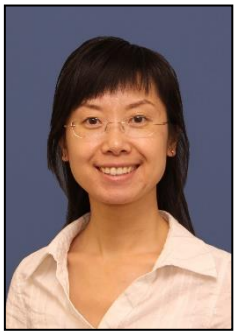
The International Research Foundation
for English Language Education

Title of Project:
Understanding Essay Rating as a Socially Mediated Activity:
The Case of a High-Stakes English Test in China

Researcher:
Yi Mei
Queen's University
yi.mei@queensu.ca

Research Supervisor:
Dr. Liying Cheng
Queen's University
liying.cheng@queensu.ca



Yi Mei

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,

177 Webster St., # 220, Monterey, CA 93940 USA
Web: www.tirfonline.org / **Email:** info@tirfonline.org

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

177 Webster St., # 220, Monterey, CA 93940 USA
Web: www.tirfonline.org / **Email:** info@tirfonline.org

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

177 Webster St., # 220, Monterey, CA 93940 USA
Web: www.tirfonline.org / **Email:** info@tirfonline.org

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.

177 Webster St., # 220, Monterey, CA 93940 USA
Web: www.tirfonline.org / **Email:** info@tirfonline.org

References

Abdul Kadir, K. (2008). *Framing a validity argument for test use and impact: The Malaysian public service experience* (Unpublished doctoral dissertation). University of Illinois at Urbana-Champaign, Champaign, Illinois, United States.

Alderson, J. C., & Wall, D. (1993). Does washback exist? *Applied Linguistics*, *14*(2), 115–129.

Attali, Y., Lewis, W., & Steier, M. (2012). Scoring with the computer: Alternative procedures for improving the reliability of holistic essay scoring. *Language Testing*, *30*, 125–141. http://doi.org/10.1177/0265532212452396

Au, W. (2007). High-stakes testing and curricular control: A qualitative metasynthesis. *Educational Researcher*, *36*(5), 258–267. http://doi.org/10.3102/0013189X07306523

Axel, E. (1997). One developmental line in European activity theories. In M. Cole, Y. Engeström, & O. Vasquez (Eds.), *Mind, culture, and activity: Seminal papers from the laboratory of comparative human cognition* (pp. 128–146). Cambridge, UK: Cambridge University Press.

Baird, J.-A., Greatorex, J., & Bell, J. F. (2004). What makes marking reliable? Experiments with UK examinations. *Assessment in Education: Principles, Policy & Practice*, *11*(3), 331–348. http://doi.org/10.1080/0969594042000304627

Baker, B. (2010). Playing with the stakes: A consideration of an aspect of the social context of a gatekeeping writing assessment. *Assessing Writing*, *15*, 133–153. http://doi.org/10.1016/j.asw.2010.06.002

Baker, B. (2012). Individual differences in rater decision-making style: An exploratory mixed methods study. *Language Assessment Quarterly*, *9*, 225–248. http://doi.org/10.1080/15434303.2011.637262

Barkaoui, K. (2007a). Participants, texts , and processes in ESL/EFL essay tests: A narrative review of the literature. *The Canadian Modern Language Review*, *64*, 99–134.

Barkaoui, K. (2007b). Rating scale impact on EFL essay marking: A mixed-method study. *Assessing Writing*, *12*, 86–107. http://doi.org/10.1016/j.asw.2007.07.001

Barkaoui, K. (2008). *Effects of scoring method and rater experience on ESL essay rating processes and outcomes* (Unpublished doctoral dissertation). The Ontario Institute for Studies in Education of the University of Toronto, Toronto, Ontario, Canada.

Barkaoui, K. (2010a). Do ESL essay raters' evaluation criteria change with experience? A mixed methods, cross-sectional study. *TESOL Quarterly*, *44*, 37–41. http://doi.org/10.5054/tq.2010.214047

Barkaoui, K. (2010b). Variability in ESL essay rating processes: The role of the rating scale and rater experience. *Language Assessment Quarterly*, *7*, 54–74. http://doi.org/10.1080/15434300903464418

Barkaoui, K. (2011a). Effects of marking method and rater experience on ESL essay scores and rater performance. *Assessment in Education: Principles, Policy & Practice*, *18*, 279–293. http://doi.org/10.1080/0969594X.2010.526585

Barkaoui, K. (2011b). Think-aloud protocols in research on essay rating: An empirical study of their veridicality and reactivity. *Language Testing*, *28*, 51–75. http://doi.org/10.1177/0265532210376379

Barrett-Mynes, J. (2013). *Literacy instruction in the wake of Common Core State Standards* (Unpublished doctoral dissertation). Georgia State University, Atlanta, Georgia, United States.

Bejar, I. I. (2012). Rater cognition: Implications for validity. *Educational Measurement: Issues and Practice*, *31*(3), 2–9.

Benson, A., Lawler, C., & Whitworth, A. (2008). Rules, roles and tools: Activity theory and the comparative study of e-learning. *British Educational Research Journal*, *39*, 456–467. http://doi.org/10.1111/j.1467-8535.2008.00838.x

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability (Formerly: Journal of Personnel Evaluation in Education)*. http://doi.org/10.1007/s11092-008-9068-5

Boeije, H. (2002). A purposeful approach to the constant comparative method in the analysis of qualitative interviews. *Quality & Quantity*, *36*, 391–409.

Bowen, G. (2009). Document analysis as a qualitative research method. *Qualitative Research Journal*, *9*(2), 27–40.

Bowles, M. A. (2010). *The think-aloud controversy in second language research*. New York, NY: Routledge.

Broad, B. (2003). *What we really value: Beyond rubrics in teaching and assessing writing*. Logan, UT: Utah State University Press.

Brown, J. D. (1991). Do English and ESL faculties rate writing samples differently? *TESOL Quarterly*, *25*, 587–603.

Cardon, P. W., & Scott, J. C. (2003). Chinese business face: Communication behaviours and teaching approaches. *Business Communication Quarterly*, *66*(4), 9–22.

Charney, D. (1984). The validity of using holistic scoring to evaluate writing: A critical overview. *Research in the Teaching of English*, *18*(1), 65–81.

Cheng, L. (2005). *Changing language teaching through language testing: A washback study*. Cambridge, UK: Cambridge University Press.

Cheng, L. (2008). The key to success: English language testing in China. *Language Testing*, *25*(1), 15–37. http://doi.org/10.1177/0265532207083743

Cheng, L., & Curtis, A. (2010). The impact of English language assessment and the Chinese learner in China and beyond. In L. Cheng & A. Curtis (Eds.), *English language assessment and the Chinese learner* (pp. 267–273). New York, NY: Routledge.

Cheng, L., & Qi, L. (2006). Description and examination of the National Matriculation English Test. *Language Assessment Quarterly*, *3*, 53–70. http://doi.org/10.1207/s15434311laq0301_4

Cheng, L., Sun, Y., & Ma, J. (2015). Review of washback research literature within Kane's argument-based validation framework. *Language Teaching*, *48*, 436–470. http://doi.org/10.1017/S0261444815000233

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, MA: The Massachusetts Institute of Technology Press.

Cole, M. (1976). Foreword. In M. Cole (Ed.), *Cognitive development its cultural and social foundation*. Cambridge, MA: Harvard University Press.

Cole, M., & Engeström, Y. (1993). A cultural-historical approach to distributed cognition. In G. Salomon (Ed.), *Distributed cognitions: Psychological and educational considerations* (pp. 1–46). Cambridge, UK: Cambridge University Press.

Commonwealth of Australia. (2009). *Research on China's National College Entrance Examination (the Gaokao)*. Retrieved from http://sydney.edu.au/ab/committees/admissions/2011/AEI_Gaokao_Report.pdf

Connor-Linton, J. (1995). Crosscultural comparison of writing standards: American ESL and Japanese EFL. *World Englishes*, *14*, 99–115.

Coughlan, P., & Duff, P. (1994). Same task, different activities: Analysis of SLA task from an activity theory perspective. In J. P. Lantolf & G. Appel (Eds.), *Vygotskian approaches to second language research* (pp. 173–194). Norwood, NJ: Ablex.

Crisp, V. (2008). Exploring the nature of examiner thinking during the process of examination marking. *Cambridge Journal of Education, 38,* 247–264. http://doi.org/10.1080/03057640802063486

Crisp, V. (2010). Towards a model of the judgement processes involved in examination marking. *Oxford Review of Education, 36*(1), 1–21. http://doi.org/10.1080/03054980903454181

Crisp, V. (2012). An investigation of rater cognition in the assessment of projects. *Educational Measurement: Issues and Practice, 31*(3), 10–20. http://doi.org/10.1111/j.1745-3992.2012.00239.x

Crossouard, B. (2009). A sociocultural reflection on formative assessment and collaborative challenges in the states of Jersey. *Research Papers in Education, 24*(1), 77–93. http://doi.org/10.1080/13669870801945909

Cumming, A. (1990). Expertise in evaluating second language compositions. *Language Testing, 7*(1), 31–51. http://doi.org/10.1177/026553229000700104

Cumming, A. (Ed.). (2006). *Goals for academic writing: ESL students and their instructors*. Amsterdam, the Netherlands: John Benjamins Publishing Company.

Cumming, A., Kantor, R., Baba, K., Erdosy, U., Eouanzoui, K., & James, M. (2005). Differences in written discourse in independent and integrated prototype tasks for next generation TOEFL. *Assessing Writing, 10,* 5–43. http://doi.org/10.1016/j.asw.2005.02.001

Cumming, A., Kantor, R., & Powers, D. E. (2002). Decision making while rating ESL/EFL writing tasks: A descriptive framework. *The Modern Language Journal, 86*(1), 67–96. http://doi.org/10.1111/1540-4781.00137

Deane, P. (2013). On the relation between automated essay scoring and modern views of the writing construct. *Assessing Writing, 18,* 7–24. http://doi.org/10.1016/j.asw.2012.10.002

Delaruelle, S. (1997). Text type and rater decision-making in the writing module. In G. Brindley & G. Wigglesworth (Eds.), *Access: Issues in language test design and delivery* (pp. 215– 242). Sydney, Australia: National Center for English Language Teaching and Research, Macquarie University.

DeRemer, M. L. (1998). Writing assessment: Raters' elaboration of the rating task. *Assessing Writing, 5,* 7–29.

Diederich, P. B., French, J. W., & Carlton, S. T. (1961). *Factors in judgments of writing ability* (RB-61-15). Princeton, NJ: Educational Testing Service.

Earley, P. C. (1997). *Face, harmony, and social structure: An analysis of organizational behavior across cultures*. New York, NY: Oxford University Press.

Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing*, *25*(2), 155–185. http://doi.org/10.1177/0265532207086780

Elder, C., Barkhuizen, G., Knoch, U., & von Randow, J. (2007). Evaluating rater responses to an online training program for L2 writing assessment. *Language Testing*, *24*, 37–64. http://doi.org/10.1177/0265532207071511

Elder, C., Knoch, U., Barkhuizen, G., & von Randow, J. (2005). Individual feedback to enhance rater training: Does it work? *Language Assessment Quarterly*, *2*, 175–196.

Elliot, N., & Williamson, D. M. (2013). *Assessing Writing* special issue: Assessing writing with automated scoring systems. *Assessing Writing*, *18*, 1–6. http://doi.org/10.1016/j.asw.2012.11.002

Elliott, V. (2013). Empathetic projections and affect reactions in examiners of "A" level English and History. *Assessment in Education: Principles, Policy & Practice*, *20*, 266–280. http://doi.org/10.1080/0969594X.2013.768597

Engelhard, G. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement*, *31*(2), 93–112.

Engeström, Y. (1987). *Learning by expanding: An activity-theoretical approach to developmental research*. Helsinki, Finland: Orienta-Konsultit.

Engeström, Y. (1993). Developmental studies of work as a testbench of activity theory: The case of primary care medical practice. In S. Chaiklin & J. Lave (Eds.), *Understanding practice: Perspectives on activity and context* (pp. 64–103). New York, NY: Cambridge University Press.

Engeström, Y. (1999). Activity theory and individual and social transformation. In Y. Engeström, R. Miettinen, & R.-L. Punamaki (Eds.), *Perspectives on activity theory* (pp. 19–38). New York, NY: Cambridge University Press.

Engeström, Y. (2000). Activity theory as a framework for analyzing and redesigning work. *Ergonomics*, *43*, 960–974. http://doi.org/10.1080/001401300409143

Engeström, Y. (2001). Expansive learning at work: Toward an activity theoretical reconceptualization. *Journal of Education and Work*, *14*(1), 133–156. http://doi.org/10.1080/13639080020028747

Engeström, Y. (2015). *Learning by expanding: An activity-theoretical approach to developmental research* (2nd ed.). New York, NY: Cambridge University Press. http://doi.org/10.1016/j.intcom.2007.07.003

Erdosy, U. (2005). *Responding to non-native and native writers of English: A history professor's indigenous criteria for grading and feedback in an undergraduate sinology course* (Unpublished doctoral dissertation). The Ontario Institute for Studies in Education of the University of Toronto, Toronto, Ontario, Canada.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data* (Rev.). Cambridge, MA: MIT Press.

Freedman, S. W., & Calfee, R. C. (1983). Holistic assessment of writing: Experimental design and cognitive theory. In P. Mosenthal, L. Tamor, & S. A. Walmsley (Eds.), *Research on writing: Principles and methods* (pp. 75–98). New York, NY: Longman.

Furneaux, C., & Rignall, M. (2007). The effect of standardisation-training on rater judgements for the IELTS writing module.pdf. In L. Taylor & P. Falvey (Eds.), *IELTS collected papers: Research in speaking and writing assessment* (pp. 422–445). Cambridge, UK: Cambridge University Press/Cambridge ESOL.

Gao, H. Y. (2011). Study on construct validity of CET4 writing: An attempt based on internal structure of test and assessment process. *Foreign Language Testing and Teaching*, (4), 33–41.

Gass, S. M., & Mackey, A. (2000). *Stimulated recall methodology in second language research*. Mahwah, NJ: Lawrence Erlbaum Associates.

Gass, S. M., & Mackey, A. (2007). *Data elicitation for second and foreign language research*. Mahwah, NJ: Lawrence Erlbaum Associates.

Gebril, A., & Plakans, L. (2014). Assembling validity evidence for assessing academic writing: Rater reactions to integrated tasks. *Assessing Writing*, *21*, 56–73. http://doi.org/10.1016/j.asw.2014.03.002

Gentil, G. (2006). Variations in goals and activities for multilingual writing. In A. Cumming (Ed.), *Goals for academic writing: ESL students and their instructors* (pp. 142–156). Amsterdam, the Netherlands: John Benjamins Publishing Company.

Goffman, E. (1967). *On face-work: An analysis of ritual elements in social interaction*. *Interaction ritual: Essays in face-to-face behavior*. London, England: Penguin Books.

Green, A. (1998). *Verbal protocol analysis in language testing research*. Cambridge, UK: Cambridge University Press.

Greig, G. (2008). *The role and importance of context in collective learning: Multiple case studies in Scottish primary care* (Unpublished doctoral dissertation). University of St. Andrews, Fife, Scotland, UK.

Gu, P. Y. (2013). The unbearable lightness of the curriculum: What drives the assessment practices of a teacher of English as a Foreign Language in a Chinese secondary school? *Assessment in Education: Principles, Policy & Practice, 21*, 286–305. http://doi.org/10.1080/0969594X.2013.836076

Gu, X. (2007). *Positive or negative: An empirical study of CET washback*. Chongqing, China: Chongqing University Press.

Hall, A. L., & Rist, R. C. (1999). Integrating multiple qualitative research methods (or avoiding the precariousness of a one-legged stool). *Psychology & Marketing*, *16*, 291–304.

Hamp-Lyons, L. (1989). Raters respond to rhetoric in writing. In H. W. Dechert & M. Raupach (Eds.), *Interlingual Processes* (pp. 229–244). Tubingen, Germany: Gunter Narr Verlag Tubingen.

Hamp-Lyons, L. (1991). Basic concepts. In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 5–15). Norwood, NJ: Ablex Publishing Corporation.

Hamp-Lyons, L. (2007). Worrying about rating. *Assessing Writing*, *12*, 1–9. http://doi.org/10.1016/j.asw.2007.05.002

Hamp-Lyons, L., & Zhang, W. (2001). World Englishes: Issues in and from academic writing assessment. In J. Flowerdew & M. Peacock (Eds.), *Research perspectives on English for academic purposes* (pp. 101–116). Cambridge, UK: Cambridge University Press.

Haneda, M. (2007). Modes of engagement in foreign language writing: An activity theoretical perspective. *The Canadian Modern Language Review*, *64*, 301–331.

Ho, D. Y. (1976). On the concept of face. *American Journal of Sociology*, *81*(4), 867–884.

Homburg, T. J. (1984). Holistic evaluation of ESL compositions: Can it be validated objectively? *TESOL Quarterly*, *18*, 87–107.

Huang, J. (2008). How accurate are ESL students' holistic writing scores on large-scale assessments?—A generalizability theory approach. *Assessing Writing*, *13*, 201–218. http://doi.org/10.1016/j.asw.2008.10.002

Huang, X. (2011). *The washback effects of national college entrance examination: Chinese language teachers' beliefs and practices* (Unpublished doctoral dissertation). The Chinese University of Hong Kong, Hong Kong, China.

Huot, B. (1990). The literature of direct writing assessment: Major concerns and prevailing trends. *Review of Educational Research*, *60*, 237–263. http://doi.org/10.3102/00346543060002237

Huot, B. (1993). The influence of holistic scoring procedures on reading and rating student essays. In M. Williamson & B. Huot (Eds.), *Validating holistic scoring for writing assessment: Theoretcial and empirical foundations* (pp. 206–236). Cresskill, NJ: Hampton Press, Inc.

Huot, B. (2002). *(Re)Articulating writing assessment for teaching and learning*. Logan, UT: Utah State University Press.

Hwang, A., Francesco, A. M., & Kessler, E. (2003). The Relationship between individualism-collectivism, face, and feedback and learning processes in Hong Kong, Singapore, and the United States. *Journal of Cross-Cultural Psychology*, *34*(1), 72–91. http://doi.org/10.1177/0022022102239156

Hwang, K. (1987). Face and favor: The Chinese power game. *American Journal of Sociology*, *92*, 944–974.

Johnson, J., & Lim, G. S. (2009). The influence of rater language background on writing performance assessment. *Language Testing*, *26*, 485–505. http://doi.org/10.1177/0265532209340186

Karasavvidis, I. (2009). Activity Theory as a conceptual framework for understanding teacher approaches to Information and Communication Technologies. *Computers & Education*, *53*, 436–444. http://doi.org/10.1016/j.compedu.2009.03.003

Knoch, U., Read, J., & von Randow, J. (2007). Re-training writing raters online: How does it compare with face-to-face training? *Assessing Writing*, *12*, 26–43. http://doi.org/10.1016/j.asw.2007.04.001

Kondo-Brown, K. (2002). A FACETS analysis of rater bias in measuring Japanese second language writing performance. *Language Testing*, *19*, 3–31. http://doi.org/10.1191/0265532202lt218oa

Lantolf, J. P., & Thorne, S. L. (2006). *Sociocultural theory and the genesis of second language development*. New York, NY: Oxford University Press.

Lee, H. K. (2004). A comparative study of ESL writers' performance in a paper-based and a computer-delivered writing test. *Assessing Writing*, *9*, 4–26. http://doi.org/10.1016/j.asw.2004.01.001

Lei, X. (2008). Exploring a sociocultural approach to writing strategy research: Mediated actions in writing activities. *Journal of Second Language Writing*, *17*, 217–236. http://doi.org/10.1016/j.jslw.2008.04.001

Lei, X. (2009). *Understanding writing strategy use from a sociocultural perspective: A multiple case study of Chinese EFL learners of different writing abilities* (Unpublished doctoral dissertation). The University of Hong Kong, Hong Kong, China.

Leont'ev, A. N. (1981). *Problems of the development of the mind*. Moscow, Russia: Progress Publishers.

Li, H., & He, L. (2015). A comparison of EFL raters' essay-rating processes across two types of rating scales. *Language Assessment Quarterly*, *12*, 178–212. http://doi.org/10.1080/15434303.2015.1011738

Li, J. (2016). The interactions between emotion, cognition, and action in the activity of assessing undergraduates' written work. In D. S. P. Gedera & P. J. Williams (Eds.), *Activity theory in education: Research and practice* (pp. 107–119). Rotterdam, the Netherlands: Sense Publishers. http://doi.org/10.1017/CBO9781107415324.004

Lim, G. S. (2011). The development and maintenance of rating quality in performance writing assessment: A longitudinal study of new and experienced raters. *Language Testing*, *28*, 543–560. http://doi.org/10.1177/0265532211406422

Lin, Y. (1936). *My country and my people*. Kingswood, UK: William Heinemann Ltd.

Liu, H. (2010). Where should the National College Entrance Examination go? *Peking University Education Review*, *8*(2), 2–13.

Liu, H., & Guo, D. (2003). A research of the relationship between pretest anxiety, achievement goal orientation and test performance. *Psychological Development and Education*, (2), 64–68.

Liu, J. (2005). 我国高考英语测试的历史与现状 [History and status quo of NMET]. *Shanghai Research on Education*, (3), 40–43.

Liu, Q. (2010). The National Education Examinations Authority and its English language tests. In L. Cheng & A. Curtis (Eds.), *English language assessment and the Chinese learner* (pp. 29–43). New York, NY: Routledge.

Liu, X. (2011). 高考户籍制的历史镜像、现实困境与反思[History, contemporary dilemmas and reflections on Gaokao Hukou system]. *Journal of National Academy of Education Administration*, (11), 57–61.

Liu, X., & Liu, H. (2012). Migration examination problems of Imperial Examinations in Qing Dynasty and its enlightenment. *Education Research, 1*, 141–147.

Lu, Y. (2010). *Rater bias studies in online TEM4 essay marking* (Unpublished doctoral dissertation). Shanghai International Studies University, Shanghai, China.

Lukmani, Y. (1996). Linguistic accuracy versus coherence in assessing examination answers in content subjects. In M. Milanovic & N. Saville (Eds.), *Performance testing, cognition and assessment* (pp. 130–150). Cambridge, UK: Cambridge University Press.

Lumley, T. (2002). Assessment criteria in a large-scale writing test: What do they really mean to the raters? *Language Testing, 19*, 246–276. http://doi.org/10.1191/0265532202lt230oa

Lumley, T. (2005). *Assessing second language writing: The rater's perspective*. New York, NY: Peter Lang.

Lyle, J. (2003). Stimulated recall: A report on its use in naturalistic research. *British Educational Research Journal, 29*, 861–878. http://doi.org/10.1080/0141192032000137349

Mackey, A., & Gass, S. M. (2005). *Second language research: Methodology and design*. Mahwah, NJ: Lawrence Erlbaum.

Marken, J. A. (2006). An application of activity theory: A case of global training. *Performance Improvement Quarterly*, *19*(2), 27–50. http://doi.org/10.1111/j.1937-8327.2006.tb00364.x

McNamara, T., & Roever, C. (2006). *Language testing: The social dimension*. Malden, MA: Blackwell Publishing.

Mei, Y., & Cheng, L. (2014). Scoring fairness in large-scale high-stakes English language testing: An examination of the National Matriculation English Test. In D. Codium (Ed.), *English language education and assessment: Recent developments in Hong Kong and Chinese Mainland* (pp. 171–187). Singapore: Springer Science+BusinessMedia LLC. http://doi.org/10.1007/978-981-287-071-1_11

Merriam, S. B. (2009). *Qualitative research: A guide to design and implementation*. San Francisco, CA: Jossey-Bass.

Messick, S. (1996). Validity and washback in language testing. *Language Testing*, *13*, 241–256. http://doi.org/10.1177/026553229601300302

Milanovic, M., Saville, N., & Shen, S. (1996). A study of the decision-making behaviour of composisiton markers. In M. Milanovic & N. Saville (Eds.), *Performance test, cognition and assessment: Selected papers from the 15th Lanugage Testing Research Colloquium (LTRC), Cambridge and Arnhem* (pp. 92–114). Cambridge, MA: Cambridge University Press.

Miller, F. A., & Alvarado, K. (2005). Incorporating documents into qualitative nursing research. *Journal of Nursing Scholarship*, 348–353.

Ministry of Education (MoE). (2013). 教育部关于推进中小学教育 质量综合评价改革的意见*[MoE: Promoting evaluation reform in primary and secondary education]*. Retrieved from http://www.moe.edu.cn/publicfiles/business/htmlfiles/moe/s7054/201306/153185.html

Mwanza, D. (2002). Conceptualising work activity for CAL systems design. *Journal of Computer Assisted Learning*, *18*, 84–92.

Mwanza, D., & Engeström, Y. (2003). Pedagogical adeptness in the design of e-learning environments: Experiences from the Lab @ Future project. In A. Rossett (Ed.), *Proceedings of E-Learn 2003 International Conference on E-Learning in Corporate, Government, Healthcare, & Higher Education* (pp. 1344–1347). Phoenix, AZ. Paper retrieved from https://www.researchgate.net/publication/42795915_Pedagogical_Adeptness_in_the_Design_of_E-learning_Environments_Experiences_from_the_LabFuture_Project

Myford, C. M., & Mislevy, R. J. (1995). *Monitoring and improving a portfolio assessment system* (CSE Technical Report 402). Los Angeles, CA: National Centre for Research on Evaluation, Standards, and Student Testing (CRESST).

Parks, S. (2000). Same task, different activities: Issues of investment, identity, and use of strategy. *TESL Canada Journal*, *17*(2), 64–88.

Patton, M. Q. (2002). *Qualitative research and evaluation methods* (3rd ed.). Thousand Oaks, CA: SAGE Publications, Inc.

Patton, M. Q. (2015). *Qualitative research and evaluation methods: Integrating theory and practice* (4th ed.). Thousand Oaks, CA: SAGE Publications, Inc.

Pryor, J., & Crossouard, B. (2008). A socio-cultural theorisation of formative assessment. *Oxford Review of Education*, *34*(1), 1–20. http://doi.org/10.1080/03054980701476386

Pula, J., & Huot, B. (1993). A model of background influences on holistic raters. In M. M. Williamson & B. Huot (Eds.), *Validating holistic scoring for writing assessment:*

Theoretcial and empirical foundations (pp. 237–265). Cresskill, NJ: Hampton Press, Inc.

Qi, L. (2005). Stakeholders' conflicting aims undermine the washback function of a high-stakes test. *Language Testing*, *22*, 142–173. http://doi.org/10.1191/0265532205lt300oa

Qi, L. (2007). Is testing an efficient agent for pedagogical change? Examining the intended washback of the writing task in a high-stakes English test in China. *Assessment in Education: Principles, Policy & Practice*, *14*, 51–74. http://doi.org/10.1080/09695940701272856

Qi, L. (2010). Should proofreading go? Examining the selection function and washback of the proofreading sub-test in the National Matriculation English Test. In L. Cheng & A. Curtis (Eds.), *English language assessment and the Chinese learner* (pp. 219–233). New York, NY: Routledge.

Qi, X. (2011). Face: A Chinese concept in a global sociology. *Journal of Sociology*, *47*, 279–295. http://doi.org/10.1177/1440783311407692

Redding, S. G., & Ng, M. (1982). The role of "face" in the organizational perceptions of Chinese managers. *Organisation Studies*, *3*, 92–123.

Rinnert, C., & Kobayashi, H. (2001). Differing perceptions of EFL writing among readers in Japan. *The Modern Language Journal*, *85*, 189–209.

Russo, J. E., Johnson, E. J., & Stephens, D. L. (1989). The validity of verbal protocols. *Memory & Cognition*, *17*, 759–769.

Sakyi, A. A. (2000). Validation of holistic scoring for ESL writing assessment: How raters evaluate compositions. In A. Kunnan (Ed.), *Fairness and validation in language assessment* (pp. 129–152). Cambridge, UK: Cambridge University Press.

Sakyi, A. A. (2003). *A Study of the holistic scoring behaviours of experienced and novice ESL instructors* (Unpublished doctoral dissertation). The Ontario Institute for Studies in Education of the University of Toronto, Toronto, Ontario, Canada.

Saldaña, J. (2009). *The coding manual for qualitative researchers*. Thousand Oaks, CA: SAGE Publications, Inc.

Sanderson, P. (2001). *Language and differentiation in examining at A level* (Unpublished doctoral dissertation). University of Leeds, Leeds, United Kingdom.

Sasaki, T. (2003). Recipient orientation in verbal report protocols: Methodological issues in concurrent think-aloud. *Second Language Studies*, *22*(1), 1–54.

Schaefer, E. (2008). Rater bias patterns in an EFL writing assessment. *Language Testing*, *25*, 465–493. http://doi.org/10.1177/0265532208094273

Shi, L. (2001). Native- and nonnative-speaking EFL teachers' evaluation of Chinese students' English writing. *Language Testing*, *18*, 303–325.

Shohamy, E., Gordon, C. M., & Kraemer, R. (1992). The effect of raters' background and training on the reliability of direct writing tests. *The Modern Language Journal*, *76*, 27–33.

Smagorinsky, P. (1994). Think-aloud protocol analysis: Beyond the black box. In *Speaking about writing: Reflections in research methodology* (pp. 3–19). Thousand Oaks, CA: SAGE Publications, Inc.

Smagorinsky, P. (1998). Thinking and Speech and Protocol Analysis. *Mind, Culture, and Activity*, *5*(3), 157–177. http://doi.org/10.1207/s15327884mca0503_2

Smagorinsky, P. (2001). Rethinking protocol analysis from a cultural perspective. *Annual Review of Applied Linguistics*, *21*, 233–245.

Smith, D. (2000). Rater judgments in the direct assessment of competency-based second language writing ability. In G. Brindley (Ed.), *Studies in immigrant English language assessment studies in immigrant English language assessment* (Vol. 1, pp. 159–189). Sydney, Australia: Macquarie University Press.

Strauss, A., & Corbin, J. (1990). *Basics of qualitative research: Grounded theory procedures and techniques*. Newbury Park, CA: SAGE Publications, Inc.

Strauss, A., Schatzman, L., Bucher, R., Ehrlich, D., & Sabshin, M. (1981). *Psychiatric ideologies and institutions*. New York, NY: The Free Press of Glencoe.

Swain, M. (2006). Verval protocols: What does it mean for research to use speaking as a data collection tool? In M. Chalhoub-Deville, C. A. Chapelle, & P. A. Duff (Eds.), *Inference and Generalizability in Applied Linguistics: Multiple perspectives* (pp. 97–114). Philadelphia, PA: John Benjamins Publishing Company.

Swain, M., Kinnear, P., & Steinman, L. (2010). *Sociocultural theory in second language education: An introduction through narratives*. Bristol, UK: Multilingual Matters.

Twiselton, S. (2004). The role of teacher identities in learning to teach primary literacy. *Educational Review*, *56*(2), 157–164. http://doi.org/10.1080/0031910410001693245

Vaughan, C. (1991). Holistic assessment: What goes on in the raters' minds. In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 111–126). Norwood, NJ: Ablex Publishing Corporation.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. London, England: Harvard University Press.

Wang, D., & Gao, M. (2013). Educational equality or social mobility: The value conflict between preservice teachers and the Free Teacher Education Program in China. *Teaching and Teacher Education, 32*, 66–74. http://doi.org/10.1016/j.tate.2013.01.008

Wang, H., & Zhan, X. (2011). Multidimensional analysis on fairness of English testing in College Entrance Examination. *Curriculum, Teaching Material and Method*, *31*(5), 49–53.

Warschauer, M., & Ware, P. (2006). Automated writing evaluation: Defining the classroom research agenda. *Language Teaching Research*, *10*, 157–180.

Webb, M., & Jones, J. (2009). Exploring tensions in developing assessment for learning. *Assessment in Education: Principles, Policy & Practice*, *16*, 165–184. http://doi.org/10.1080/09695940903075925

Weigle, S. (1999). 1 Investigating rater/prompt interactions in writing assessment: Quantitative and qualitative approaches. *Assessing Writing*, *6*, 145–178.

Weigle, S. C. (1994). Effects of training on raters of ESL compositions. *Language Testing*, *11*, 197–223. http://doi.org/10.1177/026553229401100206

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing*, *15*, 263–287. http://doi.org/10.1177/026553229801500205

Weigle, S. C. (2002). *Assessing writing*. Cambridge, UK: Cambridge University Press.

Wertsch, J. V. (1985). *Vygotsky and the social formation of mind*. Cambridge, MA: Harvard University Press.

Wertsch, J. V. (1998). *Mind as action*. New York, NY: Oxford University Press.

White, E. M. (1993). Holistic scoring: Past triumphs, future challenges. In M. M. Williamson & B. Huot (Eds.), *Validating holistic scoring for writing assessment: Theoretical and empirical foundations* (pp. 79–108). Cresskill, NJ: Hampton Press, Inc.

Wiseman, C. S. (2012). Rater effects: Ego engagement in rater decision-making. *Assessing Writing*, *17*, 150–173. http://doi.org/10.1016/j.asw.2011.12.001

177 Webster St., # 220, Monterey, CA 93940 USA
Web: www.tirfonline.org / **Email:** info@tirfonline.org

Wolfe, E. W. (1997). The relationship between essay reading style and scoring proficiency in a psychometric scoring system. *Assessing Writing*, *4*, 83–106. http://doi.org/10.1016/S1075-2935(97)80006-2

Wolfe, E. W., Kao, C., & Ranney, M. (1998). Cognitive differences in proficient and nonproficient essay scorers. *Written Communication*, *15*, 465–492.

Xi, X. (2010). Automated scoring and feedback systems: Where are we and where are we heading? *Language Testing*, *27*, 291–300. http://doi.org/10.1177/0265532210364643

Yamagata-Lynch, L. C. (2003). Using activity theory as an analytic lens for examining technology professional development in schools. *Mind, Culture, and Activity*, *10*, 100–119. http://doi.org/10.1207/S1532-7884MCA1002_2

Yamagata-Lynch, L. C. (2010). *Activity systems analysis methods: Understanding complex learning environments*. Boston, MA: Springer. http://doi.org/10.1007/978-1-4419-6321-5

Yamagata-Lynch, L. C., & Haudenschild, M. T. (2009). Using activity systems analysis to identify inner contradictions in teacher professional development. *Teaching and Teacher Education*, *25*, 507–517. http://doi.org/10.1016/j.tate.2008.09.014

Yin, R. K. (2009). *Case study research: Design and methods* (4th ed.). Thousand Oaks, CA: SAGE Publications, Inc.

Yin, R. K. (2011). *Qualitative research from start to finish*. *Qualitative research from start to finish*. New York, NY: The Guilford Press. http://doi.org/10.1007/s13398-014-0173-.2

Yin, R. K. (2014). *Case study research: Design and methods* (5th ed.). Thousand Oaks, CA: SAGE Publications, Inc.

Zeng, Y. (2010). The Computerized Oral English Test of the National Matriculaton English Test. In L. Cheng & A. Curtis (Eds.), *English language assessment and the Chinese learner* (pp. 234–247). New York, NY: Routledge.

Zhang, J. (2009). *Exploring rating process and rater belief: Seeking the internal account for rater variability* (Unpublished doctoral dissertation). Guangdong University of Foreign Studies, Guangzhou, China.

Zhang, W. (1999). *The rhetorical patterns found in Chinese EFL student writers ' examination essays in English and the influence of these patterns on rater response* (Unpublished doctoral dissertation). The Hong Kong Polytechnic University, Hong Kong, China.

Zheng, R. (2010). The National College Entrance Examination Reform: Concerns and practice. *Peking University Education Review*, *8*(2), 14–29.

Zheng, R., & Chen, W. (2013). A study on negative washback of large-scale high-stakes test. *Journal of Huazhong Normal University (Humanities and Social Sciences)*, *52*(1), 147–154.