Project Summary

Motivation for the Research

Essay rating research in language assessment has largely focused on human raters’ essay rating as a solely cognitive process of information processing (e.g., Bejar, 2012; Freedman & Calfee, 1983) or problem-solving (e.g., Crisp, 2010; DeRemer, 1998). These studies described how individual raters’ rating processes and 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; Li & He, 2015; Milanovic, Saville, & Shen, 1996; Weigle, 1994). However, the decontextualized view taken in these studies does not address the interactions between raters and the sociocultural contexts where the essay rating takes place, often leading to inconsistent findings across different contexts. Attention has also recently been given to the social aspect of essay rating in language assessment research (e.g., Baker, 2010; Lumley, 2005). The oversimplification of the sociocultural contexts in these studies fails to fully address rater-context interactions, thus reiterating the need for a comprehensive, situated understanding of essay rating. Essay raters are social beings who rate written work composed by other social beings. Essay rating is hence a socially situated activity with socially constructed meanings, motives, and consequences (Barkaoui, 2008). A study that situates this activity within its sociocultural context can make a valuable contribution to the literature.

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). This study explored raters’ essay rating activity (ERA) at one provincial rating center in China within the context of the high-stakes National Matriculation English Test (NMET). NMET is situated in the Chinese testing-driven society (Cheng & Curtis, 2010), and is the English component of the university entrance examination (known as Gaokao). Each year, over nine million test takers write Gaokao, and their results from four component examinations (English, Chinese, mathematics, sciences or social sciences) exclusively determine their university admission decisions. Gaokao exerts a huge impact on numerous stakeholders (e.g., students, teachers,
parents, high schools, universities), and strongly influences teaching and learning (i.e., washback effect) in Chinese secondary education (Cheng & Qi, 2006; Gu, 2013; Qi, 2005). To obtain a more situated understanding of the ERA involved in its English component test at one provincial rating center, three research questions were addressed:

1. How do raters assess NMET essays to achieve their goals?
2. What are the broader (Chinese society) and immediate (rating center 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?

Methodology
The study adopted a multiple-method, multiple-perspective qualitative case study design, including data collected through think-aloud protocols, stimulated recalls, interviews, and the analysis of documents. There were 25 participants involved from two settings (the rating center and high schools), including rating center directors, team leaders, NMET essay raters who were high school teachers, and these raters’ school principals and teaching colleagues. Data were analyzed using open and axial coding techniques (Strauss & Corbin, 1990), and CHAT for data integration.

Research Findings
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 system);
- high-stakes consequences of raters’ ratings to student writers;
- saving “face” by staying close to others’ ratings;
- prior teaching and rating experience; and
- advice from colleagues.

These findings were then situated in findings of the social layer in the context of the rating center, where NMET raters performed their ERA, and the high schools, where raters taught. The social layer findings showed that the high-stakes nature of 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. They thought the NMET ERA they participated in was a sacred mission with grave responsibility, and were under high stress, challenged by the rating center requirements and the pressure to save “face” by staying within the interrater agreement of their peers. On the other hand, raters thought their NMET ERA experiences were beneficial to their teaching practices, where one of their priorities
was to improve student NMET performance. Findings from this layer contributed to understanding the sociocultural context of NMET ERA under study.

A CHAT analysis of these findings further revealed the interaction between raters and the NMET sociocultural context. The cognitive layer reflects what raters’ decision-making looked like, and the social layer explains how and why raters’ decision-making worked in that way. The two layers are interrelated through a series of interactions between raters’ cognition and the activity sociocultural context, subsequent rating tensions, and raters’ corresponding solutions. This bilayer conceptualization can explain why raters may take similar actions (e.g., attending to similar essay features, following similar rating sequence, and considering similar factors) to solve different tensions, or adopt different actions to solve similar tensions. For example, raters regarded their rating quality indicators (e.g., valid rating rates, serious rating error rates) to identify whether or not their rating performance was deviant from other raters. 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 looked less competent than other raters, a potential threat to their “face”. These concerns were associated with raters’ rating goals (rating accurately and fast, holding accountable for student writers) and the institutional and sociocultural rules (following rating criteria, the Gaokao having high stakes for test-takers, saving “face”) that guided their behaviours. These concerns would subsequently lead raters to adopt various actions and try to keep their ratings close to their peer raters, which is more complex and richer than the “play it safe” concern identified in previous studies (e.g., Knoch, Read, & von Randow, 2007; Myford & Mislevy, 1995). The raters may consult team leaders, communicate with other raters, and sometimes think of possible scores assigned by second raters. The 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). These findings highlighted the roles of goals and rules in rater decision-making, in addition to rating tensions and raters’ corresponding solutions.

Another unique finding of this study is the relationship between essay rating and teaching. In support of previous findings about the presence of influences associated with raters’ teaching experiences (e.g., Cumming, 1990; Eckes, 2008; Hamp-Lyons, 1989; Pula & Huot, 1993), my study revealed a more interactional and dynamic relationship between essay rating and teaching in this NMET context. From their preparticipation context in their teaching communities, teachers brought not only their teaching experience to assist in decision-making, but were also incentivized to participate for the sake of their professional development, which was associated with the high importance of NMET to their teaching careers; then during their ERA at the rating center, these teachers not only completed the task of rating essays by relying on their prior teaching experiences, 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.

This study 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. The study has three major contributions to language assessment research. First, it highlights the value of a sociocultural view to essay rating research and postulates a bilevel 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. Finally, based on findings from the CHAT analysis, this study provides a direction for future washback studies, implying that teachers’ involvement in high-stakes rating may lead to a potential washback effect on teaching practices.

Implications

My study also has two major implications for language education practitioners and policy makers. First, this study provides support for improving NMET rating practices and potentially positive washback in Chinese high school English language 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 teachers’ professional development, which could potentially bring positive washback to high school teaching. NMET rating centers should consider providing every high school teacher with periodic opportunities to participate in NMET essay rating for professional development purposes. Second, this study suggests the practical value of applying CHAT to improving essay-rating practices. The study demonstrates that a CHAT analysis can help researchers better understand what works and what does not in essay rating practices and provide corresponding support, with implications for practices in other contexts.
References


