less competent than other raters, a potential threat to their “face”. These concerns were associated with the first two of raters’ objects and the institutional and sociocultural rules outlined in Figure 15. Consequently, these concerns would lead raters to adopt various mediating tools and try to keep their ratings close to their peer raters. The tools they adopted included consulting team leaders, communicating with other raters, and sometimes thinking of possible scores assigned by second raters. This CHAT analysis of how raters solved rating tensions revealed a far more complex interaction between raters’ cognition and the context, whereas previous research on essay rating provided few insights (e.g., Lumley, 2005).

The outcome of this activity system was that (a) most raters managed to maintain high rating accuracy and speed that satisfied institutional requirements and ensured accountability to
student writers; and as a result of their 9-day intensive NMET rating experience, (b) raters felt they obtained rich information and experiences to inform their future teaching and NMET preparation training practices. The outcome (b) met these teacher raters’ expectations for participation, and explained why participants considered the NMET rating a rewarding experience. It also provided subsequent resources for these raters’ postparticipation practices (including teaching) at schools.

This CHAT analysis of the central NMET ERA revealed the complex interaction between raters’ cognition and the Chinese NMET context when raters were rating essays, an underresearched area by previous essay rating scholarship that has been dominated by the cognitive approach. This CHAT analysis also indicated a potential influence of NMET essay rating on these NMET raters’ teaching practices upon their return to their respective teaching communities. I elaborate this influence in the following subsection of postparticipation consequences.

Postparticipation Consequences (at High Schools)

In addition to the two immediate outcomes indicated in Figure 15, the influence of central NMET ERA went beyond essay rating itself and led to three consequences to participating teachers’ subsequent practices at schools. These consequences included: the potential washback effect of NMET essay rating, an enhanced sense of authority and pride, and intention to continue participating.

The first consequence was concerning the potential washback effect of NMET essay rating. One of the outcomes of central ERA, information and experience obtained from NMET rating experience, potentially informed participating teachers’ future teaching and NMET preparation training practices. Such information was obtained through rating NMET essays and
communicating with teachers from other schools through various channels (e.g., face-to-face, WeChat, shared documents). During the 9-day rating session, teachers were exposed to several thousand NMET essays by students provincewide. This experience provided them with a big pool of student essays than before in their schools with their own students. These teachers studied good (and poor) NMET essays, which allowed them an opportunity to reflect on their own English writing instruction. They exchanged information with raters from other schools, which served as another source of information. As a result, participants all had positive perceptions of their NMET rating experiences. They seemed to develop a better understanding about the NMET writing test construct and their students’ learning needs, subsequently outlining specific teaching plans or goals focused on improving students’ English writing skills. This experience also appeared to enhance these teachers’ teaching-to-the-test practices. For example, teachers summarized various instrumental strategies of pleasing the assumed raters (e.g., neat handwriting, only including relevant major details, use of transitions), teaching students for possibly higher essay scores. Some of those strategies (e.g., neat handwriting, communicating effectively to the imaginary rater instead of the supposed reader required in the writing task), however, did not match with the intended NMET writing test construct. These reported teaching-to-the-test practices were similar to L. Qi’s (2005) findings on the negative washback of NMET essay writing test, where she found high school teachers and students had a narrow interpretation of NMET essay writing construct. L. Qi suggested that, among other reasons, the “insider” information from high school teachers with prior NMET essay rating experience might be responsible for this narrowed interpretation. These prior NMET raters revealed and shared some NMET essay rating criteria, such as handwriting quality, inclusion of major details, and linguistic accuracy, without stressing the intended communicative appropriacy as an evaluation criteria. As
they were considered “insiders”, what they said became valuable information for test preparation at high schools, misled or confirmed the other teachers’ narrowed interpretation of the NMET writing test construct. Although L. Qi examined the washback of NMET writing test as a whole, her findings implied a potential negative washback effect of essay rating on high school English writing instruction. My findings partially confirmed hers: The participating teachers planned to apply or have applied the abovementioned strategies they summarized to guide their teaching-to-the-test practices, with little attention to the intended communication purposes with the supposed reader, and participants all confirmed potential improvements in their students’ NMET writing performance. However, unlike L. Qi’s study indicating that high school teachers’ participation may only assist in high school teaching-to-the-test practices with negative washback, this postparticipation consequence suggested that in this study high school teachers’ participation also seemed to bring about some positive influences on teaching and learning, such as deepened understanding of the test construct and student learning needs.

The second consequence was that these participating teachers felt an enhanced sense of authority and pride after their NMET rating experiences. It was a common practice that participants shared with students and other teachers at their schools this experience and the collected information, which was also discussed in L. Qi’s (2005) study. These participants’ rating experiences turned them into NMET preparation authority in students’ eyes: Their students were more willing to take their NMET preparation advice. Being part of the sacred, mysterious NMET rating mission brought the participating teachers a sense of pride in front of their teaching colleagues and students. This second consequence in conjunction with the first consequence explained why participants considered NMET rating experience beneficial.

These first two consequences helped to a certain extent to relieve participants’ external and
internal pressure of improving student Gaokao performance and Gaokao Zhibiao achievement prior to participation. Teachers who saw the benefits of participating hoped to further reduce the pressure through repeated participation. As a result, most teachers expressed their intention to continue participating in NMET rating if offered the opportunity again, which was the third postparticipation consequence of NMET ERA. This analysis revealed direct connections between NMET essay rating and high school teaching and learning in the NMET context, which has been previously underresearched in the washback scholarship.

Summary

This CHAT analysis synthesized the major findings from Chapter 4 and provided a more situated understanding of the NMET ERA within the context of high-stakes testing in China, which was a result of interactions between the two settings (high schools and provincial rating centre). This NMET ERA consisted of three components, which were presented chronologically:

Preparticipation context (at high schools) → Central NMET ERA (at rating centre) →
Postparticipation consequences (at high schools).

This way of presentation provided a broader view of the sociocultural context related to the NMET ERA, highlighting the relationship between participants’ activity and practices at the two settings, not only in terms of sequence of happening, but also their interactions. As shown in Figure 16, the NMET ERA situated within the broader context of the Chinese education system was made up of a central activity system (Figure 15) at the rating centre and its related context and consequences at the high schools. This central activity system was filled with systemic tensions. These tensions linked the cognitive layer with the social layer of NMET ERA, and deepened the understanding of the complexity of NMET raters’ decision-making, where raters
Figure 16. Relationship between the two activity-related settings (high schools and rating centre)
kept interacting with the sociocultural context. These tensions, alongside participants’ pre- and post-participation practices, were the forces driving the development of NMET ERA. Links existed between the two settings. In the preparticipation context at high schools, teachers who were willing and qualified to participate later became NMET essay raters, which provided the subject for the central ERA. Their incentive to participate included obtaining professional development, which fed into the object of central ERA. These teachers’ teaching experience fed into the tool component in the central ERA, because teaching experience was one of the NMET rater selection criteria. In the central ERA at the rating centre, these teachers’ object of professional development was fulfilled, and the outcome of the activity was the acquisition of experiences and information that potentially shaped participants’ future teaching practices at high schools. Apart from the immediate outcomes, the central NMET ERA led to three consequences to participants’ practices at high schools afterward: the potential washback effect (both positive and negative) of NMET essay rating on high school teaching and learning, participants’ enhanced sense of authority and pride, and their intention to continue participating in NMET rating. This CHAT analysis provided a more situated understanding of NMET essay rating by revealing its cyclical interaction with practices at high schools and its relationship with teaching and learning within the context of Chinese education system.

Discussion

To better understand the relationship between study findings and existing literature, the discussion is organized by my study’s three research questions.

RQ1: How do raters assess NMET essays to achieve their goals?

The findings reveal a surface structure and a deep structure concerning how raters assessed NMET essays to achieve their goals. I borrowed the linguistics terms “surface structure” and
“deep structure” from Chomsky (1965), who postulated the relationship between phonetic forms and semantic meanings in the area of transformational grammar. In transformational grammar, deep structures are generated by phrase-structure rules that form the basis of abstract (semantic) meanings, while surface structures are specific sentences (phonetic forms) that express those meanings derived from deep structures (through a series of transformations). With this bi-level conception of grammatical structures, Chomsky attempted to explain why sentences with different phonetic forms may have similar semantic meanings, while similar phonetic forms may have different semantic meanings. His conception of the two-levelled structure resonates with how I see the relationship between the cognitive and social layers of ERA revealed in my study.

Table 1 summarizes and compares the characteristics of surface and deep structures in the NMET ERA. The surface structure reflects what raters’ decision-making looked like on the cognitive layer, including findings about NMET raters’ rating focuses, rating sequence, and factors influencing their decision-making. Such findings are based on data from think-aloud protocols, stimulated recalls, and interviews. The deep structure explains how and why raters’ decision-making worked in that way, by drawing on evidence on the social layer visualized as an activity system highlighting interactions between raters’ cognition and the activity sociocultural context. Its findings are based on interview and archival data. The surface and deep structures are interrelated through a series of tensions and raters’ corresponding solutions within the ERA system. This bi-level conceptualization can explain why raters may turn to the same tools (e.g., by attending to similar essay features, following similar rating sequence, and considering similar factors) to solve different tensions derived from interactions within the activity system, or adopt different tools to solve similar systemic tensions.
My findings concerning the surface structure showed that NMET essay raters appeared to follow a similar rating sequence (i.e., first deciding on a score band range based on initial impressions, then refining a score decision within the band range) while attending to three writing aspects in making their scoring decisions. Their decision-making was affected by factors including institutional requirements, high-stakes consequences to test-takers, saving “face”, prior teaching and rating experience, and collegial advice. These findings largely support the previous cognitive approach studies on essay rating. Those studies have identified a similar surface structure of rater decision-making, taking the form of rating focuses, rating sequence, and factors influencing decision-making, as reviewed in Chapter 2. For example, raters were found to focus on a wide range of writing aspects when assessing essays, such as content, grammar, handwriting, organization, coherence, task realization (e.g., Barkaoui, 2010a; Broad, 2003; Erdosy, 2005; Homburg, 1984; Lee, 2004; H. Li & He, 2015; Sakyi, 2003; Vaughan, 1991). As to rating sequence, some researchers found raters tended to employ different approaches or styles.
(e.g., DeRemer, 1998; Lukmani, 1996; Vaughan, 1991; Wolfe et al., 1998); some found raters may go through a similar process, either sequential (e.g., Crisp, 2012; Freedman & Calfee, 1983; Homburg, 1984), or complex and interactive (e.g., Lumley, 2005; Sanderson, 2001). Regardless of the rating sequence raters may employ, their decision-making primarily involved their interactions with three key elements in essay rating (writing task, essay, and rating scale), and hence was affected by factors related to raters themselves and variation in those three elements (e.g., Barkaoui, 2010a; Rinnert & Kobayashi, 2001; Weigle, 1999). At the same time, some studies started to notice raters’ concurrent interactions with some sociocultural factors, such as consequences of scoring decisions to test-takers, institutional constraints, test purposes, raters’ discourse community, and raters’ community of practice (e.g., Baker, 2010; Erdosy, 2005; Lumley, 2005; Mei & Cheng, 2014). There was one factor, saving “face”, uniquely identified in my study. This factor was highly salient in this NMET context, and an analogous yet different concern than to “play it safe”, the latter having been documented in a few studies (e.g., Knoch, Read, & von Randow, 2007; Myford & Mislevy, 1995). Their similarities and differences are discussed when I address RQ3.

My findings concerning the deep structure demonstrated that, according to the CHAT analysis of the central NMET ERA (Figure 15), NMET raters’ decision-making was not a fixed, linear process as the surface structure showed, but rather a complex, socially mediated activity by nature within the NMET context. These findings indicated the specificity of NMET raters’ essay rating in the Chinese context. For example, besides the rating centre’s institutional rules, raters’ decision-making was also dictated by the sociocultural rules (high-stakes nature of Gaokao and saving “face”) unique in this Chinese context. When rating NMET essays, these raters were driven by two goals (i.e., to maintain high rating accuracy and speed and to remain
accountable to students). When these raters were performing their goal-directed decision-making actions, they constantly interacted with the context while trying to alleviate six interrelated tensions coupled with these interactions. These findings exhibited a complex, dynamic relationship between rater cognition and the Chinese sociocultural context in these NMET raters’ decision-making. Such a relationship has not been identified in the previous literature.

My study expands on the existing literature and highlights the importance of adopting a sociocultural view to study the relationship between raters and sociocultural contexts in rater decision-making. The CHAT analysis reveals a “deep structure” of essay rating, which demystifies rater-context interactions and their relationships with rater decision-making and helps to understand what is hidden from the “surface structure” and what is behind raters’ cognitive functioning. The guiding CHAT framework allows this situated view and forms the basis to explore the rater-context relationship. For example, in addition to using think-aloud protocols (TAPs) and stimulated recalls (SRs) to examine raters’ decision-making processes, which alone is usually the study focus of essay rating research, I followed up with raters about any factor(s) they perceived having influenced their decision-making and any goal(s) they held when assessing NMET essays, instead of assuming raters held the same goal as required by the rating centre. The results demonstrated that these raters had their own goals when rating essays, which were associated with the institutional and sociocultural rules and their responsibilities specific to this NMET context. Also, interviewing raters about their interactions with the sociocultural context showed that instead of passively being affected by the context, these raters’ decision-making was both affecting and affected by the contexts, resulting in a series of interconnected tensions and raters’ subsequent solutions before they ultimately achieved their goals. Exploring such relationships generated a contextualized and more meaningful
understanding of NMET raters’ decision-making. Therefore, my study suggests that future research on essay rating should adopt a sociocultural view, instead of relying on cognitive evidence alone, which enables not only the examination of “what” but also the explanation of “how” and “why” in rater decision-making.

**RQ2: What are the broader (Chinese society) and immediate (rating centre and school) sociocultural contexts in which the NMET ERA is situated?**

The sociocultural contexts of NMET ERA in this study are summarized as follows.

At the Chinese societal level, Gaokao is recognised as so far a relatively fair means of selecting talented individuals for university admission, the decisions of which are based on test-taker’s Gaokao test results rather than on their socioeconomic status. Student Gaokao performance holds high stakes to all stakeholders, including test-takers, their families, teachers, and high schools. Gaokao can provide a life-changing opportunity for test-takers, and their Gaokao performance is often associated with family glory. Besides, student Gaokao performance is socially recognized as an indicator of a high school’s success as well as the success of a teacher’s teaching career. As a result, Gaokao rating gains attention and is highly regarded nationwide. The social impact of Gaokao is passed onto secondary education. Gaokao shapes teaching and learning in high schools, leading to pervasive teaching-to-the-test practices in these schools, and motivates teachers to participate in NMET rating to collect first-hand information about NMET essay writing and rating, in the hopes that such information would assist them in training their students for the NMET.

Bearing the above incentive to participate in NMET ERA at the rating centre, teacher raters were driven by two additional goals (objectives) when rating NMET essays. One goal was to rate essays fast and accurately, to meet the rating centre’s institutional requirements, without
appearing less competent than other raters in terms of interrater agreement and hence losing “face”. A second goal was to be accountable for students, associated with raters’ awareness of the importance of Gaokao and high-stakes consequences of their ratings to these students. To best achieve the three objectives, especially the latter two, these teacher raters made use of all resources available to cope with the challenges they encountered. These raters ultimately managed to achieve all three objectives.

Further, these teacher raters found their NMET rating experience beneficial to their careers. They felt that NMET rating experience deepened their understanding of the NMET writing test construct and their students’ learning needs, and helped them to form teaching plans directed at improving student writing skills. Meanwhile, their teaching-to-the-test practices were more targeted, resulting in their perceived improvement in student NMET writing performance. Moreover, they gained a sense of authority and pride, thus fulfilling their original objective to participate in NMET rating.

In sum, as the CHAT analysis of NMET ERA revealed, the broader and immediate sociocultural contexts were viewed as an activity system within the Chinese education system. This system consisted of three interconnected components (preparticipation context, central NMET ERA, postparticipation consequences) at the two activity-related settings, with the central activity composed of four mediators (i.e., tools, rules, community, division of labour) and inherent systemic tensions. This analysis uncovers the context where raters made their rating decisions and explains why raters assessed NMET essays in the way they did.

Although most prior literature on essay rating (e.g., Crisp, 2010; Homburg, 1984; H. Li & He, 2015; Vaughan, 1991) ignored interactions between raters and sociocultural contexts, some studies started to examine influences of the factors elicited from sociocultural contexts on rater
decision-making. These factors included test stakes (e.g., Baker, 2010), rater training (e.g., Elder, Knoch, Barkhuizen, & von Randow, 2005; Furneaux & Rignall, 2007; Lu, 2010; Weigle, 1998), and institutional constraints (e.g., Lumley, 2005). My findings confirmed that the above few factors were related to the sociocultural context affecting NMET raters’ decision-making. However, those factors should not be considered in isolation, because each factor not only contributed to the sociocultural context, but also interacted with each other and other elements within the context, altogether forming the sociocultural context where raters assessed NMET essays. This study hence embraces a broader view of the sociocultural context which goes beyond “the particular social milieu in which the assessment takes place (i.e., the school or institutional setting), which determines the goals of the assessment and the broader social and cultural context and relates to cultural norms about writing assessment” (Weigle, 2002, p. 60). Instead, adopting the CHAT perspective, my study regarded the social world outside NMET raters as the sociocultural context, which consisted of the activity system shown earlier in the CHAT analysis. As illustrated in Figure 16, the sociocultural context of NMET ERA included the complete activity system (with the three interconnected components at the two settings) within the Chinese education system. It included (a) raters’ preparticipation context at their schools that accounted for raters’ motivation to join, for professional development purposes; (b) the mediators in their central NMET ERA (various tools raters used, rating centre’s institutional rules and unique sociocultural rules, and raters’ shared responsibilities within the rating community) and the tensions generated by their interactions, while raters were striving to be good raters and collecting information to inform future teaching practices; and (c) raters’ postparticipation consequences upon return to their schools, where they intended to apply their NMET rating experience to their teaching practices, and which benefited them such that they
were inclined to participate again. This broadened view of sociocultural context assigns more meaning to how and why raters thought and acted in certain ways, and is therefore the key to understanding what was hidden beneath what Engeström (2001) called the “tip of the iceberg”.

**RQ3: What is the nature of NMET ERA as an activity system within the above sociocultural contexts?**

The NMET ERA was by nature a complex, socially mediated activity that involved both cognitive and social layers, where raters constantly interacted with the NMET sociocultural context when rating essays, as I explained in answering RQ1 and RQ2. The CHAT analysis highlights three key findings concerning the unique nature of NMET ERA in this study: roles of goals and rules, rating tensions and raters’ solutions, and the relationship between essay rating and teaching.

First of all, both goals and rules played important roles in the NMET ERA of study. Previous studies tended to assume—explicitly or implicitly—that raters held the rating goals of assessing essay quality based on prescribed evaluation criteria (e.g., Crisp, 2010; DeRemer, 1998) and maintaining high rater reliability as expected by the institutions (e.g., Lumley, 2005). This study found that NMET raters participated in NMET ERA with three goals or objectives to achieve, including (a) satisfying the requirements by the rating centre (as previous studies have assumed). In addition, raters pursued self-generated goals, such as (b) being accountable to students and (c) acquiring professional development. These three goals showed that raters’ NMET ERA was a purposeful activity and they were the key to understanding raters’ interactions with their rating and teaching communities within the NMET context. The raters’ first two goals, (a) rating accurately and fast and (b) being accountable, were related to their rating community, interacting with different components in the central ERA and contributing to the rating tensions
raters felt. The third goal (c) professional development was related with raters’ teaching communities, linking their teaching and rating communities together and supporting the development of NMET ERA from raters’ preparticipation context to the central NMET ERA, and later to their postparticipation consequences.

While the goal of (c) professional development originated from raters’ practices in their teaching communities, the other two goals—(a) and (b)—were closely related to the institutional and sociocultural rules that raters observed when carrying out the central NMET ERA. Previous studies also suggested that rater decision-making can be bound by certain rules (e.g., Baker, 2010; Lumley, 2005). Lumley observed that when assessing essays used for immigration decisions, raters tried hard to meet institutional goals of rater reliability in spite of various rating difficulties. That was one of the first studies that explicitly discussed the role of institutional rules in governing raters’ rating behaviours. Baker observed that when raters were invited to reassess essays for a teacher certification test, raters seemed to be preoccupied with being consistent with their previous ratings. She suggested that raters may feel the pressure of consistency due to perceived desirability of intrarater reliability and possible consequences to future employment. In support of those prior studies, my study confirmed the presence of rules dictating raters’ activity, and suggested that there may be multiple rules in play. My findings showed that NMET raters did more than follow the institutional rules established by the rating centre, as in Lumley’s study. As social beings, NMET raters also conformed to two sociocultural rules characteristic of the Chinese high-stakes testing culture, the high-stakes nature of Gaokao to students and saving “face”. The first sociocultural rule was associated with high-stakes consequences of ratings to test-takers in China, consistent with previous findings where researchers noticed some raters expressed such concern when scoring essays (e.g., Lumley, 2005;
Smith, 2000). My study explored further and found that the high-stakes nature of Gaokao was perceived by all the participants as one rule, because of the omnipresent impact of Gaokao in the country, on high school education, and on participants’ personal life experiences. This rule strengthened raters’ sense of accountability as NMET raters. Another sociocultural rule that was uniquely identified in this study was the rule of saving “face”. As the famous Chinese writer Lin Yutang (1936) commented, saving “face” is “the most delicate standard by which Chinese social intercourse is regulated” (p. 190). Lin’s comment remains true in China to this day. As a Confucian legacy of valuing “face” for morality and ideals of human relationship, the sociological “face” was identified as a major dimension of Chinese culture (Cardon & Scott, 2003), and a key to understanding Chinese spirit in almost every aspect of social life in China (X. Qi, 2011). “Face” was conceptualized as

The respectability and/or deference which a person can claim for himself from others, by virtue of the relative position he occupies in his social network and the degree to which he is judged to have functioned adequately in that position as well as acceptably in his general conduct . . . In terms of two interacting parties, face is the reciprocated compliance, respect, and/or deference that each party expects from, and extends to, the other party. (Ho, 1976, p. 883)

Although “face” has been well-acknowledged as a universal phenomenon of the social anchoring of self in the gaze of others (Ho, 1976; A. Hwang, Francesco, & Kessler, 2003; X. Qi, 2011), it has gained high salience in the Chinese society and in particular, this NMET context. The “face”-saving rule explained why raters cared about how their rating statistics looked like in front of others. They did their task not only to meet the institutional expectations, but also, if not more importantly, to avoid losing “face”. This concept of saving “face” is analogous to raters’ concern
to “play it safe”, the latter being identified by a few previous studies where raters were found to be afraid of being singled out under certain conditions (e.g., after training) and hence may demonstrate a central tendency stance (e.g., Knoch et al., 2007; Myford & Mislevy, 1995). In this Chinese context, however, the concept of saving “face” is richer and broader than playing it safe. In this study, when raters felt their “face” was threatened, they may have assigned a middle ground score to “play it safe”. Meanwhile, they also made great efforts and employed various resources (e.g., team leaders, anchor essays, colleagues’ advice) to learn how to improve their understanding of the rating specifications to facilitate their applications. Therefore, the “play it safe” concern can only account for part of the phenomenon, while saving “face” is a deeply-rooted sociocultural rule that dictated these Chinese raters’ behaviours and provided a fuller picture for the phenomenon. These findings indicate a need for taking into account all the rules guiding raters, especially those sociocultural rules unique in each context, for a comprehensive and more in-depth understanding of raters’ decision-making.

The second unique finding from this study was the tensions and raters’ corresponding solutions when raters made scoring decisions. The central NMET ERA (Figure 15) was filled with inherent conflicts and systemic tensions in the NMET context where the essay rating took place. Lumley (2005) commented that, as an idealized tool, any rating criteria must remain inadequate. In my study, this inadequately defined tool provoked the conflict within the two institutional rules (i.e., rating by following the rating criteria vs. maintaining high rating accuracy and speed), which further created a series of tensions in the central NMET ERA. Lumley identified a tension between the inadequate rating scale and the intuitive impression in rater decision-making, whereby raters had to reconcile their impression of an essay and the rating scale wording before producing a final essay score. My findings confirmed the presence of this
tension, and revealed that such a tension was masked by a more complex tension mechanism in rater decision-making that involved constant interactions between different mediators in the activity system.

This study not only identified how those tensions operated in NMET raters’ decision-making, but further explained how raters attempted to reduce the tensions, which has not previously been identified in the literature. When encountering tensions, raters turned to various tools (e.g., possible scores by second rater, prior teaching and rating experiences, advice from colleagues) to supplement the ones provided by the rating centre (e.g., rating scale and specifications, anchor essays, essays, writing tasks, team leaders, rating quality indicators) in order to resolve the tensions that interfered with their objectives. Raters either used the same tools to alleviate different tensions or used different tools to alleviate the same tensions. As a result, those inherent tensions were alleviated but never resolved, accounting for why those raters were still struggling by the end of their rating experiences, despite their satisfactory rating accuracy and speed evaluated by the OSM system. All those identified tensions and the corresponding solutions taken by raters provided a plausible explanation to what Lumley (2002) described as “a somewhat indeterminate process” (p. 246) of how raters resolve rating difficulties. These findings demonstrated the complexity of these raters’ decision-making and furthered the understanding of how raters’ cognition interacted with this unique testing context.

The third unique finding was the relationship between essay rating and teaching. Previous studies on essay rating examined this relationship and often concluded with a unidirectional relationship, that is, raters’ teaching experience affected their essay rating processes and/or results (e.g., Cumming, 1990; Eckes, 2008; Hamp-Lyons, 1989; Pula & Huot, 1993). In support of those findings about the presence of influences associated with raters’ teaching experience,
through CHAT analysis my study revealed a more interactional and dynamic relationship between essay rating and teaching in this NMET context. As shown in Figure 16, from their pre-participation context in their teaching communities, teachers who participated in NMET ERA brought not only their teaching experience to assist in decision-making, but also an incentive to participate for the sake of professional development, which was associated with the high importance of NMET to their teaching careers; then during the central ERA at the rating centre, these teachers not only completed the task of rating essays with the assistance of their prior teaching experience, but also collected information about NMET essay writing and rating to achieve their objective of professional development; from their NMET rating experience, these teachers brought the collected information back to their respective communities to inform their future teaching practices. These findings indicate that NMET essay rating shapes and is shaped by teaching in this testing-driven educational context, suggesting a potential washback effect, that is, a change in teachers’ approaches to teaching from pre- to post-participation.

The testing and teaching cycle is unique within the Chinese context as mentioned by previous washback studies (e.g., Cheng, 2005; Gu, 2007; L. Qi, 2005, 2010). This line of research usually focuses on examining the impact of testing as a whole, on aspects such as achieving intended educational reforms, on racial and ethnic minority learners, and on teaching and learning (see Cheng et al., 2015, for a detailed review). One of the few exceptions was a study by Abdul Kadir (2008), who examined evidence from scoring procedures, among other types of evidence, as part of her argument concerning the washback of an occupational English test on robustness of a testing program. Therefore, my study is one of the few first indicating that essay rating, as one important step of testing, can affect teaching and learning too.

My findings suggested that NMET essay rating may bring both positive and negative
washback to high school teaching. Messick (1996) included washback as one type of consequential evidence of construct validity, and defined the quality of washback (positive or negative) in relation to construct validity. Following his categorization, if the use of a test leads to interpretations of test construct underrepresentation or construct-irrelevance and subsequent teaching-to-the-test practices, it creates a negative washback effect. Otherwise, it creates a positive washback effect. In this study, teacher participants confirmed one benefit of participation was gaining access to a bigger pool of student essays. They were able to identify good and bad qualities of essays by studying a variety of essays, which deepened their understanding of the NMET writing test construct, and to determine the kind of support their students may need to improve their essay writing. In this sense, high school teachers’ involvement in the NMET ERA seemed able to bring about positive impact on classroom teaching and learning. On the other hand, NMET rating experience seemed to encourage and solidify those teachers’ teaching-to-the-test practices. They focused on improving the aspects of writing they believed to be useful in raising NMET scores and instructed their students how to please the imaginary NMET raters in order to obtain higher essay scores. Among their summarized strategies for obtaining higher scores, some led to test construct-irrelevant and construct-underrepresentative interpretations (e.g., neat handwriting, communicating to the imaginary rater). In this sense, their NMET rating experience may have led to negative washback effect on their teaching. These findings about the washback of NMET essay rating seemed partly consistent with earlier findings about washback of the NMET writing test where only negative washback was identified (e.g., L. Qi, 2005, 2007). For example, L. Qi reported that high school teachers who participated in NMET essay rating provided insider information about essay rating and were responsible for other teachers’ misinterpretations of the test construct that seemed to
contribute to the negative washback in high school English writing instruction. Such negative washback included enhanced English writing teaching-to-the-test practices, being strategic about raising test scores, emphasizing the testing situation and the assumed preferences of raters for higher NMET test scores, while ignoring the communicative purpose of NMET writing task. The examination of washback is beyond the scope of my study, however, my findings do suggest that the washback of NMET essay rating is an underresearched area.

To sum up, examining NMET essay rating through the CHAT perspective helped to gain a more comprehensive picture of the NMET ERA. This study unraveled not only how raters interacted with the NMET sociocultural context when rating essays, but also how their practices in teaching communities interacted with their activity in the rating community. My findings concerning the nature of NMET ERA highlight the richness and complexities of this human activity and deepen the understanding of raters’ decision-making in this context.

**Implications**

This section discusses implications of my study findings for research and practice on essay rating and teaching.

**Implications for Research**

This study provides a novel and important view for conceptualizing and understanding essay rating as a socially mediated activity and postulates a bi-level (i.e., surface and deep structures) conceptualization concerning ERAs. The findings highlight the value of adopting a sociocultural view of cognition, where social interaction plays a fundamental role in cognitive functioning. My study acknowledges the value of looking at essay rating from a cognitive view, as revealed by findings concerning the cognitive layer (i.e., surface structure) of NMET ERA. Meanwhile, findings concerning the social layer (i.e., deep structure) suggest that findings of the
cognitive layer alone might just be the “tip of the iceberg”. Looking into the social layer provides nuanced insights into rater decision-making embedded in the sociocultural context. Raters’ interactions with the sociocultural context—especially the interactions among raters’ goals or objectives of rating, rules, systemic tensions caused by different mediators, and raters’ corresponding solutions to those tensions—may provide a more plausible, and perhaps more meaningful, explanation to the “indeterminate process” of how raters resolved tensions to which Lumley’s (2002, 2005) study first drew our attention. By situating the cognitive layer within the social layer of ERA in this study, findings of the deep structure explained the complexity of rater decision-making, no matter how straightforward it may appear in its surface structure. Therefore, researchers need to be aware of the situated nature of ERA, and be cautious of examining essay rating in isolation and overgeneralizing their findings. From a sociocultural perspective, any research, even experimental research, is socially situated (Smagorinsky, 1998). My findings echo this need for an expanded research scope to obtain a more in-depth understanding of essay rating, which necessitates the examination of raters’ interactions with their sociocultural contexts.

This sociocultural view of essay rating requires our current understanding of rater reliability and interpretation of essay scores (i.e., based on cognitive perspectives) to be reconsidered. In performance assessment, rater reliability is usually predicted in terms of interrater agreement, rater self-consistency, severity/leniency, and bias patterns (e.g., Barkaoui, 2010a; Elder, Barkhuizen, Knoch, & von Randow, 2007; Kondo-Brown, 2002; Schaefer, 2008). Ideally, the scores assigned by any rater to one essay are interchangeable and consistent over time, regardless of rater background (e.g., gender, age, L1, education, profession, work experience), writer background (e.g., gender, L1, language proficiency, personal experience, cultural and educational background), presentation of writing tasks (e.g., task choice,
instructions, task discourse mode), characteristics of essays (e.g., writing mode, rhetoric, language errors), types of rating scale in use (e.g., holistic, analytic), and the sociocultural context (e.g., test stakes). The assumption behind this cognitive view of rater reliability is consistency and standardization to be maintained across time and location (Huot, 2002). Any rater variability associated with the aforementioned differences is considered a measurement error that lowers rater reliability (e.g., Engelhard, 1994). The contribution of different factors and their interactions to rater variability is often accounted for by using different statistic models, such as ANOVA and MANOVA (e.g., Cumming, 1990; Freedman & Calfee, 1983; Lukmani, 1996; Shi, 2001), generalizability theory (e.g., Baird, Greatorex, & Bell, 2004; Barkaoui, 2007b; Huang, 2008), multi-faceted Rasch measurement (e.g., Johnson & Lim, 2009; Schaefer, 2008; Wiseman, 2012), and multilevel modelling (e.g., Barkaoui, 2010a). The present study, however, reveals that the relationship between raters and essay scores is complex, not linear. Every rater decision-making action tends to involve interactions with different mediators. Such findings indicate that there may not exist a “true” reliability of ratings, or that a seemingly perfect reliability may be achieved for different reasons, masking a lot of rater-context interactions behind the reliability. For example, most participants’ rating performance (i.e., interrater agreement and rater self-consistency) was considered satisfactory according to the OSM system during the operational NMET rating session. This outcome was a result of constant interactions between these raters and the context, as I explained earlier. These findings, however, do not mean we need to reject reliability in writing assessment, as some researchers suggested (e.g., Huot, 2002); but rather, to adopt a sociocultural perspective so that reliability can be examined within contexts and what is behind reliability can be better understood. In the same vein, this study does not intend to question the use of those statistical models, but to suggest that
researchers give consideration to rater-context interactions when interpreting statistical results, which would otherwise provide oversimplified answers to complex questions.

This sociocultural perspective requires researchers to interpret rating results (i.e., essay scores) differently. From a cognitive view, essay scores are often considered indicators of the fixed and isolated writing ability residing in a test-taker (Huot, 2002; Swain et al., 2010). Following Swain et al., I propose that essay scores are viewed as a result of writing assessment coconstructed by writers, raters, and the sociocultural context. From a sociocultural perspective, essay writing is a socially mediated activity where writers actively interact with the context to achieve their own goals of writing (e.g., Gentil, 2006; Haneda, 2007; Lei, 2008). Essay rating—as my findings showed—is also socially situated. Therefore, writing assessment is a mediated, coconstructed activity by both parties and the sociocultural context. Hence the interpretation of derived essay scores needs to take into account the impact of writer-rater-context interactions on the scores. To this end, researchers ought to collect evidence on raters’ and writers’ activities, alongside essay scores, for a deeper understanding of what constitutes a score. When reporting scores, such information needs to be provided to essay score users and caution against overgeneralization of score use.

In addition to the essentiality of a sociocultural view (of understanding essay rating, rater reliability, and essay scores), this study demonstrates the value of employing CHAT as a guiding sociocultural theoretical framework, by outlining how to conduct essay rating research by adopting a CHAT approach: What kind of questions does CHAT lead us to ask about essay rating? How can we address those questions in terms of research design, data collection, and analysis strategies? And how do we interpret the findings? This study highlights considerations of adopting a CHAT approach versus a cognitive approach. These considerations include: How
CHAT leads us to ask research questions differently; what kind of evidence other than the cognitive evidence is needed to address those research questions; who else, besides raters, should be included as participants; how to use the research tools (e.g., TAPs, SRs, interviews) beyond the cognitive approach; what other research tools (e.g., documents) can be used and how to use them; how to analyze the collected data and integrate the results by using additional and different analysis strategies; and how to interpret the findings. The present study adopted a multiple-method multiple-perspective case study approach to researching ERA. Collectively, different sources of evidence helped to capture a comprehensive picture of essay rating, identifying the trajectory of participating raters’ ERA, from their practices in their respective teaching communities before participation, their activity at the rating community, and back to their practices at their respective teaching communities after participation. This research approach facilitated a situated understanding of NMET raters’ decision-making within the Chinese sociocultural context.

Lastly, my findings could potentially contribute to understanding influences of NMET essay rating on teaching, or washback. A relationship between essay rating and teaching was not apparent until the CHAT analysis revealed a trajectory of participating raters’ ERA. The findings suggested that NMET rating experience may not just help these teachers teach to the NMET writing test. Their NMET experience also allowed them to have a closer view of the test construct by applying the NMET rating criteria, and a broader view of student writing by examining the essays written by other students beyond their own schools and classrooms. This experience could potentially help them better understand the NMET essay writing construct, their own students’ learning needs, and their own writing instruction. Hence, NMET essay rating could potentially bring both positive and negative washback to high school teaching. These
findings bring attention to a potential research area in the field of washback studies to explore influences of NMET essay rating on classroom teaching, examining how teachers learn from this professional development experience and apply what they have learned to classroom teaching.

**Implications for Practices**

My study has several implications for essay rating practices. First and foremost, based on my findings, I have three recommendations for improving NMET essay rating practices at this provincial rating centre and across the country. The first recommendation is to provide a more detailed and regular rater training. As the six tensions in central ERA showed, raters struggled with underdefined rating criteria while maintaining high rater reliability, which they compensated for by turning to various mediating tools (e.g., prior teaching and rating experiences, collegial advice, rating quality indicators). This highlights a need for more detailed training as a supporting tool during the formal training that allots more time for raters to discuss with each other and their trainers their understanding of the rating criteria, and communicate their thoughts and difficulties with certain score band descriptors and anchor essays. As raters reported that the insufficiently defined rating criteria were a consistent challenge throughout their nine days of NMET rating, regular training sessions in this duration would be helpful to reduce the associated tensions. Detailed rater training is not only important for improving raters’ rating quality, but also important for high school teachers’ professional development which could potentially bring positive influences to classroom teaching. The high school teachers in this study lacked teaching resources in their schools to support their English writing instruction; therefore, they participated in NMET ERA as professional development. My findings revealed that other than the possibly negative washback of enhanced teaching-to-the-test practices, their NMET rating experience could potentially bring about positive washback, as I discussed earlier. In light
of this potential positive washback, NMET rating centres can take advantage of essay rating as a professional development session for high school teachers, providing them with more detailed training and helping them to better understand the intended NMET essay writing construct. This may provide a solution to addressing the negative washback of NMET writing that L. Qi (2005) identified, where she found the intended communicative writing construct of the NMET was unsuccessfully translated in high school teaching. Based on my findings, it was possible that the unsuccessful translation was also caused by, among other reasons, insufficient teaching resources available to teachers to support their understanding of communicative English writing and subsequent translation into their classroom teaching.

Given this potential positive washback, my second recommendation is that NMET rating centres should consider providing all high school teachers with periodic opportunities to participate in NMET essay rating as professional development. Although the challenge remains as to how to curb the negative washback and help teachers use NMET rating experience to affect their teaching practices positively, this may not be a unique challenge to this context. These first two recommendations may also have important implications for rating other Gaokao subjects.

My third recommendation for NMET essay rating practices is trying not to overemphasize interrater agreement to raters. As my findings indicated, when raters saw their rating quality indicators were not good enough, or not as good as other raters, three types of concerns arose (i.e., failing to meet institutional requirements, failing to be accountable for the students, losing “face”). Due to these concerns associated with interrater agreement, raters turned to different tools and strived to keep their ratings in line with those of others, which led to both positive (e.g., trying to improve their understanding of the rating criteria) and negative consequences (e.g., assigning middle ground scores). Therefore, the NMET rating centre should
be aware of such consequences, avoiding overemphasizing interrater agreement to reduce the potentially negative impact as much as possible.

The CHAT approach this study adopted also opens up avenues for efforts to improve essay rating practices in other contexts. The use of CHAT is not limited to describing and explaining complex human activities such as professional development (Karasavvidis, 2009; Yamagata-Lynch, 2003), writing (Cumming, 2006; Gentil, 2006; Haneda, 2007; Lei, 2008), grading (J. Li, 2016), and essay rating, the latter being evidenced in this study. CHAT analyses can also be used to design and implement interventions based on identified tensions, for the sake of betterment and positive change (Engeström, 1993, 2000, 2015). Any testing institution can make use of CHAT analyses to generate information to guide essay rating practices. For example, when recruiting raters, understanding the context where raters work or come from would be helpful in identifying possible interactions that may influence their performance. When raters participate in essay rating, a CHAT analysis can help to check the alignment between raters’ goals and the intended institutional ones, and identify systemic tensions that detract raters from the intended institutional goals. More importantly, CHAT analyses can help identify what kind of support for which mediator is needed so as to reduce unnecessary tensions and subsequently redirect raters’ personal goals to the intended ones. This entails the need for a more customized approach to rater training, instead of conventional practices of standardized training.

Testing institutions should also be cautious when making any change in existing activity systems. When new elements or materials are introduced, or old ones are removed from an activity system, new tensions may arise. The new tensions can either alleviate current contradictions in existing systems or produce new ones, and both can influence raters’ ERAs or even change the nature of their activities. Testing institutions should thus be aware of this
challenge, and take corresponding measures to promote positive changes while mitigating negative ones. For example, when there is a change in the writing task or rater training mode, raters’ attitudes toward the change should not simply be understood as raters’ personal preferences. An examination of if and how such change leads to new tensions may provide insight into the reasons behind raters’ attitudes. Once emerging tensions are identified, supporting or introducing corresponding elements in related mediators may help to reduce tensions.

Lastly, the renewed sociocultural view of essay scores I proposed earlier has important implications for the use of automated writing evaluation (AWE) systems. In recent years, AWE systems have gained popularity in large-scale testing, where AWE software is often used together with human raters to score essays (Deane, 2013; Elliot & Williamson, 2013). Although AWE software was found to be able to achieve high agreement with human raters (e.g., Warschauer & Ware, 2006), a machine can never function as does a human rater (Attali, Lewis, & Steier, 2012). Testing institutions that adopt such a combined approach need to be aware of the different meanings behind scores from the two sources, rather than assuming they are interchangeable.

**Limitations**

There were two major limitations to this study. The first limitation was that due to practical constraints and feasibility issues in this high-stakes testing context, data could not be collected during the operational NMET essay rating session. All the influences claimed by participating raters were thus introspection- and perception-based. No quantitative evidence of authentic NMET essay scores was available to support their claims. NMET sample essays used in this study were collected by the researcher, instead of authentic NMET essays raters encountered in operational NMET rating sessions. TAPs and SRs were conducted under a simulated research...
setting and not under an authentic NMET rating setting. Effects of NMET rating experience and those of TAP method on raters’ essay rating were examined based on perception, instead of more rigorous experimental designs. Although most participants claimed their essay rating actions in this study were similar or close to how they performed in the operational sessions, from a CHAT perspective, those differences between real-life NMET essay rating and the research condition can affect the research findings, as research methods themselves are socially situated practices (Barkaoui, 2011b; Smagorinsky, 2001). Practical constraints also made onsite observation at the rating centre unattainable and unavailable to this study, which could have provided first-hand, direct experience of participants’ activities (Merriam, 2009). In spite of these practical constraints, my efforts to collect data as close to operational NMET rating sessions as possible and using a multiple-method multiple-perspective approach, make me confident that the findings from this study were plausible inferences of what happened in the NMET ERA based on the multiple perspectives and sources of data.

Another limitation was that findings from this study emerged from a qualitative case study alone involving a small sample of raters’ NMET ERA at one provincial rating centre out of 31 provincial centres in China. Due to the focus I placed on this rating centre and the sensitivity of CHAT to the sociocultural context, my findings may not be generalizable to other NMET rating centres. My research goal was instead to showcase the value of adopting a CHAT perspective in identifying and unpacking the complexity of rater decision-making, in an attempt to promote further exploration in essay rating research. To capture a bigger picture of NMET ERAs in China, future CHAT-based studies should be conducted in multiple provincial centres across the country to survey a larger sample of NMET raters.
Recommendations for Future Research

Use of the CHAT framework as a sociocultural approach opens up a range of opportunities for essay rating research. First of all, future studies can replicate the current study in different contexts or conditions. These different contexts or conditions may include: essay rating with different levels of consequences (e.g., high- or low-stakes essay tests), or different purposes (e.g., selection, certification, immigration, placement, diagnosis); when raters come from different cultural contexts (e.g., Australia, Brazil, China, Canada), or work under different types of dynamics (e.g., solo, collaboratively, at home, in group setting). Comparison studies may be needed to see the above differential influences on ERAs. Conducting CHAT analysis may help better understand differences, if any, in raters’ rating processes and results that are related to the different contexts or conditions.

Future essay rating research can apply CHAT analysis to help understand inconsistent findings from cognitive evidence. For example, previous studies found that raters may employ different strategies for reading essays, interpreting rating scales, and assigning scores (e.g., Milanovic et al., 1996; Smith, 2000; Vaughan, 1991). The cognitive perspective those studies adopted was not able to satisfactorily explain such differences among raters. A CHAT perspective may bring insights into how and why raters adopt different styles to approach the same rating task. Another example is inconsistent findings concerning influences on rating processes and results associated with differences in the four key elements (i.e., rater, writing task, rating scale, essay) in essay rating as I discussed in Chapter 2. From a CHAT perspective, any difference in these key elements may lead to changed nature of ERAs, turning the seemingly same rating task into different activities. Incorporating a CHAT analysis can generate a situated
view and include considerations from the social layer to explain differences in rating processes and results.

Researchers can also reanalyze and interpret from a CHAT perspective the previous essay rating studies that followed the cognitive line of thinking to yield new insights, as I exemplified in Chapter 2. Some second language education researchers reexamined their previous studies through an activity theoretical lens, and obtained more in-depth understanding of the research phenomenon they had examined (e.g., Coughlan & Duff, 1994; Haneda, 2007). Likewise, reinspecting previous essay rating studies from a new angle, with additional information about the social layer of ERAs, may bring researchers a better, if not different, understanding of rater decision-making in the set contexts.

Future essay rating research can apply CHAT analysis to examine the effectiveness of new rating procedures or materials as they are developing and/or implemented, that is, to conduct developmental research. Given that CHAT has the potential of identifying tensions and supporting transformation, CHAT analyses can be used when new essay rating procedures or materials are introduced in ERAs, to provide timely, continuous feedback on their effectiveness in supporting intended purposes. Such feedback can include information about how well raters are interacting with the new procedure or material, what works and what does not, what kinds of tensions the new procedure or material has created or alleviated, what kinds of support may be needed. The research findings can offer solutions and alternatives to subsequent development. This kind of CHAT analysis can be reiterated until the new procedure or material functions satisfactorily.

In addition to the above recommendations for future CHAT-based essay rating research, this study also yields recommendations for future washback studies to explore the impact of
teachers’ involvement in high-stakes rating (e.g., their NMET rating experience in this study) on their teaching practices. In this study, participants cited that their teaching practices were influenced by their participation in NMET rating. It is beyond the scope of the current study but a worthwhile research area to explore with more empirical evidence concerning the influence of teacher raters’ essay rating experiences, especially high-stakes ones, on their teaching practices. In many contexts, teachers are hired as raters for various high-stakes tests. From a professional development perspective, their rating experiences may have an impact on their teaching practices afterward. Future washback studies can examine such influences by comparing teacher raters’ teaching practices before and after their participation in high-stakes rating activities and addressing questions such as: (a) whether there is truly a change in their practices; (b) if so, how their practices have changed; (c) what kinds of change have taken place; and (d) what caused such change. This line of research will help enrich our understanding of washback and essay rating from another angle.

Finally, other sources of data (e.g., essay scores, rater questionnaires, direct observation) collected under real-life conditions can enrich the information about the activity under study for more comprehensive findings. For example, quantitative data (e.g., essay scores, rater questionnaires) can complement qualitative evidence when exploring ERAs, which may lead to a better understanding of the complexities in ERAs. Future research can also include direct observation of ERAs while they are happening to collect first-hand information about raters’ activities, in conjunction with data collection methods used in this study (i.e., TAPs, SRs, interviews, and document analysis), to enrich the researcher’s understanding of the phenomenon and thus enhance research trustworthiness.
Conclusion

Essay raters are social beings who rate written work composed by other social beings. To better understand how they make scoring decisions that have an impact on essay writers and other stake-holders, it is important to study how raters’ decision-making interacts with the sociocultural context in which their essay rating practices are situated. My research explored how raters scored essays for a high-stakes English test used for university entrance decisions in China—a country with a test-driven culture—during which raters interacted with their rating community at the provincial rating centre, their teaching communities at the high schools where they taught, and the education system. In particular, this research applied the CHAT framework to examine NMET essay rating from a sociocultural perspective that incorporated the examination of both cognitive and social processes of rater decision-making, revealing the socially mediated nature of NMET essay rating.

My study has three major research contributions. First, it highlights the value of a sociocultural view to essay rating research and postulates a bi-level (surface and deep structures) conceptualization concerning ERAs. A sociocultural view understands raters’ cognitive functioning by situating it into its sociocultural context, rather than as an isolated event. This view could help to understand not only the “what” (surface structure) but also the “how” and “why” (deep structure) in raters’ decision-making, thus making the findings more meaningful. Second, this study demonstrates how to use the CHAT framework as a sociocultural approach to conduct essay rating research and the value of doing so. Last, based on findings from the CHAT analysis, this study provides a direction for future washback studies, implying that teachers’ involvement in high-stakes rating, such as the NMET in this study, may lead to potential washback effect on their teaching practices.
My study also has two major practical implications. First, this study provides support for improving NMET rating practices and potentially positive washback in Chinese high school English teaching. The findings stress the need for a detailed and regular rater training during NMET rating sessions, for the purposes of improving rating quality and supporting high school teachers’ professional development that could potentially bring positive washback to high school teaching. NMET rating centres should consider providing every high school teacher with periodic opportunities to participate in NMET essay rating for professional development purposes. My findings also highlight that raters’ concerns about interrater agreement could bring both positive and negative impact on raters’ essay rating, and hence NMET rating centres should be careful and avoid overemphasizing interrater agreement. Second, this study suggests the practical value of applying CHAT to improving essay rating practices. The study demonstrates that a CHAT analysis can help to better understand “what works and what does not” in essay rating practices and provide corresponding support, with implications for practices in other contexts.
References


http://doi.org/10.1111/j.1467-8535.2008.00838.x


http://doi.org/10.1016/j.intcom.2007.07.003


http://doi.org/10.1016/j.asw.2014.03.002


http://doi.org/10.1016/j.asw.2004.01.001


http://doi.org/10.1016/j.jslw.2008.04.001


http://doi.org/10.1080/15434303.2015.1011738


http://doi.org/10.1017/CBO9781107415324.004


Qi, L. (2005). Stakeholders’ conflicting aims undermine the washback function of a high-stakes


Warschauer, M., & Ware, P. (2006). Automated writing evaluation: Defining the classroom


Appendix A Letter of Approval from the General Research Ethics Board

May 28, 2014

Ms. Yi Mei
Ph.D. Candidate
Faculty of Education
Queen’s University
Duncan McArthur Hall
511 Union Street West
Kingston, ON, K7M 5R7

GREB Ref #: GEDUC-730-14; Research # 6012788
Title: “GEDUC-730-14 Understanding Essay Rating Activity: A Sociocultural Perspective”

Dear Ms. Mei,

The General Research Ethics Board (GREB), by means of a delegated board review, has cleared your proposal entitled “GEDUC-730-14 Understanding Essay Rating Activity: A Sociocultural Perspective” for ethical compliance with the Tri-Council Guidelines (TCGS) and Queen’s ethical policies. In accordance with the Tri-Council Guidelines (article D.1.6) and Senate Terms of Reference (article 3), your project has been cleared for one year. At the end of each year, the GREB will ask if your project has been completed and, if not, what changes have occurred or will occur in the next year.

You are reminded of your obligation to advise the GREB, with a copy to your unit REB, of any adverse event(s) that occur during this one year period (access this form at https://services.queens.ca/emu/researcher/ and click Events - GREB Adverse Event Report). An adverse event includes, but is not limited to, a complaint, a change or unexpected event that alters the level of risk for the researcher or participant(s) or situation that requires a substantial change in approach to a participant(s). You are also advised that all adverse events must be reported to the GREB within 48 hours.

You are also reminded that all changes that might affect human participants must be cleared by the GREB. For example, you must report changes to the level of risk, participant characteristics, and implementation of new procedures. To make an amendment, access the application at https://services.queens.ca/emu/researcher/ and click Events - GREB. You must submit an Approved Study Form. These changes will automatically be sent to the Ethics Coordinator, Gail Irving, at the Office of Research Services or研科研@saces.ca for further review and clearance by the GREB or GREB Chair.

On behalf of the General Research Ethics Board, I wish you continued success in your research.

Yours sincerely,

Joan Stevenson, Ph.D.
Chair, General Research Ethics Board

c Dr. Ling Cheng, Faculty Supervisor
Dr. Benjamin Bolden, Chair, Unit REB
Ms. Stacey Boulton, Office of Graduate Studies and Bureau of Research
Appendix B.1 Letter of Information/Consent Form – Team Leaders and Directors

Understanding Essay Rating Activity: A Sociocultural Perspective
从社会文化视角理解写作评分活动

This research is being conducted by Yi Mei (Doctor of Philosophy, Candidate) under the supervision of Liying Cheng in the Faculty of Education at Queen’s University in Kingston, Ontario. This study has been granted clearance according to the recommended principles of Canadian ethics guidelines and Queen’s policies.

What is this study about? Essay rating activity is a socially-situated practice but has been researched from a decontextualized view point (Barkaoui, 2008). The purpose of this study is to understand the rating practices of the National Matriculation English Test (NMET) essay raters while they rate NMET essays in a provincial NMET rating centre. To be specific, I am interested in raters’ NMET rating experiences, and what individual NMET raters’ decision-making behaviours are when they rate NMET sample essays. Your experience with NMET rating operation procedures will be used to facilitate our understanding of the NMET raters’ activities.

What will this study require? If you agree to participate in this research you will be interviewed about NMET rating operation procedures such as contents of rater training and your interactions with raters and/or (other) team leaders for quality control reasons. The entire process will last 1 hour and be scheduled at the end of the NMET rating session. You have the full choice of the time and location for the interview. And the interview will be recorded in digital audio files.

Is participation voluntary? Your participation is completely voluntary and choosing not to participate will not result in any adverse consequences. There are no known physical, psychological, economic, or social risks associated with this study. Further, you are free to choose, without reason or consequence, to refuse to answer any questions. You may withdraw from the study within 30 days after your participation with no negative consequences. If you withdraw from the study, you may choose to have your data removed.

What will happen to my responses? The interview recording will be transcribed and then the recording will be destroyed. All electronic files will be password protected. Only the researcher has the access to the files. Paper and audio data will be secured in a locked cabinet. I will maintain copies of the transcripts for a minimum of 5 years and may use the data (with names removed) in subsequent research. Confidentiality will be protected to the extent possible. None of the data will contain your name or the identity of your place of work. To protect your identity a pseudonym will replace your name on all data files and in any dissemination of findings. This research may result in publications of various types, including journal articles or other professional publications.

What if I have concerns? Any questions about study participation or a request to withdraw from the study may be directed to Yi Mei at +86 139-9423-1170; yi.mei@queensu.ca or my supervisor Liying Cheng at +1 613-533-6000 ext. 77431; liying.cheng@queensu.ca. Any ethical concerns about the study may be directed to the Chair of the General Research Ethics Board at (613) 533-6081 or chair.GREB@queensu.ca.

215
CONSENT FORM

Understanding Essay Rating Activity: A Sociocultural Perspective
从社会文化视角理解写作评分活动

Name (please print clearly): ____________________________________________________

1. I have read and retained the Letter of Information and have had any questions answered to my satisfaction.

2. I understand that I will be participating in the study called Understanding Essay Rating Activity: A Sociocultural Perspective (从社会文化视角理解写作评分活动). I understand the purpose of this research is to understand the NMET essay raters’ practices while rating NMET essays in a provincial NMET rating centre. I understand that participation in this study will entail a maximum of 1 hour of my time involving an interview.

3. I agree that the individual interview can be audio recorded.

4. I understand that my participation in this study is voluntary and I may withdraw within 30 days after my participation without adverse consequences. I understand that if I withdraw from the study, I may choose to have my data removed. I understand that the data may also be published in professional journals or presented at academic conferences. I understand that every effort will be made to maintain confidentiality to the extent possible now and in the future.

5. I am aware that any questions about study participation or a request to withdraw from the study may be directed to Yi Mei at +86 139-9423-1170; yi.mei@queensu.ca or her supervisor Liying Cheng at +1 613-533-6000 ext. 77431; liying.cheng@queensu.ca. Any ethical concerns about the study may be directed to the Chair of the General Research Ethics Board at (613) 533-6081 or chair.GREB@queensu.ca.

Please sign one copy of this Consent Form and return to the researcher.
Retain the second copy for your records.

I have read the above statements and had any questions answered. I freely consent to participate in this study.
Participant’s Signature: ______________________________________________________

Date: __________________________ E-mail address: _______________________________
Appendix B.2 Letter of Information/Consent Form – NMET Raters

Understanding Essay Rating Activity: A Sociocultural Perspective
从社会文化视角理解写作评分活动

This research is being conducted by Yi Mei (Doctor of Philosophy, Candidate) under the supervision of Liying Cheng in the Faculty of Education at Queen’s University in Kingston, Ontario. This study has been granted clearance according to the recommended principles of Canadian ethics guidelines and Queen’s policies.

**What is this study about?** Essay rating activity is a socially-situated practice but has been researched from a decontextualized viewpoint (Barkaoui, 2008). The purpose of this study is to understand the rating practices of the National Matriculation English Test (NMET) essay raters while rating NMET essays in a provincial NMET rating centre. To be specific, I am interested in raters’ NMET rating experiences, and what individual NMET raters’ decision-making behaviours are when they rate NMET sample essays.

**What will this study require?** You will also engage in a series of tasks: (a) a training session on think-aloud protocols (TAPs); (b) rating the eight sample essays as you would in the operational NMET rating session but while thinking aloud; (c) a stimulated recall (SR) interview to talk retrospectively about your thought processes while rating the eight sample essays; and (d) an interview about your goal-directed actions involved in rating the sample essays. The entire process will last 4 hours and be roughly scheduled on three separate days. You have the full choice of the time and location.

Within four weeks after the NMET rating session, I will contact you for another individual interview, and it will last 1 hour. You will be asked to elaborate/comment on how your NMET essay decision-making actions influenced/were influenced by other interrelated activities. You have the full choice of the time and location for the second individual interview. All the interviews and TAPs will be recorded in digital audio files. You will also help the researcher to recruit your colleagues at school as participants.

**Is participation voluntary?** Your participation is completely voluntary and choosing not to participate will not result in any adverse consequences. There are no known physical, psychological, economic, or social risks associated with this study. Further, you are free to choose, without reason or consequence, to refuse to answer any questions. You may withdraw from the study within 30 days after your participation with no negative consequences. If you withdraw from the study, you may choose to have your data removed.

**What will happen to my responses?** The TAP and interview recordings will be transcribed and then the recording will be destroyed. All electronic files will be password protected. Only the researcher has the access to the files. Paper and audio data will be secured in a locked cabinet. I will maintain copies of the transcripts for a minimum of 5 years and may use the data (with names removed) in subsequent research. Confidentiality will be protected to the extent possible. None of the data will contain your name or the identity of your place of work. To protect your identity a pseudonym will replace your name on all data files and in any dissemination of
findings. This research may result in publications of various types, including journal articles or other professional publications.

**Will I be compensated for my participation?** Yes, after full participation in the above five tasks, you will receive a token of appreciation.

**What if I have concerns?** Any questions about study participation or a request to withdraw from the study may be directed to Yi Mei at +86 139-9423-1170; yi.mei@queensu.ca or my supervisor Liying Cheng at +1 613-533-6000 ext. 77431; liying.cheng@queensu.ca. Any ethical concerns about the study may be directed to the Chair of the General Research Ethics Board at (613) 533-6081 or chair.GREB@queensu.ca.
CONSENT FORM

Understanding Essay Rating Activity: A Sociocultural Perspective
从社会文化视角理解写作评分活动

Name (please print clearly): ____________________________________________________

1. I have read and retained the Letter of Information and have had any questions answered to my satisfaction.

2. I understand that I will be participating in the study called Understanding Essay Rating Activity: A Sociocultural Perspective (从社会文化视角理解写作评分活动). I understand the purpose of this research is to understand the NMET essay raters’ practices while rating NMET essays in a provincial NMET rating centre. I understand that participation in this study will entail a maximum of 4 hours of my time involving: a TAP training (maximum 30 minutes), TAPs while scoring eight essays (maximum 30 minutes), a SR interview (maximum 60 minutes), two individual interviews (maximum 60 minutes each). I will also help the researcher to recruit my colleagues at school as participants.

3. I agree that the TAPs, the SR interview, and the two individual interviews can be audio recorded.

4. I understand that my participation in this study is voluntary and I may withdraw within 30 days after my participation without adverse consequences. I understand that if I withdraw from the study, I may choose to have my data removed. I understand that the data may also be published in professional journals or presented at academic conferences. I understand that every effort will be made to maintain confidentiality to the extent possible now and in the future.

5. I am aware that any questions about study participation or a request to withdraw from the study may be directed to Yi Mei at +86 139-9423-1170; yi.mei@queensu.ca or her supervisor Liying Cheng at +1 613-533-6000 ext. 77431; liying.cheng@queensu.ca. Any ethical concerns about the study may be directed to the Chair of the General Research Ethics Board at (613) 533-6081 or chair.GREB@queensu.ca.

Please sign one copy of this Consent Form and return to the researcher.
Retain the second copy for your records.

I have read the above statements and had any questions answered. I freely consent to participate in this study.
Participant’s Signature: ______________________________________________________

Date: __________________________ E-mail address: _______________________________
Appendix B.3 Letter of Information/Consent Form – Participants at School Context

Understanding Essay Rating Activity: A Sociocultural Perspective
从社会文化视角理解写作评分活动

This research is being conducted by Yi Mei (Doctor of Philosophy, Candidate) under the supervision of Liying Cheng in the Faculty of Education at Queen’s University in Kingston, Ontario. This study has been granted clearance according to the recommended principles of Canadian ethics guidelines and Queen’s policies.

What is this study about? Essay rating activity is a socially-situated practice but has been researched from a decontextualized view point (Barkaoui, 2008). The purpose of this study is to understand the rating practices of the National Matriculation English Test (NMET) essay raters while rating NMET essays in a provincial NMET rating centre. To be specific, I am interested in raters’ NMET rating experiences, and what individual NMET raters’ decision-making behaviours are when they rate NMET sample essays. Your experience with how the activities in your school have been related with NMET essay rating activity will be used to facilitate our understanding of the NMET raters’ activities.

What will this study require? If you agree to participate you will be interviewed about activities in your school that are related with NMET essay rating activity, such as NMET writing test preparation, recommending/supporting teachers to be NMET raters, and sharing NMET essay rating tips and strategies. The entire process will last 1 hour. You have the full choice of the time and location for the interview. The interview will be recorded in digital audio files.

Is participation voluntary? Your participation is completely voluntary and choosing not to participate will not result in any adverse consequences. There are no known physical, psychological, economic, or social risks associated with this study. Further, you are free to choose, without reason or consequence, to refuse to answer any questions. You may withdraw from the study within 30 days after your participation with no negative consequences. If you withdraw from the study, you may choose to have your data removed.

What will happen to my responses? The interview recording will be transcribed and then the recording will be destroyed. All electronic files will be password protected. Only the researcher has the access to the files. Paper and audio data will be secured in a locked cabinet. I will maintain copies of the transcripts for a minimum of 5 years and may use the data (with names removed) in subsequent research. Confidentiality will be protected to the extent possible. None of the data will contain your name or the identity of your place of work. To protect your identity a pseudonym will replace your name on all data files and in any dissemination of findings. This research may result in publications of various types, including journal articles or other professional publications.

What if I have concerns? Any questions about study participation or a request to withdraw from the study may be directed to Yi Mei at +86 139-9423-1170; yi.mei@queensu.ca or my supervisor Liying Cheng at +1 613-533-6000 ext. 77431; liying.cheng@queensu.ca. Any ethical concerns about the study may be directed to the Chair of the General Research Ethics Board at (613) 533-6081 or chair.GREB@queensu.ca.
CONSENT FORM

Understanding Essay Rating Activity: A Sociocultural Perspective
从社会文化视角理解写作评分活动

Name (please print clearly): ____________________________________________________

1. I have read and retained the Letter of Information and have had any questions answered to my satisfaction.

2. I understand that I will be participating in the study called Understanding Essay Rating Activity: A Sociocultural Perspective (从社会文化视角理解写作评分活动). I understand the purpose of this research is to understand the NMET essay raters’ practices while rating NMET essays in a provincial NMET rating centre. I understand that participation in this study will entail a maximum of 1 hour of my time involving an interview.

3. I agree that the individual interview can be audio recorded.

4. I understand that my participation in this study is voluntary and I may withdraw within 30 days after my participation without adverse consequences. I understand that if I withdraw from the study, I may choose to have my data removed. I understand that the data may also be published in professional journals or presented at academic conferences. I understand that every effort will be made to maintain confidentiality to the extent possible now and in the future.

5. I am aware that any questions about study participation or a request to withdraw from the study may be directed to Yi Mei at +86 139-9423-1170; yi.mei@queensu.ca or her supervisor Liying Cheng at +1 613-533-6000 ext. 77431; liying.cheng@queensu.ca. Any ethical concerns about the study may be directed to the Chair of the General Research Ethics Board at (613) 533-6081 or chair.GREB@queensu.ca.

Please sign one copy of this Consent Form and return to the researcher.
Retain the second copy for your records.

I have read the above statements and had any questions answered. I freely consent to participate in this study.

Participant’s Signature: ____________________________________________________

Date: __________________________ E-mail address: _______________________________
Appendix B.4 Letter of Information/Consent Form – NMET Sample Essay Writers

Understanding Essay Rating Activity: A Sociocultural Perspective
从社会文化视角理解写作评分活动

This research is being conducted by Yi Mei (Doctor of Philosophy, Candidate) under the supervision of Liying Cheng in the Faculty of Education at Queen’s University in Kingston, Ontario. This study has been granted clearance according to the recommended principles of Canadian ethics guidelines and Queen’s policies.

What is this study about? Essay rating activity is a socially-situated practice but has been researched from a decontextualized viewpoint (Barkaoui, 2008). The purpose of this study is to understand the rating practices of the National Matriculation English Test (NMET) essay raters while they rate NMET essays in a provincial NMET rating centre. To be specific, I am interested in raters’ NMET rating experiences, and what individual NMET raters’ decision-making behaviours are when they rate NMET sample essays. Your reproduced NMET essays will be used to facilitate our understanding of the NMET raters’ activities.

What will this study require? If you agree to participate in this research you will write an essay based on an NMET writing task. You will receive a rough estimate of your essay scores based on the official NMET scoring rubrics. Your essay will be rated by NMET raters, and may be selected for further study as part of this research project. You have the full choice of the time and location to reproduce your NMET essays. The total estimated time required for participation is 30 minutes at maximum.

Is participation voluntary? Your participation is completely voluntary and choosing not to participate will not result in any adverse consequences. There are no known physical, psychological, economic, or social risks associated with this study. Further, you are free to choose, without reason or consequence, to refuse to answer any questions. You may withdraw from the study within 30 days after your participation with no negative consequences. If you withdraw from the study, you may choose to have your data removed.

What will happen to my responses? Your reproduced essays will be photocopied, and all paper data will be secured in a locked cabinet. Only the researcher has the access to the files. Confidentiality will be protected to the extent possible. None of the data will contain your name or the identity of your school. To protect your identity a pseudonym will replace your name on all data files and in any dissemination of findings. Once the study is complete, the files, without identifying information, will be destroyed after five years upon the completion of the study. This research may result in publications of various types, including journal articles or other professional publications.

What if I have concerns? Any questions about study participation or a request to withdraw from the study may be directed to Yi Mei at +86 139-9423-1170; yi.mei@queensu.ca or my supervisor Liying Cheng at +1 613-533-6000 ext. 77431; liying.cheng@queensu.ca. Any ethical concerns about the study may be directed to the Chair of the General Research Ethics Board at (613) 533-6081 or chair.GREB@queensu.ca.
CONSENT FORM

Understanding Essay Rating Activity: A Sociocultural Perspective
从社会文化视角理解写作评分活动

Name (please print clearly): ____________________________________________________

1. I have read and retained the Letter of Information and have had any questions answered to my satisfaction.

2. I understand that I will be participating in the study called Understanding Essay Rating Activity: A Sociocultural Perspective (从社会文化视角理解写作评分活动). I understand the purpose of this research is to understand the NMET essay raters’ practices while rating NMET essays in a provincial NMET rating centre. I understand that participation in this study will entail a maximum of 30 minutes of my time involving reproducing my NMET essay.

3. I understand that my participation in this study is voluntary and I may withdraw within 30 days after my participation without adverse consequences. I understand that if I withdraw from the study, I may choose to have my data removed. I understand that the data may also be published in professional journals or presented at academic conferences. I understand that every effort will be made to maintain confidentiality to the extent possible now and in the future.

4. I am aware that any questions about study participation or a request to withdraw from the study may be directed to Yi Mei at +86 139-9423-1170; yi.mei@queensu.ca or her supervisor Liying Cheng at +1 613-533-6000 ext. 77431; liying.cheng@queensu.ca. Any ethical concerns about the study may be directed to the Chair of the General Research Ethics Board at (613) 533-6081 or chair.GREB@queensu.ca.

Please sign one copy of this Consent Form and return to the researcher.
Retain the second copy for your records.

I have read the above statements and had any questions answered. I freely consent to participate in this study.

Participant’s Signature: _________________________________________________________

Date: __________________________ E-mail address: _________________________________
Appendix C Think-Aloud Training Guidelines and Instructions
(Adapted from Barkaoui, 2008, pp.263-264)

Guidelines for think-aloud training:

I would like you to rate the eight NMET sample essays just as you rate them as in the operational NMET rating sessions as far as possible. However, there will be one important difference in rating this batch of essays. As I have previously explained, I am conducting a study of the processes used by raters when they rate NMET essays; and I would now like you to talk and think aloud as you read and decide what score to give each essay, while the tape recorder records what you say. Think-aloud protocols ask people to say everything they think about while they perform a task, with the aim of documenting and better understanding what you pay attention to and consider important when you do a task. I want you to say out loud everything that you say to yourself silently. Just act as if you are talking to yourself.

To help you do this, I will provide you with training on thinking aloud. I will give you some tasks and you have to perform them while thinking aloud. You have to tell me everything that you are thinking from the time you first see the task until you give an answer. I would like you to talk aloud constantly from the time I present each task until you have given your final answer to the question. I don't want you to try to plan out what you say or try to explain to me what you are saying. Just act as if you are alone in the room speaking to yourself. It is most important that you keep talking. If you are silent for any period of time I will ask you to talk. Do you understand what I want you to do?

Now let's begin with some practices. I want you to solve some math problems and tell me what you are thinking as you get the answer.

- What is the result of multiplying 23 by 72?
- A father, a mother, and their son are 80 years old together. The father is twice as old as the son. The mother has the same age as the father. How old is the son?

Instructions for rating while thinking aloud (for participants):

Please read these instructions carefully before you begin rating the essays.

You will rate a set of eight NMET sample essays while you think aloud. I would like you to rate these essays just as you rated the authentic NMET essays as far as possible. The purpose of the think-aloud protocols for this study is to find out in as much detail as possible what you are thinking about, deciding, and doing while you read and decide what score to give each essay. The most important thing to emphasize is, say everything you are thinking about, and make certain this is recorded clearly onto the tape recorder. What you say will become important data for my research. Score the essays as quickly as possible based on the NMET scale. I want you to think aloud through the entire scoring process. Tell me everything you are thinking from the time you first see each essay until you give a score. Please read the following guidelines carefully.
1. Take a few minutes to study the topics and scale level descriptors. Please record this as well.

2. As you assess each essay, indicate clearly the code number of the essay that you are rating when you start to read it. Then when you have made a rating decision, indicate the score you assign to it. Then indicate the code number of the next one as you start to assess it, and soon.

3. Say as much as you can while you are reading the composition and deciding on how to rate it, and make sure the number you assign to each composition is recorded along with your ongoing impressions of it. Please write the scores on the top of each essay.

4. Report your first impression of each essay and if it influences your rating. Then continue talking, saying what you are thinking about, as you are making your assessment decisions.

5. It is not necessary for you to write anything for this task. However, if you write something down (i.e., marks, comments, corrections), say that you are writing, and report what you are writing. Feel free to write on the essay, if you like, but we want to know on the tape recording what you might write.

6. It is important that you keep talking all the time, registering your thoughts all the time. If you spend time reading the essay, the topic, or the rating scale, then you should do that aloud also, so that I can understand what you are doing at that time. Make sure you report exactly what you are doing all the time.

7. If you happen to reconsider any of your ratings (e.g., for a second or third time), verbalize your reason(s) for doing so, and indicate on the tape that this is what you are doing and the code of the essay. Please revise the final score on the page of the essay.

8. Keep talking, conveying your thoughts continuously, while you assess the essays, from the initial point when you first see each essay until you have completed rating it, and indeed until you rate the whole set of essays.

9. Speak continuously. Report fully, even what might seem trivial. Do not assume that others know what you are doing or thinking.

10. Try to avoid speech fillers (i.e., uh, urn) as much as possible. Try to use words instead, so that I can understand what your thoughts have been.

11. Talk and make your assessment as naturally and as honestly as you can, according to what you usually do when you assess students’ essays. Don't start rationalizing your ideas at length; I am just interested in your decisions as you make them, and as you would do if you were doing this task on your own.

12. If you have to take a break while you are assessing the compositions, indicate on the tape that
you are doing this, turn the tape recorder off, and then, when you start again, indicate this clearly onto the tape.

13. When you have completed assessing all of the compositions in the package, then indicate this clearly on the tape, so we know when your think-aloud protocol ends.

Thank you for your cooperation!
Appendix D Stimulated Recall and Follow-Up Interview Guide

Guidelines and instructions for raters conducting SRs:

You just completed marking the eight sample essays while thinking aloud, i.e., talk aloud to yourself, but I do not know how you came to the scores you assigned to them. I am interested in knowing what you were thinking at the time you were marking each essay and deciding on the score you gave. Now I would like you to listen to your TAP recording while rereading the eight NMET sample essays one by one, and then tell me what was in your mind at that time while you were assessing each essay and assigning each score. After that, I will ask questions about your thought process.

Questions:

1. Follow-up questions about perceptions, reasons, and underlying purposes:
   a) What do you think of this essay? Why did you mark this essay in this particular way?
   b) What were your rating priorities? Why?

   | Assess task completion (what do you mean by this?) | Assess fulfillment of communicative writing purposes |
   | Assess coverage of major details | Assess effectiveness of cohesive devices |
   | Consider lexis, syntax/morphology, gravity and frequency of errors | Consider handwriting/eligibility, text length, punctuation, spelling |

c) What do you think has influenced the way you marked the essays (e.g., consideration about consequences of your rating/any political issues or considerations/any requirements by the rating centre/your NMET test-prep training experience)? How? Why?

2. This rating task vs. authentic NMET rating task: If you were rating these essays in the operational NMET rating session, will you assess them in a different way? If yes, how and why?

3. Perceived effects of TAPs:
   a) What are your thoughts about thinking aloud?
   b) Do you think thinking aloud affected the way you rated the essays? Your views of essay rating? Your rating criteria? The scores you assigned? If yes, why/how?
   c) Did thinking aloud make you aware of things you had not thought about before regarding essay rating?
d) Did you discover anything new or learn anything about your rating from thinking aloud? Did thinking aloud make you think of new things to try when rating essays in the future?

e) Do you have any thoughts or comments that you would like to add about your experience thinking aloud while rating the essays?

4. **Perceived effects of SRs:**
   a) What are your thoughts about stimulated recalls?
   b) Did you discover anything new or learn anything about your rating from stimulated recalls? Did stimulated recalls make you think of new things to try when rating essays in the future?
   c) What did you think about my interaction with you? Did you notice anything specific about my interaction with you? If yes, at that time, why did you think I was doing that?

5. **Other Comments:** Do you have any thoughts or comments that you would like to add?

Thank you for your cooperation!
Appendix E Individual Interview Guide

SECTION ONE: THE BROADER CONTEXT

(FOR ALL PARTICIPANTS)

1. Perceptions about Gaokao & its rating:
   1) What do you think about Gaokao (e.g., its purposes, role, impact, etc.)?
   2) Gaokao rating as a political mission: What do you think about Gaokao rating (e.g., its purposes, role, impact, etc.)? Someone says Gaokao rating is a political mission. What do you think about this comment?
   3) Openness and transparency of rating: What do you think about the procedures and processes of rating Gaokao papers? In recent years, during Gaokao rating sessions news reporters, representatives of test-takers and their parents are allowed to visit the rating venues and get to know how the Gaokao test papers are marked. What do you think of these initiatives?
   4) Problems with Gaokao rating: What do you think about the monitoring of rating quality? The Gaokao rating venue in Guangxi province was reported to have speed rating issue while marking Gaokao papers. What comments do you have regarding this news report?
   5) Slogans: What do you think of the slogans posted at the rating centre and in your school/classroom? What do you think of these slogans (show to participants the photo or script of slogans found at the rating centre)?
      a) 铁肩担道义 分分不负寒窗功
      b) 心系千万考生 细心公正阅卷
      What do you think of the slogans below? Which one do you feel most relevant to you? Why?
      a) 今生只为高考狂, 冲进重点孝爹娘
      b) 提高一分, 干掉千人
      c) 考过高富帅, 战胜“官二代”!

SECTION TWO: AT THE RATING CENTRE

(FOR DIRECTORS, TEAM LEADERS, & RATERS)

2. Self-perceived role and responsibility: Could you summarize your role at the rating centre? What are your responsibilities? How many years have you been serving this role? What is your work at the rating centre like every year?

3. Motives: What motivated you to serve this role? What goals do you want to achieve? Why?

4. Perceived sociocultural context at the rating centre:
   1) What do you think about this rating centre?
   2) Qualifications and selection criteria of raters: What do you think about the essay raters at this centre (e.g., qualifications, quality, professionalism)?
3) **Organizational dynamics & security and confidentiality issues:** What do you think about the sociopolitical atmosphere at this centre (e.g., organizational structure and dynamics, security and confidentiality measures and instruction)?

4) **Measures/efforts taken to ensure rating quality:** What do you think about the monitoring of rating quality at this rating centre (e.g., measures, procedures, effectiveness)? If there is any measure you can suggest the centre to take to improve rating quality, what would that be?

5. **Relationship with others:** What roles and responsibilities do you think the director/team leaders/raters should take? What types of relationship do you have with them? What types of relationship do you have with your colleagues at the rating centre?

6. **Use of documents and artifacts** (including NMET rating criteria used to guide the NMET rating): What documents and artifacts have you created, collected, and/or used during/after the NMET rating session? What are the purposes or importance of these documents (e.g., rating scale and specifications, test specifications, work notes)? How useful do you find these documents are to your rating work at the centre? Please explain with examples.

7. **Challenges:** What common challenges or problems, if any, did you encounter when rating the NMET essays? How did you solve them? If you did not, why? Please give an example.

8. **Efforts for improvement:** What strategies did you take to improve your own NMET essay rating quality? Why did you choose these strategies? Are you finding them as effective as you expected? Why do you think these strategies are effective or ineffective? How do you think the situation can be improved?

**SECTION THREE: AT HIGH SCHOOLS**

(FOR VICE PRINCIPALS, TEACHER RATERS, & TEAM LEADERS/TEACHERS)

9. **Self-perceived role and responsibility:** Could you summarize your role at school? How long have you been serving this role? What are your responsibilities related to Gaokao/Gaokao rating? What is your work at school like every year that is related to NMET/NMET rating?

10. **Motives:** What motivated you to recommend/support XX (name of the NMET rater)/serve as NMET rater? What goals do you want to achieve? Why?

11. **Perceived sociocultural context at school:**
   1) What do you think about work related with Gaokao rating at your school?
   2) **Importance of Gaokao rating:** What do you think about the sociopolitical atmosphere regarding Gaokao rating at your school (e.g., importance attached, the kind of support before/during/after Gaokao rating session)? Why do you think your school attach (not) much importance to Gaokao rating?
   3) **Gaokao rating informing test preparation training:** What do you think about the Gaokao rating experience (e.g., purpose, importance, impact)? Some raters found NMET rating experience informs their test preparation training afterward. What comments do you have regarding this opinion? Are you finding any changes caused by your/other raters’
NMET rating experience to teaching practices? If yes, can you describe them? Why has that experience led to such changes? Please explain with examples.

12. **Relationship with others:** What roles and responsibilities do you think your vice principal/teachers with NMET rating experience should take regarding your NMET rating task? What types of relationship do you have with them?

13. **Use of documents and artifacts:** What documents and artifacts related to NMET rating have you created, collected, and/or used in your teaching practices (e.g., rating scale and guidelines, test specifications, work notes)? What are the purposes or importance of these documents? How useful do you find these documents are to your teaching? Please explain with examples.

**CONCLUSION:**

14. **Concluding questions (at the end of the interview):** Do you have anybody in mind that you think I ought to speak to for my research? Is there anything I should know that I have not asked you?

Thank you for your participation!
# Appendix F Transcription Conventions for Verbal Protocols

(Adapted from Barkaoui, 2008, p.269)

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>[ ]</td>
<td>For uncertain transcription</td>
</tr>
<tr>
<td>x</td>
<td>Incomprehensible item, one word only</td>
</tr>
<tr>
<td>xx</td>
<td>Incomprehensible item of phrase length</td>
</tr>
<tr>
<td>xxx</td>
<td>Incomprehensible item beyond phrase length</td>
</tr>
<tr>
<td>, (comma)</td>
<td>Shorter pauses of less than 3 seconds</td>
</tr>
<tr>
<td>?</td>
<td>Questioning intonation</td>
</tr>
<tr>
<td>…(3 dots)</td>
<td>Incomplete or interrupted utterances (rising, prolonged intonation)</td>
</tr>
<tr>
<td>CAPITAL LETTERS</td>
<td>(in English) Text read directly from essay (including single words read from essay)</td>
</tr>
</tbody>
</table>
| Bold type | (a) (if in Chinese) Text translated directly from essay (including single words translated from essay);  
(b) (if in English) Rising intonation for emphasis/stress (e.g., stress an error while reading) |
| Italics | Text read from the prompt Text |
| Underlined | Text read directly from the rating scale |
| ( )    | Procedural and other behaviors (e.g., sorting essays, laughs, coughs) |
| - (hyphen) | An incomplete word (e.g., wait plea-, wo-) |
Appendix G Selected Sample Essays

Essay 14-04

Dear Sir,

I am a student from China and I want to study English in the U.K. for 6 weeks in summer vacation. Your advertisement drew my attention. I want to know something about you. First, when would I start my class? Second, what is the size of class? Third, how many hours per week would I take? Then, how much fees I will pay for? Last, what’s the type of accommodation? I want you can reply me that I could know something more about you.

Yours,

Li Hua
To whom it may concern.

During this summer vacation, I'm Li Hua, a senior school student. I'm going to UK to promote my English study for six weeks. When I surfed the Internet, I found your advertisement which made me excited. So I decided to participate in your school's education plan.

Would you mind answering my questions? First and foremost, I know it has 3 weeks, 6 weeks and 12 weeks courses. Could you please tell me when it starts? Meanwhile I also want to know how many students in a class. What's more, how many the hours the class take? Last but not least, it said that it has reasonable fees so how much should I pay for it? I really concerned about the type of accommodation. If possible, I want to have an English roommate. Thank you in advance.

Looking forward to your reply.

Yours sincerely,

Li Hua
Dear Peter,

I’d like to ask you to write an article for our school’s English newspaper. The cultures overseas part is one of the most popular parts of our newspaper. Every time we post a few articles and pictures to introduce two specific cultural events, costumes or daily life abroad.

We’d appreciate it if an American citizen can write an article to post on our paper. And we want you to discuss about two topics: One is American customs of national holidays. And the other is daily life of your high school. We want you to make comments vividly as much as you can in around 400 words.

Our deadline is June 28th. If you are pleased to write the article, please send it to me before the date.

Yours,
Li Hua
Dear Peter,

I’d like to ask you to write an article for our school’s English newspaper.

“The Fongch Culture” is the most attractive part in our school’s English newspaper, saying about 20 cultures and lifestyles of 20 different culture, which is our favorite part.

The article There are two points in the article I ask you write. First of all, you are supposed to introduce the festival culture in America to us. Then, I hope you can write some experience in you Senior High students’ daily life. Besides, there are two asks to you. The one is that the article need to contain about 400 words. The other is that you must send your article to me before June 28th.

Thanks for writing my letter.

Yours,
Li Hua
Appendix H Selected Think-Aloud Protocols

Ri-juan, Essay 14-04

接下来[我]看四号卷，四号卷，哦，这个[卷子]...第一眼看上去...文章没有分段，没有分段这个是会影响到我打分的，我觉得一个作文它应该是分一下段，第一段说出为什么写这封信，然后第二段具体的内容，到第三段说出自己的一点点...比如说美好的祝愿xx，所以这样的话这个文章...第一眼看上去那就应该是[归档]在四档往下也就是及格往下，及格往下，那么我需要再看一下它的...这个要点，6 WEEKS IN SUMMER VACATION，xxx FIRST，x WHEN WOULD I START MY CLASS，SECOND，WHAT IS THAT？THIRD，HOW MANY HOURS PER WEEK WOULD I TAKE，THEN HOW MUCH FEES WILL I PAY FOR？LAST，WHAT'S THE TYPE OF ACCOMMODATION.I WANT YOU CAN REPLY ME THAT I COULD KNOW SOMETHING MORE ABOUT YOU. FOR SIX WEEKS...xxx[文章]倒是要点是写出来了，要点写出来的话，而且[文章]几乎没有...很大的语言错误，那这样的话...就是...[归档]上一档吧，原来定在三档，那么可以到四档，当然四档也就是一个16分，因为这个它里面的...里面的要点倒是写出来了，文章[似乎是口误，应为“评分标准”]要求的内容，而且当然最后[文章]也写出了一点点希望，就是...中间这里，它就是没有分段，[这]一种[给我]感觉不是很好，当然[它还有]涂抹，还有就是句子与句子之间...逗号太多，没有断句，[所以]这是这个四号的16分。

Ri-xiu, Essay 15-57

57号卷，这个作文整体上来看，卷面不是很整洁，[我]看一下这个文章中有没有包含[评分标准]要求的6个点在内...（reads the essay silently）它这里头用了一个CULTURES OVERSEAS MODEL，这个表述首先就不准确，不应该算，就是说它缺少了国外文化栏目这个要点。再一个就是它这里头说到什么TWO SPECIFIC CULTURAL EVENTS，COSTUMES OR DAILY BASIS ABROAD，这些都表述得不是很清楚，[这个要点]基本上算没有包含进去。再一个就是（reads the essay）AN AMERICAN CITIZEN CAN WRITE AND ARTICLE TO POST ON OUR PAPER, AND WE WANT YOU TO DISCUSS ABOUT TWO TOPICS: ONE IS AMERICAN CUSTOMES OF NATIONAL HOLIDAYS, THE OTHER IS DAILY LIFE OF YOUR HIGH STUDENTS。这个DAILY LIFE OF YOUR...OUR DEADLINE IS JULY，我们说的是6月28[日]，他这是7月28[日]，然后字数要求他这是AS YOU CAN IN AROUND 400 WORDS，整体上这个作文不行，[我]归档可以归到应该是个二档作文6到10分的作文[，归到二档]也是有难度的。DISCUSS ABOUT…(scans the essay again)几乎这个内容就没有表达清楚，[这]就是一个一档作文，而且最多也就是...信息也漏掉很多，就是[包括了]400字、日期，而且就是关于美国风俗和中学生生活这个[要点]引出也是到后面才引出了这么几个，所以最多[给打]5分吧。这个作文最多5分。
Appendix I Coding Scheme (with Examples) for Think-Aloud Protocols and Stimulated Recalls

<table>
<thead>
<tr>
<th>Code</th>
<th>Examples¹⁰</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Self-monitoring Focus</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Interpretation strategies</strong></td>
<td></td>
</tr>
</tbody>
</table>
| (Re)Read or translate composition | (Translating) 我不想去…我不想太多学生在班里。我感到应该是这个单词写错了。我觉得是…想知道多少…付多少钱，多少小时在这个多少。(Ri-wei, T, 14-03)  
(Translating) I DON’T WANT TOO…I DON’T WANT TOO MANY STUDENTS IN CLASS. I FEEL… This word must be spelled wrong. I FEEL… WANT TO KNOW HOW MUCH… PAY HOW MUCH MONEY, HOW MANY HOURS DURING THIS. |
|                               | DESCRIBE THE AMERICAN FESTIVALS AND THE STUDENTS’ DAILY LIFE, SUCH AS, WHAT YOU DO IN WEEKEND, WHAT YOU LEARN IN YOUR SCHOOL DAYS. WITH YOUR ARTICLE, I THINK THE STUDENTS IN OUR SCHOOL WOULD WANT TO BE A…这块看不清。EXCHANGE STUDENT 想成为一个交换生，去美国的交换生，这可能．．．(Ri-dong, T, 15-63)  
DESCRIBE THE AMERICAN FESTIVALS AND THE STUDENTS’ DAILY LIFE, SUCH AS, WHAT YOU DO IN WEEKEND, WHAT YOU LEARN IN YOUR SCHOOL DAYS. WITH YOUR ARTICLE, I THINK THE STUDENTS IN OUR SCHOOL WOULD WANT TO BE A… I cannot recognize this word. WANT TO BE AN EXCHANGE STUDENT, AN EXCHANGE STUDENT TO AMERICA, WHICH CAN… |
| Envision personal situation of the writer | I FEEL 后面这个 NOTICE 我感觉应该是用错词了这个，应该感到紧张什么的。(Ri-wei, S, 14-03)  
The word NOTICE after I FEEL, I think [the student] must be nervous and have used the wrong word. |
|                               | There are two points, (translating) 我要求你写的这篇文章有两个点。这完全就是汉语化的汉式英语，没有转变成英语思维，完全就是把汉语翻译成英语的这种语言，基础不行。(Ri-xiu, T, 15-64)  
THERE ARE TWO POINTS, (translating) THERE ARE TWO POINTS IN THE ARTICLE I ASK YOU TO WRITE. This is |

¹⁰ Each quote is followed by rater’s pseudonym, data source code (T = Think-aloud protocol, S = Stimulated recall), essay code, and translation.
completely Chinglish. [The student] was not thinking in English, just translated the Chinese into English. [He] must have a weak foundation for English.

Scan whole composition

The last is Essay 19. This essay looks, see its handwriting, the handwriting is very neat. And the structure, at a glance, it is divided into three sections. Now let me take a look at this essay.

六号作文它这个分段字迹也比较清晰，这篇作文第一印象就是感觉就是比较符合四档作文。(Ri-wei, S, 14-06)

Refer to or interpret rating scale

Essay 6 has several paragraphs and writes with clear handwriting. My initial impression of this essay is it matches in Band 4.

(Translating)"FOREIGN CULTURE" IS THE MOST ATTRACTIVE PART IN OUR SCHOOL’S ENGLISH NEWSPAPER, TALKING ABOUT CULTURES AND LIFESTYLES OF DIFFERENT CULTURE, WHICH IS OUR FAVORITE PART. This is talking about introduction to the column.

这篇作文书写比较认真，内容都提到了，印象还是不错的，所以这个考虑是第五档作文，就是它就是完成了试题规定的任务。(Ri-wei, T, 14-12)

Judgement strategies

When I am rating the essays, I would ignore the details added by the writer, and use my eyes to scan and search for the major details. Now I skip to his last paragraph.

首先这个文章关键应该看第二段，第二段就是是否把要求
<table>
<thead>
<tr>
<th>Section</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>First of all, the key of this essay lies in the second paragraph.</strong></td>
<td>I will see whether the aspects required to consult were included in this paragraph, then make a band range decision.</td>
</tr>
<tr>
<td><strong>Compare with other essays/test-takers</strong></td>
<td>Look at all these mistakes he made, that is quite a lot, but he is slightly better than the worst student.</td>
</tr>
<tr>
<td><strong>Define own rating criteria</strong></td>
<td>Because I pay more attention, especially in letters, I pay more attention to the first and last paragraphs.</td>
</tr>
<tr>
<td><strong>Articulate general impression</strong></td>
<td>The only drawback of this essay is the student did not add a bit more details. Otherwise, it could have been place in Band 5.</td>
</tr>
<tr>
<td><strong>Articulate general impression</strong></td>
<td>This essay writes at one stretch. This one is also well-written.</td>
</tr>
<tr>
<td><strong>Summarize, distinguish, or tally judgments collectively</strong></td>
<td>看完之后我会觉得要点全了。但是在最后这个段落中，他有一个谢谢你写错了，这个也会影响一定的分数。整个文章的文章要点是全的，包括美国学生的生活，也包括字数，</td>
</tr>
</tbody>
</table>
After reading [this essay], I think [it] includes all the major details. But in the last paragraph, he wrote THANK YOU, it is wrong. This will affect his score. The entire essay has all the major details, including American students’ life, also including word count and due date.

细节错误不超过3处，很好，而且用了很多高级词汇，有大错误有两处。但总的来说多处用到高级词汇，卷面非常整洁，字迹非常好。 （Ri-xia, T, 14-19）

No more than three errors in details. Very good. [This student] also used many advanced-level vocabulary, had two major errors. But generally speaking, [he] used advanced-level vocabulary in multiple places; the answer sheet is very tidy, and [his] handwriting is very neat.

<table>
<thead>
<tr>
<th>Articulate or revise scoring decision</th>
<th>给它定位到第三档左右吧，最起码得第三档。 (Ri-zhen, T, 14-12)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>[I decide] to put this essay into Band 3, should be at least Band 3.</td>
</tr>
</tbody>
</table>

那这样的话就是上一档吧，原来定在三档，那么可以到四档，当然四档也就是一个16分。 (Ri-juan, T, 14-04)

If so, [I decide] to upgrade [it] by one band. [My] original decision was in Band 3, now [it] can be put in Band 4. Well, although [it is] in Band 4, [it] only deserves 16 points.

所以这个得满分呢是…得满分是不可能了。但是它可以归在这个最高档，也就是第五档，21分到25分之间。可以说是这个卷面整体看，这篇文章应该能上21或者是22分。 (Ri-yue, S, 14-19)

So this essay would be assigned a full score. . .not possibly. But it can be put in the highest band, that is Band 5, between 21 and 25 points. Judging from the tidiness of the answer sheet and overall quality, this essay should be assigned 21 or 22 points.

Rhetorical and Ideational Focus

<table>
<thead>
<tr>
<th>Interpretation Strategies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interpret ambiguous or unclear phrases</td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

(Translating) 我们创建了这么一个栏目，项目？这就有一点
不适合了，他想表达这个意思但是词没用对。（Ri-xiu, T, 14-18）

(Translating) WE SET UP A PROGRAM, PROGRAM? This is not appropriate here. He wanted to express this idea ["column"] but used the wrong word.

Judgement Strategies

Assess major details

第一个要点在第一段要提到的是要去那干嘛，然后[第二个要点]“去上六个星期的课程”，嗯，好，在第二段出现了。（Ri-juan, T, 14-03）
The first major detail that needs to be mentioned in the first paragraph is what you want to do when you are there. Then [the second major detail] “to take a six-week course”, em, well, [is] present in the second paragraph.

然后一看他的细节，说是去哪儿去哪儿学六个星期。我第一个关注“六个星期”，然后就看看说是要求的时间有了，然后问的是内容要点：时间有了，教室的大小也有了，还有上多少节课有了，这个付多少费有了。（Ri-zhen, S, 14-06）

Then I was reading the details. He [student writer] mentioned to go somewhere to study for six weeks. I first looked for “six weeks”, then I saw he mentioned it as required. Then he asked questions, which include the major details: start date, check; class size, check; hours per week, check; how much to pay, check.]

Assess relevance to central topic

后面第二自然段写这个主要的细节的时候，也是扯到这儿扯到那儿瞎扯了一顿。（Ri-zhen, S, 14-03）

Later, in the second paragraph when [the student] wrote the major details, he talked about this and that, going off the topic.

18号的文章刚开始WITH THE DEVELOPMENT OF ECONOMY, MORE AND MORE CHINESE CHOOSE TO GO ABROAD TO EXPERIENCE A DIFFERENT CULTURE AND LIFESTYLE，就是好像写的有点稍微远了一点。但是就是发挥的过程，有到三行的内容，三行多一点的内容，没有切到主题上，一直到第四行我才看到了他这个报纸的专栏NAMED FOREIGN CULTURE。（Ri-ling, T, 15-18）

The beginning of Essay 18 writes “WITH THE DEVELOPMENT OF ECONOMY, MORE AND MORE CHINESE CHOOSE TO GO ABROAD TO EXPERIENCE A DIFFERENT CULTURE AND LIFESTYLE”. This sounds a bit off the topic. These added information occupied three lines, a bit more than three lines, not relevant to the topic. I did not
see him mentioning the section in the newspaper NAMED FOREIGN CULTURE until in the fourth line.

**Assess task completion**

这个学生没有完成试题所规定的内容，只写了两个要点，文章没有写完。(Ri-ling, T, 15-63)
This student did not complete what is required in the writing task. [He] only wrote about two major details. The essay was not finished.

这个文章尽管是只有一段，字迹稍微潦草，但是基本没有错误，而且能够清楚的表达出作者的写作意图，写作的目的。 (Ri-yue, T, 14-04)
Although this essay only has one paragraph, and the handwriting is a bit illegible, it has few errors and could clearly convey the writing purpose.

**Assess coherence and cohesion**

用了个 MEANWHILE 你看这个过渡词语，同时 I ALSO WANT TO KNOW HOW MANY STUDENTS IN THE CLASS，这个过渡词也挺好的。下一句又是 WHAT'S MORE 开头，另外，也用得非常好，HOW MANY THE HOURS THE CLASS TAKE…LAST BUT NOT LEAST 你看这些过渡用得多好。(Ri-wei, S, 14-19)
[He] used the word MEANWHILE, look at this nice transition word. MEANWHILE, I ALSO WANT TO KNOW HOW MANY STUDENTS IN THE CLASS, this is also a good transition. Next sentence starts with WHAT'S MORE, well done. HOW MANY THE HOURS THE CLASS TAKE…LAST BUT NOT LEAST. Look, how well these transitions were used.

但是他这个句与句之间的连接过渡不是特别自然。用那些"FIRST"还是那个…FIRST、SECOND、AT LAST 有点列举的感觉就不是特别好，感觉不如用很自然的过渡词要好一些。(Ri-wei, T, 14-08)
But the sentential transitions were not very natural. He used the words such as FIRST, or FIRST、SECOND、AT LAST, they read like a list, do not feel well. They do not feel as smooth as using some natural transitions.

**Assess text organization**

这个不需要分段，接下来的它这个二、三、四应该合成一个自然段就好，因为讲的都是一个内容，主要是问这个详细情况，所以它这分段的有点不合适。(Ri-wei, S, 14-10)
Here [it] does not need more than one paragraph. The following second, third, and fourth points can be combined
into one paragraph, because they all talk about the same thing, asking for more details. Dividing them into several paragraphs is a bit inappropriate here.

The first paragraph gives a brief introduction at the beginning; the second paragraph is the body, asking questions; the third paragraph is a concluding sentence. The structure is quite clear.

Language Focus

<table>
<thead>
<tr>
<th>Interpretation Strategies</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Observe layout</strong></td>
</tr>
<tr>
<td>(e.g., paragraph, text length)</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td><strong>Consider legibility</strong></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td><strong>Classify errors into types</strong></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>
Usually [we] add a definite article “the” in front of UK. Proper nouns should have “the” before them. But he did not add one. (Ri-yue, T, 14-06)

There is one error in an attributive clause in the middle part. He used interrogative word order, instead of declarative word order.

中间有一个宾语从句的错误，他用的是疑问语序，没有用陈述语序。中间有一个宾语从句的错误，他用的是疑问语序，没有用陈述语序。（Ri-yue, T, 14-06）

Edit phrases for interpretation

下边有一个 HAD BETTER TO DO，多了一个 TO。 下边有一个 HAD BETTER TO DO，多了一个 TO。（Ri-ling, T, 15-63）

Later there is a HAD BETTER TO DO; there is an extra TO in it.

HAVE SOME QUESTION 我要把 QUESTION 换成 questions. (Ri-juan, T, 14-08)

HAVE SOME QUESTION, I will change QUESTION into “questions”.

Judgement Strategies

Consider syntax or morphology

这个定到五档，而且有亮点的句子很多，五档往上，有一 个定到五档，而且有亮点的句子很多，五档往上，有一些小小的语法错误。 (Ri-juan, T, 14-19)

This [essay will be] put in Band 5. [It] contains many sentences with shining points. [It can be put] in the upper end of Band 5, with a few very minor grammatical errors.

在这儿 SO MANY 画了一个横线，MONEY 那里面不可以用 many。 (Ri-juan, T, 14-03)

I underlined SO MANY, because MONEY cannot be modified by MANY.

Consider lexis

他一直用 PROGRAM 这个词，他可能对 column, part 这些 他一直用 PROGRAM 这个词，他可能对 column, part 这些 词就不太熟悉。program 项目，他一直都是在用那个词，词就不太熟悉。program 项目，他一直都是在用那个词，这就不对了，不能是一个项目。（Ri-xiu, S, 14-18）

He kept using the word PROGRAM. He is probably not familiar with the words such as “column” and “part”.

PROGRAM indicates a plan. He kept using this word. This is not right; it is not a plan.

WHEN I SURFED…I FOUND…I FOUND…WHICH MADE ME EXCITED…WHICH MADE ME EXCITED…DECIDED TO PARTICIPATE…DECIDED TO PARTICIPATE…WHICH MADE ME EXCITED…DECIDED TO PARTICIPATE…

WHEN I SURFED…I FOUND…I FOUND…WHICH MADE ME EXCITED…WHICH MADE ME EXCITED…DECIDED TO PARTICIPATE…DECIDED TO PARTICIPATE…WHICH MADE ME EXCITED…DECIDED TO PARTICIPATE…
<table>
<thead>
<tr>
<th>Consider frequency and gravity of error</th>
<th>她的错误太多了，大的错误有好几处，然后这个细节错误也有好几处。 (Ri-xia, S, 14-03)  She made too many errors, including several grave errors, as well as several minor errors in details.</th>
</tr>
</thead>
</table>
| Consider shining points (advanced-level lexical and syntactic resources) | 前边是亮点，这个 AS WELL AS，这个很地道。 (Ri-dong, S, 15-53)  There is a shining point at the beginning. This AS WELL AS, this is very native-like.  
第一个首先知道同位语的这种使用办法，然后一个句子非常的长，但是信息全部包含了，而且没有任何的错误，这儿还知道用 WHICH 这种定语从句，这也算是一个亮点吧。 (Ri-xia, T, 14-19)  
First, [this student] knew using appositives in this way, and there is a very long sentence, and all the information was included, with no error. [He] also knew to use an attributive clause guided by WHICH. These can be counted as a shining point.  
应用了大量的高级词汇，这就有了亮点了。 (Ri-xia, S, 14-19)  
[The student] used a lot of advanced-level vocabulary, this is a shining point. |
| Assess comprehensibility, fluency, and naturalness | WHETHER ARE WE LIVE 这个句子看不懂，不知道他在写什么。 (Ri-juan, S, 14-10)  
WHETHER ARE WE LIVE, I do not understand this sentence, not knowing what he was writing about. |
I see BESIDES (translating)除了这个之外...GET ALONG WITH OTHERS TO IMPROVE OUR ENGLISH...HOW MUCH I HAVE TO PAY, this sentence also works, fluent. (Ri-zhen, T, 14-12)

Let me see, BESIDES, (translating) BESIDES...GET ALONG WITH OTHERS TO IMPROVE OUR ENGLISH...HOW MUCH I HAVE TO PAY. This sentence reads fine; the English is fluent.

However, PAY HOW MUCH this is purely Chinglish. It should be something like HOW MUCH HOURS SHOULD I SPEND. She used a purely Chinglish expression.

**Consider handwriting and answer sheet tidiness**

你看这学生字迹写的也非常好，就像印出来地那么工整。 (Ri-xia, T, 14-19)

Look, this student’s handwriting is very neat, looks like printed work.

他的字也不整洁，而且划成这样。 (Ri-xiu, S, 15-40)

His handwriting is not neat, and he crossed out many places like this.

这篇作文字迹工整，给人感觉非常清爽。 (Ri-wei, T, 15-21)

The handwriting of this essay is neat, making me feel very refreshed.

**Consider spelling and punctuation**

写错单词了，”六月”写成”七月”。(Ri-wei, T, 15-57)

This word spelled wrong. JUNE was written into JULY.

这个文章句子与句子之间的句号特别少，断句不行。(Ri-juan, S, 14-04)

This essay has very few full stops between sentences; its punctuation is very poor.